Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

MINERÍA WEB COMO HERRAMIENTA DE ANÁLISIS DE FICHEROS LOG EN SERVIDORES WEB

Jorge Luis Say Valdez

Asesorado por el Ing. Osberto Alejandro Pineda Gonzáles

Guatemala, marzo de 2010

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA

MINERÍA WEB COMO HERRAMIENTA DE ANÁLISIS DE FICHEROS LOG EN SERVIDORES WEB

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA

POR

JORGE LUIS SAY VALDEZ

ASESORADO POR EL ING. OSBERTO ALEJANDRO PINEDA GONZÁLES

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, MARZO DE 2010

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	Inga. Glenda Patricia García Soria
VOCAL II	Inga. Alba Maritza Guerrero Spinola de López
VOCAL III	Ing. Miguel Angel Dávila Calderón
VOCAL IV	Br. Luis Pedro Ortíz de León
VOCAL V	Br. José Alfredo Ortíz Herincx
SECRETARIA	Inga. Marcia Ivónne Véliz Vargas

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Murphy Olympo Paiz Recinos
EXAMINADOR	Inga. Virginia Victoria Tala Ayerdi
EXAMINADOR	Ing. César Augusto Fernández Cáceres
EXAMINADOR	Ing. Edgar Estuardo Santos Sutuj
SECRETARIA	Inga. Marcia Ivónne Véliz Vargas

HONORABLE TRIBUNAL EXAMINADOR

Cumpliendo con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

MINERÍA WEB COMO HERRAMIENTA DE ANÁLISIS DE FICHEROS LOG EN SERVIDORES WEB,

tema que me fuera asignado por la Coordinación de la Carrera de Ingeniería en Ciencias y Sistemas, en agosto de 2008.

Jorge Luis Say Valdez

ACTO QUE DEDICO A:

Guatemala

Que es la tierra que me ha visto crecer y alcanzar cada una de mis metas. Para que con los conocimientos aprendidos pueda contribuir en su crecimiento y desarrollo.

Mi padre

Quien es la persona que más admiro por su perseverancia y capacidad de vencer obstáculos. Que con su ejemplo me ha inculcado el valor del trabajo, de la responsabilidad, de la honradez, de la humildad y del respeto ajeno.

Mi madre

Quien es mi eterna confidente y que con su amor siempre me ha cuidado desde que estaba en su vientre. Con su ejemplo me ha enseñado el valor de la fortaleza, de la paciencia, de la confianza, del amor a la vida y del temor a Dios.

Mis hermanos

Que me han regalado su amor y me han dado su apoyo desinteresadamente. Porque juntos hemos sido cómplices de tantas aventuras, además de compartir momentos tristes y alegres. Para que cuando llegue el momento de estar distantes no exista nada ni nadie que separe nuestros lazos de amor.

AGRADECIMIENTOS A:

Dios

Por ser el arquitecto de mi vida y darme la oportunidad de estar vivo. Porque siempre se hace presente en cada momento de mi existencia iluminándome y guiándome el sendero.

La Virgen de Guadalupe

Que es la madre que me dio la vida y la amiga fiel que siempre está a mi lado. Porque es mi refugio en los momentos de angustia y tristeza.

Mis padres

Pedro Leonardo Say Soc y *Sara Valdez Mazariegos de Say*. Que son los ángeles que Dios puso en mi camino a quienes amo y respeto. Gracias a su esfuerzo, trabajo y apoyo me encuentro en lugar que estoy. Porque a causa de ellos soy una persona con principios e integridad. Ellos son el motivo de mi superación y la fuente de mi inspiración. No tengo más dicha que el gran orgullo de ser su hijo.

Mi familia

Por el gran afecto y cariño que me guardan y por ser las personas que siempre me tienden la mano.

Mis profesores

Gracias por la pasión de compartir su experiencia y sabiduría. Desde la maestra que me enseño a leer y escribir hasta el último mentor de mi carrera de pre grado.

Mis amigos

Por su amistad, por los momentos compartidos, por su apoyo y sus palabras de aliento, ya que hasta el final fuimos un gran equipo. Espero que Dios los bendiga a cada uno por el camino que decidan tomar y que siempre logren alcanzar las metas que se tracen. Doy gracias a Dios por haberlos conocido.

El ingeniero Osberto A. Pineda

Por su paciencia y apoyo en la elaboración y culminación de este trabajo.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

PLANTEAMIENTO DEL PROBLEMA

Un problema común para cualquier organización es la falta de información disponible que sirva de respaldo para la toma de decisiones. Pero un problema aún mayor es contar con demasiada información útil y no saber cómo administrarla para sacarle provecho. Cuando ciertamente en el mundo actual se encuentra que la información es un factor intangible diferenciador y clave para alcanzar una ventaja competitiva.

Con el surgimiento de la Worl Wide Web (*WWW*) a principios de la década de los noventa y más aún con el fenómeno denominado *la burbuja* sucedido entre los años 1997 y 2001. Se ha visto un incremento exponencial en la red tanto en usuarios como en cantidad de información. La web rápidamente se ha convertido en el contenedor más amplio y conocido en información de cualquier tipo.

Cuando una organización crece en la red equitativamente también aumenta su volumen de información haciendo que el registro, control, manejo y uso de los datos como información tenga un grado mayor de dificultad. Actualmente según estudios realizados, expertos en la materia alrededor del mundo coinciden en que la información registrada en las bases de datos de las organizaciones se duplica en un período de 20 meses.

Ante esta masiva expansión de datos en la web, las técnicas conservadoras y tradicionalistas de análisis de información se han vuelto obsoletas. Esto debido a que las velocidades y tiempos entre almacenamiento y análisis no son proporcionales, siendo mayor los tiempos de almacenamiento. Debido a estas inconsistencias se pierde información y conocimiento útil que se podría extraer.

Con el crecimiento de los datos en Internet, se hace más complicado con el tiempo poder brindarle a cada usuario la información que busca conforme sus

preferencias. Ante tal situación surge la necesidad de crear sitios web que sean más persuasivos e inteligentes. Esto se puede llevar a cabo tomando como punto de partida el análisis de los ficheros históricos de cada sitio web con el fin de encontrar información oculta.

JUSTIFICACIÓN

La mayoría de organizaciones en la web por no decir todas siempre se encuentran en búsqueda de la mejora continua. Esto con fines de interés comercial para así poder aumentar los números en sus ganancias a través de sus clientes o usuarios. Ya que al final de cuentas en la red la competencia es bastante intensa y solo sobreviven los más fuertes e innovadores.

Las organizaciones deben ser capaces de entender y comprender las preferencias de sus clientes. Para lograr esto intentan analizar y explotar la gran cantidad de información que tienen a su disposición. Pero muchas veces no encuentran la manera eficaz de convertir la información en conocimiento de valor que sustente principalmente la correcta toma de decisiones a nivel gerencial y operacional.

A lo largo del tiempo la información paso de ser estadísticas de comparación a convertirse en un recurso estratégico dentro de una organización. Toda esta información en su mayoría es histórica, o sea que contiene todas las operaciones registradas en la línea del tiempo. Estos registros históricos son de vital importancia, ya que las organizaciones se basan de experiencias pasadas para tomar decisiones.

Es dentro de este contexto que surge como solución el estudio y la aplicación de la minería web. Ya que ésta a su vez emplea técnicas conocidas de la minería de datos, siendo capaz de encontrar y extraer información de los sitios web por medio de los ficheros históricos.

El análisis de un sitio web se puede llevar a cabo con facilidad, por medio de herramientas de minería web. Estas herramientas pueden trabajar debido a que todas las personas que navegan en la red dejan rastros denominados huellas

digitales. Asimismo quedan registros de todas las transacciones realizadas las cuales son almacenadas en bitácoras o ficheros logs. El resultado del análisis proporciona conocimiento útil para entender el comportamiento de los usuarios y las preferencias de navegación.

ANTECEDENTES

La inteligencia de negocios o BI (por sus siglas en inglés), tiene como función el procesamiento de los datos históricos de una organización. En los datos a procesar se encuentran los registros de ventas, las actividades de marketing, las bases de datos de los clientes, la cadena de suministros y los registros de compras entre otros.

La inteligencia de negocios es un proceso en constante evolución que tiene orígenes en la década de los ochenta. Se puede entender como parte de la inteligencia de negocios aquellas tecnologías, herramientas o aplicaciones de software que proporcionan extracción de conocimiento para la toma de decisiones.

Es dentro de la inteligencia de negocios que el descubrimiento de conocimiento a partir de las bases de datos o KDD (por sus siglas en inglés), tiene como objetivo la preparación de la información para luego interpretarla. El proceso de KDD emplea algoritmos de minería de datos para poder identificar conocimiento durante las etapas de pre procesamiento y post procesamiento.

Es entonces que para el análisis de la información que se encuentra en la red se hace uso de la minería web, que tiene implícitamente desarrollo de KDD. Las técnicas o algoritmos de minería integran otras disciplinas y se han podido desarrollar como tal gracias a la evolución de disciplinas como la estadística, la matemática, la inteligencia artificial y la computación. El avance de la minería web ha sido un proceso de investigaciones y pruebas. Las áreas que la comprenden se han desarrollado varias décadas atrás.

El camino de la minería web se inicia cuando se ve la necesidad de almacenar y procesar grandes volúmenes de información. Y es que tener la información adecuada en el momento y lugar preciso aumenta la productividad de cualquier organización.

RESUMEN

Debido a la creciente cantidad de información hoy en día, el análisis de la misma ya no puede ser manual ni incluso facilitado por herramientas de almacenes de datos u OLAP sino que este ha de ser semiautomático.

Es aquí que la minería de datos “no” transforma los datos para facilitar el acceso a la información y que el usuario la analice más fácilmente. Sino más bien que la minería de datos “analiza” los datos.

Una de las extensiones que conforma la minería de datos es la minería web, que consiste en el desarrollo de técnicas sobre el contenido de la web. Esto con el fin de encontrar patrones que puedan describir y permitir extraer conocimiento útil automáticamente de la web.

Existen diversos conceptos relacionados con la minería web (*Web mining*) pero en este caso se hace énfasis en la subdivisión llamada minería de uso de la web (*Web usage mining*), con la cual se puede evaluar y mejorar la calidad en los sitios web. Esencialmente sobre la base del comportamiento de los usuarios que la utilizan.

Para poder extraer conocimiento de un sitio web se debe seguir una serie de pasos básicos, con los cuales se descubren patrones y tendencias en los datos de un sistema a partir de los ficheros históricos (*logs*). El análisis de estos ficheros es significativo debido al valor de la información que se genera acerca de los usuarios relacionados y del sitio web en cuestión. Por mencionar algunos de estos datos estan: cantidad de usuarios, tiempos de navegación, páginas más visitadas, número de peticiones, patrones de navegación, localización geográfica, etcétera.

Las principales técnicas empleadas en la minería web y que son parte de la minería de datos son las siguientes: técnicas de clasificación, reglas de asociación, análisis de caminos, patrones secuenciales y las reglas de agrupamiento. Todas estas técnicas ayudan al descubrimiento de información y posteriormente al conocimiento intangible de los archivos en una organización.

Con el resultado obtenido del análisis del sitio se pueden llegar a obtener los patrones de comportamiento por parte de los usuarios. De esta manera se pueden agrupar y clasificarlos dichos patrones. Con el fin de mejorar aspectos tales como la personalización del servicio, la seguridad dentro del sitio y la reestructuración del contenido de las páginas entre otros. Todo esto conforme a los perfiles de usuario que previamente se han identificado y registrado.

OBJETIVOS

Generales :

1. Proporcionar los fundamentos teóricos relacionados con la minería web que tiene como fin detectar y extraer información útil de los datos históricos.

2. Analizar y evaluar la minería web como una herramienta de apoyo para la toma de decisiones de una organización en la red

Específicos :

1. Documentar los diferentes algoritmos que emplea la minería web para analizar, descubrir y predecir patrones de comportamiento en los usuarios.

2. Estudiar el proceso de conversión que siguen los datos cuando se aplica la minería de uso web en un sitio.

3. Aplicar los conceptos de minería web en un caso de estudio empleando una herramienta automatizada, con el fin de recopilar conocimiento útil.

INTRODUCCIÓN

Con el surgimiento de Internet los servidores web permiten almacenar de forma iterativa cada uno de los accesos a objetos de la red por parte de los usuarios. Los movimientos realizados quedan registrados en lo que se conoce como ficheros log o logs del servidor. Estos archivos son analizados periódicamente con la finalidad de obtener información de uso del servidor.

Los estudios tradicionales se han basado en tomar estadísticas un tanto avanzadas. Sin embargo, estas mediciones pueden presentar problemas como por ejemplo, que las operaciones registradas en el servidor sean menores a las operaciones reales, distorsionando así el control y la validez del estudio. Este problema en particular es causado por el surgimiento de cachés en la red que impide que la petición quede almacenada en el log.

Con el surgimiento de nuevos problemas que se dan cuando se analizan ficheros log, es cuando se necesitan tecnologías de vanguardia que sean más sofisticadas que las tradicionales y que puedan llegar a procesar información en tiempo real. Como respuesta a la búsqueda de estas tecnologías surge entre la comunidad informática un nuevo concepto llamado *minería web*.

La minería web es una herramienta que ayuda a las organizaciones a descubrir información y extraer conocimiento útil que no es visible a simple vista en la información que se tiene en la web. Con la aplicación de la minería web se pueden encontrar preferencias de usuarios, patrones de comportamiento y descubrir tendencias. Además sirve como apoyo para la toma de decisiones con respecto a cambios dentro del sitio.

1 MARCO TEÓRICO

1.1 Términos básicos de tecnología web

Para una mayor comprensión y entendimiento del tema, a continuación se citan algunos términos básicos que guardan relación con tecnología web.

1.1.1 Navegador web

En una arquitectura web de tipo cliente/servidor. A la aplicación que se encuentra del lado del cliente y que es usada para acceder a los recursos publicados por un servidor web se le conoce como navegador web o browser. Entre algunos navegadores web populares están: Firefox, Microsoft Internet Explorer, Netscape y Opera.

1.1.2 Servidor web

Los servidores web se encargan de publicar contenidos y recursos en la red para que sean accedidos por los usuarios. Algunos servidores web de mayor uso son: Microsoft Internet Information Server (IIS) y Apache HTTP Server.

1.1.3 Dirección IP

La dirección IP es un dato que identifica de manera única y exclusiva a cada ordenador que se encuentre conectado a Internet o a otra red. El formato de una dirección IP consta de 4 números menores a 255 separados por un punto.

Por ejemplo: *192.168.0.99*

1.1.4 Localizador uniforme de recurso (URL)

El localizador uniforme de recurso se encuentra formado por una cadena de caracteres con la cual se asigna una dirección a un recurso (página, video, audio, etcétera) que se encuentre disponible en la red. El protocolo de comunicación en la web es HTTP y el puerto por defecto es el 80. La sintaxis de una URL es:

http://dirección_de_la máquina/directorio/fichero

1.1.5 Sitio web

Un sitio web es un conjunto formado por contenidos publicados dentro de un servidor web. Estos contenidos a su vez comparten una estructura y se encuentran relacionados entre sí, con el fin de proporcionar recursos al usuario (catálogo de productos, menú de navegación, página principal, etcétera).

1.1.6 Logs

Se tratan de archivos que guardan un registro de las actividades ejecutadas por un usuario. Un log de servidor web almacena información de las peticiones realizadas (URL, fecha, sistema operativo, IP, etcétera) y el resultado de las mismas (código de respuesta). Los formatos más comunes de un archivo log de servidor web son: Common Logfile Format y Extended Logfile Format.

1.1.7 Contenido estático

Se le denomina así al contenido devuelto por un servidor web cuando se hace una petición y el servidor devuelve la misma junto con sus ficheros relacionados sin ningún tipo de tratamiento adicional.

1.1.8 Contenido dinámico

Es el tipo de contenido devuelto por un servidor web, donde la petición devuelta puede ser variable o dinámica con respecto a determinados parámetros de entrada o de usuario. Esto da una mayor flexibilidad a los servidores web y se ha logrado por medio del uso de tecnología como: Common Gateway Interface, Java Servlets, Lenguajes de scripting y servidores de dominio.

1.1.9 Sesión

Una sesión comprende todos y cada uno de los accesos registrados por un usuario. Todos estos accesos quedan almacenados en los ficheros log del servidor web. Una sesión dura el tiempo que el usuario permanezca dentro del sitio web.

1.1.10 Motor de búsqueda

Un motor de búsqueda es un sitio web, por medio del cual se llevan a cabo búsquedas o consultas. Existen dos tipos de motores de búsqueda: interno y externo. El motor de búsqueda interno se limita a efectuar búsquedas sobre las páginas internas del sitio web. El motor de búsqueda externo efectúa búsquedas sobre toda la web en general.

1.1.11 Consulta

Una consulta se encuentra formada por una o más palabras claves. Estas en conjunto representan una necesidad de búsqueda de información por parte de un usuario. Las consultas se pueden realizan en un motor de búsqueda y los resultados obtenidos dependen de las palabras clave introducidas.

1.2 Minería de datos

1.2.1 Definición

La minería de datos o data mining es un proceso de extracción de conocimiento considerablemente útil. A partir del análisis de datos pretende descubrir información oculta en los datos históricos contenidos en las bases de datos. Este análisis se lleva a cabo por medio de la aplicación de diferentes herramientas y tecnologías que logran identificar patrones en la información.

El conocimiento extraído puede predecir futuros comportamientos, cambios, estructuras, asociaciones, tendencias y anomalías. Estas situaciones en conjunto sirven de apoyo para la toma de decisiones en los procesos de negocio de los sitios web.

Para que la minería de datos funcione como tal es necesaria la integración e interacción de diversas áreas y disciplinas como la inteligencia artificial, la estadística, las bases de datos, la visualización de datos y gráficas, los procesamientos de volúmenes de información, el aprendizaje automático, la identificación y relación de patrones entre otros. Además se debe contar con un equipo de computación de alto procesamiento y rendimiento.

1.3 Historia de la minería de datos

El origen del concepto de minería de datos se puede remontar a los años sesenta. Para ese entonces algunos estadísticos empezaban a manejar términos como data fishing, data archaeology y data mining. El fin de sus investigaciones era encontrar correlaciones en las bases de datos con ruido.

En la década de los ochenta investigadores como Gregory Piatetsky-Shapiro, Rakesh Agrawal y Robert Blum entre otros, comenzaron a sentar las bases formales de la minería de datos.

A principios de la década de los noventa el grupo de empresas dedicadas al desarrollo de esta tecnología era mínimo. Para comienzos del año 2004, el número de empresas dedicadas a esta tecnología había aumentado considerablemente teniendo en el mercado aproximadamente 300 soluciones disponibles.

El desarrollo maduro de la minería de datos ha sido un proceso de investigaciones y las áreas que la comprenden se han desarrollado varias décadas atrás. El camino de la minería de datos se inicia cuando se ve la necesidad de almacenar y procesar la información. En la actualidad se encuentran soluciones para que la información sea analizada en tiempo real y de forma automática.

1.4 Aplicaciones de la minería de datos

El surgimiento de la minería de datos como tecnología ha adquirido gran popularidad en el mercado. Esto como resultado de los beneficios que la misma puede brindar a las empresas. Algunas áreas donde la aplicación de la minería de datos ha tenido una gran participación e importancia son:

- Reconocimiento y clasificación de rostros humanos.
- Investigación científica.
- Detección de actos ilícitos y delictivos.
- Soporte en bases de datos.
- Telecomunicaciones.
- Predicción del comercio y del marketing.
- Ingeniería reversa.

- Control de procesos industriales.
- Medicina.
- Análisis de la banca y de la bolsa de valores.
- Estudio de cubos sísmicos.
- Clasificación de perfil de los clientes.
- Análisis meteorológico.

1.5 Algoritmos de la minería de datos

Durante el proceso de análisis de la información la minería de datos emplea ciertos algoritmos como apoyo para el proceso y desarrollo del trabajo. A continuación se presenta una lista de aquellos métodos que se han aplicado en la minería de datos.

1.5.1 Redes neuronales

Las redes neuronales han sido de gran apoyo en la minería de datos porque estas tienen la capacidad de imitar la mente humana. Esto con el fin de encontrar patrones en la información. Las redes neuronales han tenido un papel importante en aquellas aplicaciones donde se trabaja en la clasificación de datos.

1.5.2 Reglas de inducción

Las reglas de inducción se aplican en la minería de datos, ya que se basan en reglas del tipo: "*Si un elemento A es parte de un evento X, entonces se clasifica como Y*". Todas estas reglas y relaciones son a partir de datos estadísticos de importancia.

1.5.3 Algoritmos estadísticos

Este tipo de algoritmos han sido empleados por investigadores y analistas, con el fin de controlar y detectar ciertos patrones inusuales dentro de los datos. Luego de la detección se pueden explicar los mismos por medio de modelos matemáticos y estadísticos.

1.5.4 Visualización de los datos

Con este método se puede a partir de una visualización de los datos en un momento dado, interpretar las relaciones existentes entre varios datos multidimensionales.

1.5.5 Método del vecino más cercano

Este método se aplica principalmente en aquellas bases de datos que contienen datos históricos. Este método consiste en clasificar cada uno de los registros a partir de un conjunto de datos con respecto a cierta combinación de clases de los "*n*" registros que tengan una mayor similitud con el mismo.

1.5.6 Algoritmos genéticos

Los algoritmos genéticos consisten en una serie de técnicas de optimización que usan ciertos procesos de una población de datos. Dentro de los procesos están la generación inicial de una población, la selección de datos, el cruzamiento genético y la mutación de datos. Todo esto dentro de un contorno que se basa en la evolución natural.

1.5.7 Clasificadores basados en instancias

Este método consiste en una serie de clasificadores. Estos tienen como objetivo separar y clasificar nuevos casos que se susciten a partir de casos similares que ya hayan sucedido y que se recuerden.

Cada clasificador necesita de teorías simbólicas. Hay ciertas dificultades en estos sistemas y consisten en preguntas como: ¿Cuáles casos deben ser recordados?, ¿Cómo se mide la similitud entre los casos? y ¿Cómo se debe relacionar el nuevo caso con los recordados?

1.6 Minería web

1.6.1 Definición de minería web

El concepto de minería web fue introducido por primera vez en el año de 1996 por O. Etzioni. La minería web es una de las extensiones que tiene la minería de datos. Consiste en aplicar las técnicas de la minería de datos sobre el contenido de la web. Esto con el fin de encontrar y extraer patrones de información y de comportamiento automáticamente de la web. Algunas áreas de estudio y desarrollo que abarca la minería web en la red son:

- Diseño web.
- Seguridad.
- Motores de búsqueda.
- Comercio electrónico.
- Posicionamiento web.

Con respecto al acceso, recuperación, organización y procesamiento de la información la minería web tiene un papel importante de aplicación en Internet. Ya que todas las personas que visitan un sitio web dejan huellas y pistas digitales en la red tal como: galletas, fecha y hora de acceso, navegador usado, dirección IP, URL de la página, códigos de error, protocolos de transmisión de datos, etcétera.

Toda esta información se almacena de forma automática en los servidores de los sitios por medio de bitácoras de accesos ó archivos log. Posteriormente las herramientas de la minería web analizan y procesan estos archivos para brindar información y conocimiento de cada uno de estos sitios. Por ejemplo como es que un cliente se comporta dentro del sitio antes de realizar una compra, un pago, una consulta, etcétera.

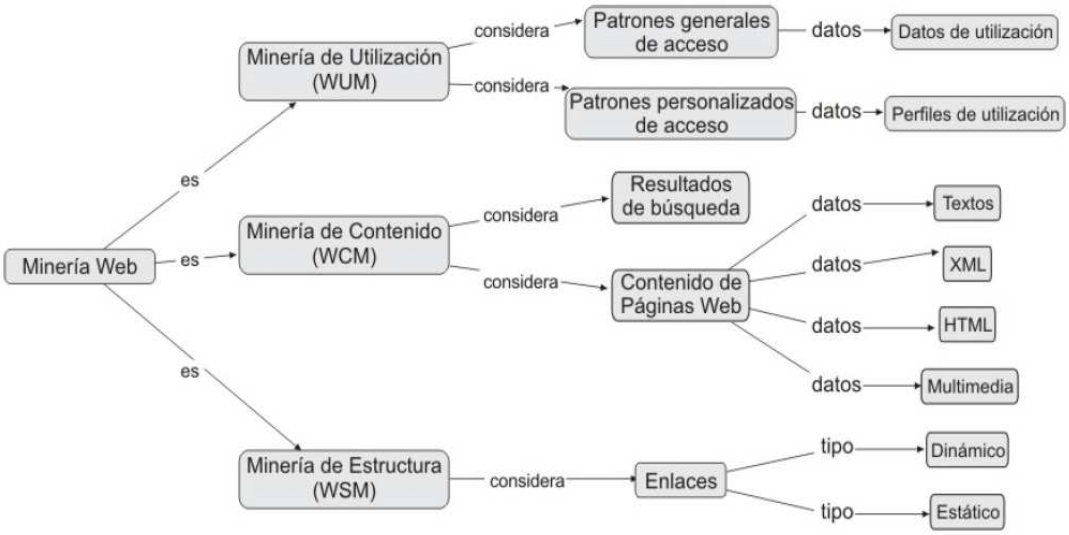
Debido a que el contenido de la red está formado por varios componentes como texto, imágenes, vídeos, metadatos entre otros. Se van formando nuevos términos para clasificar esta información, como por ejemplo minería de datos multimedia como una instancia de la minería web para ordenar los datos.

El proceso de análisis de la minería web se complementa con información como accesos totales por dominio, horarios de navegación, visitas por día, etcétera. Esta información extra es registrada y almacenada por las herramientas estadísticas.

1.6.2 Clasificación de la minería web

La minería web se divide en áreas que abarcan el contenido del sitio, la estructura de la navegación y el comportamiento de los usuarios. A continuación se presenta una gráfica relacionada.

Figura 1. Categorías de la minería web



Fuente: *http://www.infovis.net/printMag.php?num=172&lang=1*

1.6.2.1 Minería web de contenido

La minería web de contenido emplea técnicas de la minería de datos, con el fin de encontrar información precisa, sin limitarse en su búsqueda dentro del contenido. Los procesos llevados a cabo consisten en recuperar información para luego extraer conocimiento en base al contenido (documentos, imágenes, vídeos, audio, datos, links, textos, etcétera) de la web.

Dependiendo del contenido recuperado y de la consulta a contestar, se emplean técnicas de minería de texto y de minería web de estructura para organizar y clasificar el contenido. La finalidad de la minería web de contenido es perfeccionar el acceso de la información.

El mayor obstáculo que se presenta en la minería web de contenido esta dado por el contenido dinámico en la web. Otro problema es que los datos se pueden encontrar estructurados, semi estructurados ó que no se encuentren estructurados en su totalidad.

1.6.2.2 Minería web de estructura

La minería web de estructura permite conocer a fondo como se organiza y estructura la web. Su función es analizar y explorar los enlaces dentro de un sitio web. A partir de este análisis se generan grafos, luego se realizan consultas sobre los grafos con la finalidad de encontrar información relevante.

Con la aplicación de la minería web de estructura se puede saber cómo es el proceso de navegación de los usuarios. Esto ayuda al momento de reestructurar el contenido y la presentación de un sitio web. La minería web de estructura emplea algoritmos que tienen sistemas de peso para los nodos del grafo. Estos pesos permiten ponderar la información de un sitio web para saber si un tema específico tiene importancia o si no la tiene.

Un algoritmo popular en la minería web de estructura es el HITS. Este algoritmo toma una muestra de páginas web que contienen un tema común importante. En el proceso se grafican sub grafos donde cada nodo se expande hasta alcanzar un nivel pre establecido. El resultado del algoritmo es encontrar las páginas web más populares del tema.

Otro algoritmo relevante en la minería web de estructura es el que usa Google para encontrar el nivel de importancia de un sitio web sobre un tema específico. Este algoritmo se llama *PageRank*, sus creadores son Sergey Brin y Larry Page.

1.6.2.3 Minería de uso web

La minería de uso web es la encargada de detectar y extraer los patrones de comportamiento de los usuarios en la web. Además analiza los perfiles de los usuarios de un sitio web. Cuando se habla de patrones de comportamiento se hace referencia a la secuencia de grupos de páginas web que son accedidas de una manera más frecuente que otras páginas por parte de los usuarios.

Los patrones detectados ayudan a la reestructuración del sitio web. Además los patrones sirven para brindar una personalización dinámica dentro del sitio web al usuario con el ofrecimiento de contenidos acorde al perfil del mismo. Para realizar este tipo de minería se emplea comúnmente los archivos log que contienen registros de los sucesos y eventos de los servidores web.

Otros orígenes de información que sirven para realizar la minería de uso web son: bitácoras de acceso, bases de datos de usuarios, ontologías del sitio, agentes remotos y locales, bitácoras de referencia, información semántica, atributos de productos y cualquier otra fuente que aporte información del sitio web.

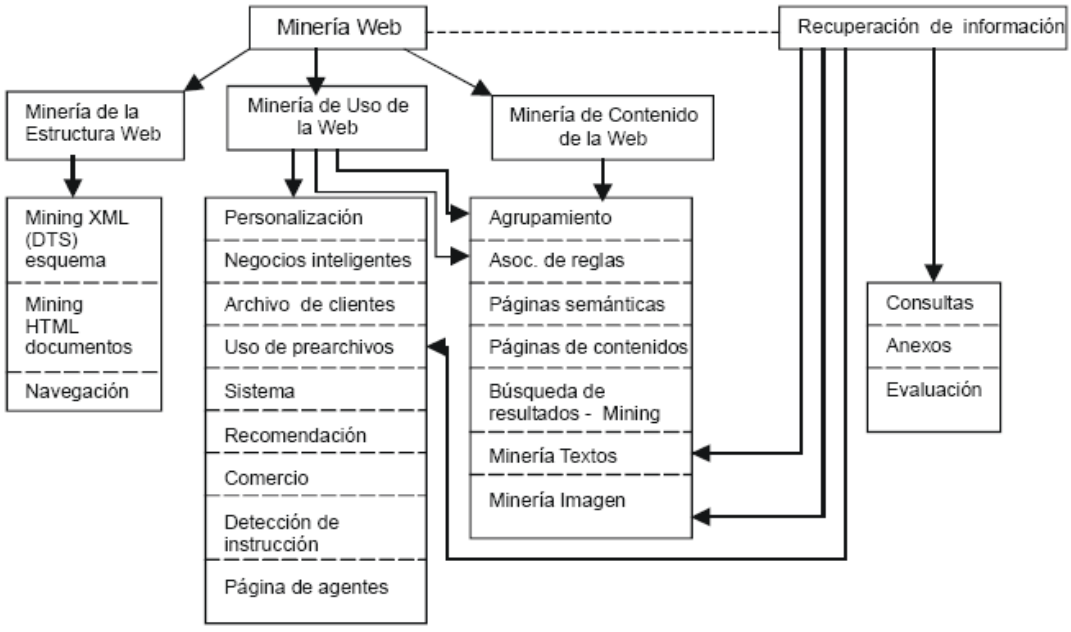
La minería de uso web tiene dos objetivos principales:

- Obtener información de los perfiles de los usuarios partiendo del comportamiento y navegación de los mismos. Para que posteriormente se pueda ofrecer una atención más personalizada.

- Extraer los patrones más generales de uso de un sitio de forma que este pueda reestructurarse. Para que el sitio sea más fácil de usar y que a la vez mejore el acceso por parte de un usuario.

1.6.3 Taxonomía de la minería web

Figura 2. Taxonomía de la minería web



Fuente: *http://sisbib.unmsm.edu.pe/BibVirtualData/publicaciones/risi/n3_2005/a01.pdf*

1.7 Aplicación de la minería web

Muchas empresas u organizaciones invierten cantidades monetarias elevadas para la construcción de un sitio web. De igual forma invierten tiempo y esfuerzos. A pesar de todo pocas organizaciones se preocupan por el proceso subsiguiente que abarca la gestión, mejora, mantenimiento y explotación del sitio.

Los sitios en la web necesitan conocer y aprender cada día sobre sus clientes o usuarios. Ya que solo de esta manera podrán brindar mejoras en los servicios y en el rendimiento del sitio web.

1.7.1 Detección de patrones de acceso

Este tipo de aplicación consiste en la detección que se realiza sobre los ficheros log para poder comprender el perfil de los usuarios y sus respectivas tendencias. Con esto se es capaz de personalizar la interacción con el usuario.

Los procesos de descubrimiento de patrones de actividad y de comportamiento guardan relación con la navegación web. Ya que se necesita de la ayuda de técnicas y algoritmos de la minería de datos que sean capaces de encontrar patrones secuenciales en los ficheros log.

1.7.2 Personalización de servicios

Este tipo de aplicación analiza las tendencias de cada usuario y modifica dinámicamente la información que se le presenta. Las páginas mostradas y el formato de los recursos se adaptan para cada visitante dependiendo de su patrón de acceso.

La personalización de la información es la característica más importante para un usuario. Ya que si un usuario al final del uso del sitio queda satisfecho es casi seguro que volverá a usar el sitio web.

1.7.3 Sitios web

Algunos sitios en la red donde la aplicación de la minería web tiene una gran relevancia son:

1.7.3.1 Sitios de venta

Los usuarios o clientes potenciales deberían poder realizar las siguientes operaciones sin ningún problema.

- Comprar productos.
- Encontrar lo que buscan.
- Hallar los productos más vendidos.

1.7.3.2 Sitios de comercio electrónico

Los clientes deberían realizar las siguientes operaciones sin complicaciones.

- Explorar el centro.
- Realizar compras impulsivas.
- Hallar productos complementarios.

1.7.3.3 Sitios promotores

Los futuros y los eventuales compradores deberían realizar sin problemas la operación siguiente.

- Obtener anuncios apropiados.

1.7.3.4 Sitios proveedores de servicios públicos

Los usuarios de estos servicios deberían realizar las siguientes operaciones sin complicación alguna.

- Encontrar el servicio que buscan.
- Encontrar servicios complementarios.
- Tener acceso a los servicios en un tiempo relativamente corto.

1.8 Estadísticas web

La web se diferencia de otros medios de distribución de información por dos características principales.

- **La web es anónima**

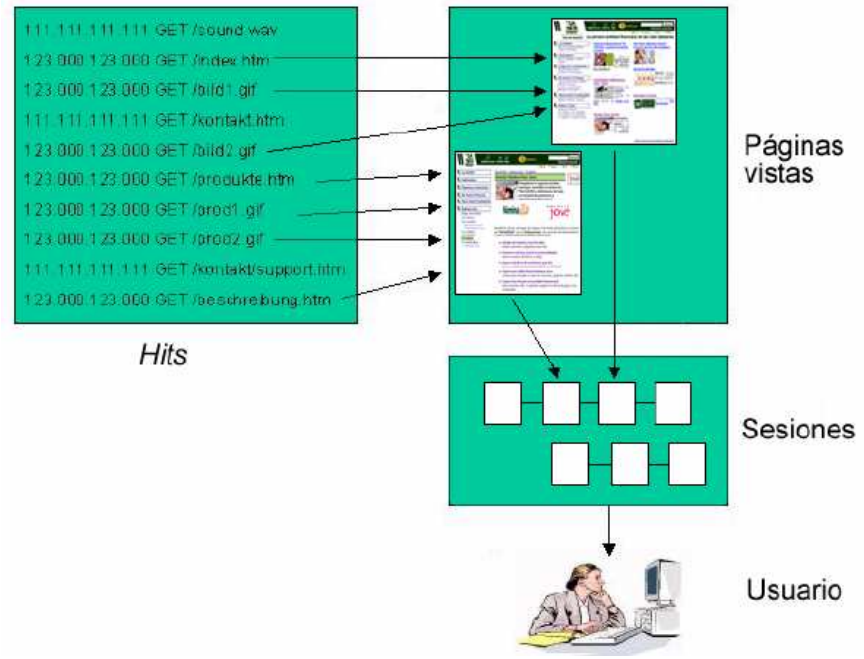
Esta característica de la web no permite a una empresa en un principio conocer información sobre sus usuarios, debido a que la gran mayoría son anónimos. Es decir que estos entran y salen de un sitio sin más detalles que los que se almacenan en el servidor del sitio.

- **La web es interactiva**

Regularmente un sitio web no solo transmite información sino que es algo recíproco. Al estudio de esta información se le conoce como “análisis del tráfico web”.

El análisis estadístico de un sitio web está formado por componentes que funcionan interrelacionados con la finalidad de recolectar datos para su posterior análisis.

Figura 3. Componentes de un análisis estadístico



Fuente: *http://jordisan.net/pfc/Memoria_PFC_web_mining_Jordi.pdf*

1.8.1 Petición o Hits

La petición consiste en la solicitud de un recurso al servidor web. El cual recibe un hit o acceso que a su vez lo guarda como una nueva línea en el fichero log. Al mismo tiempo devolverá un código de respuesta a la petición solicitada.

1.8.2 Página vista o página visitada

Es una página web solicitada por los usuarios que se encuentra formada por un conjunto de ficheros (código HTML, java scripts, imágenes, tablas, etcétera). Las páginas visitadas sirven de referencia para medir la popularidad del contenido de un sitio web.

1.8.3 Visita o sesión de usuario

Una visita o sesión de usuario se encuentra conformada por el conjunto de páginas accedidas por un usuario durante una misma y única sesión de trabajo. El estudio de las visitas proporciona información del comportamiento de los usuarios dentro del sitio web.

1.8.4 Usuario

Un usuario es la persona física que accede al servidor web durante el período de tiempo analizado.

1.9 Log de servidor web

Los logs de servidores web permiten realizar el estudio de un sitio web sin intervenir en la herramienta de publicación de contenidos. A continuación se presenta la información que se almacena en un fichero log.

Tabla I. Estructura de un fichero log

Campos de log
Dirección IP del cliente
Dirección del servidor
Nombre de usuario
Fecha y hora
Petición / Recurso
Código de respuesta
Tamaño de la respuesta
Remitente
User agent

Fuente: *http://jordisan.net/pfc/Memoria_PFC_web_mining_Jordi.pdf*

1.9.1 Dirección IP del cliente

Es la dirección IP del ordenador por medio de la cual el cliente accede al servidor. En varias ocasiones esta dirección puede ser la de una maquina que actué como intermediaria (firewall, proxy). Dificultando así la verdadera identidad del usuario.

1.9.2 Dirección del servidor

Es la dirección que el servidor web utiliza para que se pueda acceder a los recursos del mismo. Es útil cuando se ejecutan varios servidores virtuales, ya que permite separar cada uno de los accesos de cada servidor.

1.9.3 Nombre de usuario

Este campo se almacena en el fichero log solo sí el usuario se encuentra registrado dentro del servidor web. En caso contrario permanece vacío.

1.9.4 Fecha y hora

En este campo se almacena la fecha y hora del servidor al momento de realizar una petición.

1.9.5 Petición/Recurso

Es la petición solicitada por el usuario. Se incluye el método HTTP empleado (get, post, etcétera), el recurso solicitado (página, video, audio, etcétera) y la versión del protocolo HTTP empleada.

1.9.6 Código de respuesta

Este campo almacena el código HTTP de respuesta proporcionado por el servidor web que define el resultado de la petición.

1.9.7 Tamaño de la respuesta

Este campo almacena el tamaño en bytes del mensaje enviado por el servidor web como respuesta a la petición.

1.9.8 Remitente

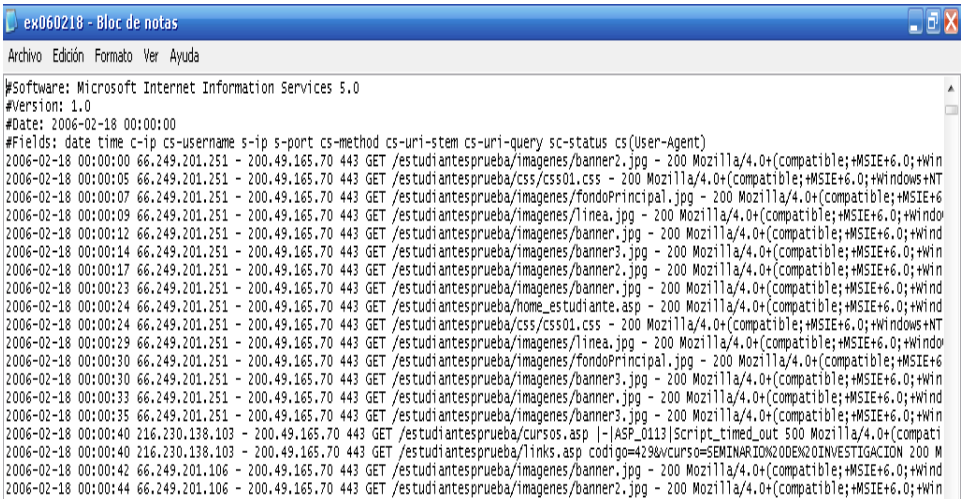
Es un campo opcional que contiene la URL del recurso origen de donde se realizó la petición. Este campo es útil para ver la secuencia de páginas visitadas por un usuario.

1.9.9 User agent

Este campo es una cadena de texto opcional que contiene la identificación de la plataforma empleada por el usuario (tipo de navegador, versión del navegador, nombre del sistema operativo, etcétera).

1.9.10 Ejemplo de un fichero log

Figura 4. Ejemplo de un fichero log



1.10 Contenidos de la minería web

1.10.1 Minería web de contenidos estáticos

Dentro del análisis de ficheros log las páginas accedidas están identificadas por las URLs de entrada. En un servidor de contenidos estáticos, las direcciones URL tienen la siguiente estructura básica:

http://www.dominio.com/path/fichero.htm

Para estos contenidos las herramientas de estudio de logs se basan en ciertos supuestos como:

- Existencia de relaciones biunívocas entre las páginas y las URLs.
- Las rutas de las URLs muestran la localización de las páginas en la estructura jerárquica del contenido.

A partir del desarrollo de las suposiciones antes mencionadas, es posible extraer conocimiento y elaborar reestructuras dentro del sitio web.

1.10.2 Minería web de contenidos dinámicos

En un servidor de contenidos dinámicos, las direcciones URL tienen un formato más variable que depende no solo del lenguaje y de la plataforma empleada sino de la implementación del sitio web. Un ejemplo puede ser:

http://www.dominio.com/productos.jsp?id=53

En los contenidos dinámicos se puede dar que una página pueda accederse empleando URLs distintas o que una misma URL muestre contenidos distintos en función de parámetros. Por ejemplo parámetros de códigos de usuario.

Esta característica en los logs de los sitios dinámicos limita un poco a las herramientas de análisis. Ya que estas no resuelven de una forma correcta la correspondencia entre URLs y las páginas con su contenido.

Sin embargo, existen productos más avanzados que permiten asociar el contenido del sitio con las respectivas URLs proporcionando datos adicionales. Estos productos ayudan a las herramientas de análisis a la identificación y agrupación de páginas generadas de modo dinámico.

2 PROCESO DE MINERÍA DE USO WEB

2.1 Funcionamiento de la minería de uso web

Para llevar a cabo un proyecto de minería de uso de la web se hace necesario establecer y seguir un proceso previamente definido por medio del cual se logre obtener resultados de utilidad.

El proceso general de la minería de uso web contiene las siguientes etapas: recolección de los datos, preparación de los datos, transformación de los datos, descubrimiento de patrones y análisis de patrones.

- La etapa de recolección de datos se puede llevar a cabo a nivel del servidor web, a nivel del servidor proxy y a nivel de los agentes del cliente. En ocasiones para obtener los datos actualizados se necesita de la sincronización de estos tres niveles.

- La etapa de preparación de los datos se divide en varias tareas: establecimiento de objetivos, limpieza de datos, identificación de usuarios, identificación de sesiones, identificación de caminos, identificación de transacciones, formateo y elaboración de inferencias.

- La etapa de transformación de datos sirve para crear estructuras de datos, tablas de datos y grafos del sitio. Estos son requeridos antes de iniciar el descubrimiento de patrones.

- La etapa de descubrimiento de patrones se lleva a cabo empleando diferentes algoritmos como: análisis estadístico, agrupación,

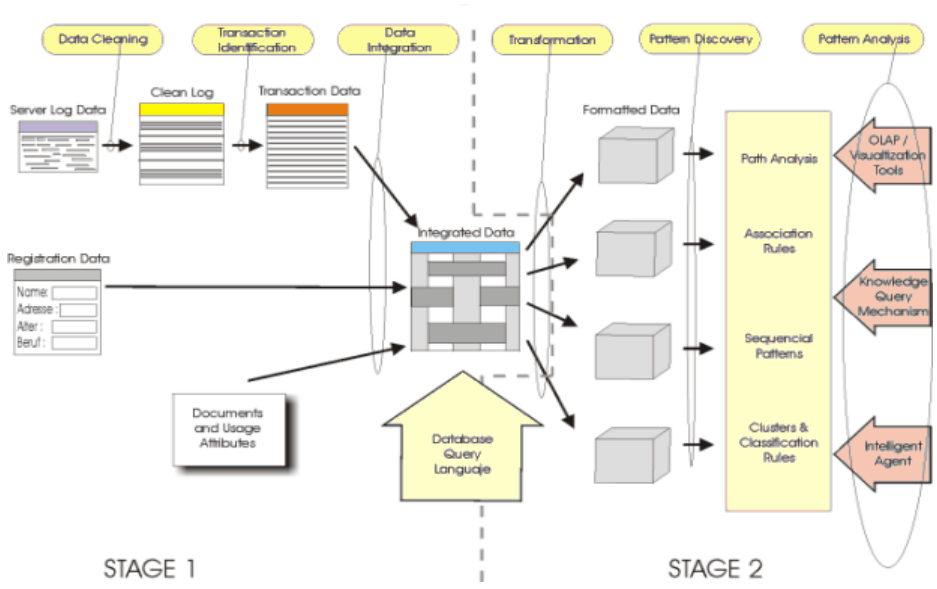
clasificación, reglas de asociación, patrones secuenciales y análisis de caminos.

- En la etapa de análisis de patrones se descubre información oculta, se muestran los resultados de las acciones realizadas por los usuarios, se identifican tendencias de los usuarios y se muestran las preferencias de los usuarios.

  Para poder realizar el análisis de patrones se emplean técnicas de visualización gráfica, herramientas OLAP, técnicas de filtración de datos y procesos KDD entre otras. Los resultados del análisis deben presentar el comportamiento del usuario con la finalidad de mejorar los sistemas, personalizar y reestructurar los sitios web.

A continuación se presenta un diagrama del proceso general de la minería de uso web.

Figura 5. Proceso de minería de uso web



Fuente: *http://galeas.de/webmining.html*

2.2 Recolección de datos

Durante este proceso se recopilan todos aquellos datos que formaran parte del análisis de minería web. En los ficheros log del servidor web se encuentran datos de los usuarios, datos demográficos, datos de marketing y datos del sitio web entre otros.

Aparte de los ficheros log del sitio web existen otras fuentes de información que sirven para realizar la recolección de datos. Por mencionar algunos estan los siguientes: bitácoras de acceso, bases de datos de usuarios, ontologías del sitio, agentes remotos y locales, bitácoras de referencia, información semántica, atributos de los productos y cualquier otra fuente de información que se considere relevante durante el estudio.

El proceso de recolección de datos se puede llevar a cabo a nivel del servidor proxy, a nivel del servidor web y a nivel de los agentes del cliente. Muchas veces para obtener datos que estén actualizados se necesita de la sincronización de estos tres niveles.

2.3 Pre procesamiento de los datos

Durante el proceso de minería de uso web se considera un pre procesamiento de los datos. Este proceso consiste en convertir la información disponible (archivos log) en información manejable. Para esto se usan diversos tipos de abstracciones y filtraciones de información que se considere redundante ó despreciable.

Durante el pre procesamiento se manejan diversos tipos de información pertenecientes al sitio web como los son: información por parte de usuarios,

información del contenido del sitio e información de la estructura de dicho contenido.

A continuación se presenta una lista de los pasos que abarca el proceso de pre procesamiento.

2.3.1 Eliminar robots de acceso web

Regularmente los ficheros logs tienen información acerca de robots de acceso web tales como spiders, índices y crawlers entre otros. Estos datos son eliminados, ya que no representan información relevante durante el proceso de minería de uso web. Los robots de acceso web son creados originalmente por los administradores de los sitios web con la finalidad de generar y brindar permisos de acceso dentro del sitio.

La información de los robots de acceso web generalmente se puede encontrar dentro de archivos con extensión “txt”. En otras ocasiones los robots de acceso web se pueden identificar fácilmente a partir de la dirección IP y del host name del usuario.

2.3.2 Filtrar datos multimedia

Cada sitio web contiene diversas páginas web. Estas páginas a su vez almacenan junto con la información algunos datos multimedia tales como sonido, imagen y video. Como el servidor web registra todas las peticiones, esté también registra todas las entradas y salidas solicitadas de datos multimedia. Entonces para llevar a cabo un análisis de minería web todas estas entradas multimedia son filtradas de los ficheros logs.

Cada una de estas entradas tiene su propia ruta de acceso. Está ruta puede proporcionar datos acerca de los usuarios ó del sitio web. Para cada acción que se registra en el fichero log hay un código que lo identifica. El encargado en filtrar estas entradas es quien decide cuales eliminar y cuáles no eliminar.

2.3.3 Extraer transacciones

Cuando se termina el proceso de filtración y eliminación de aquella información que se considera despreciable de los ficheros logs, lo que sigue es la identificación y extracción de transacciones referentes a los usuarios. Una transacción es una entrada u operación registrada en el servidor web. Dicha transacción es efectuada por un visitante en un período de tiempo.

Al período de tiempo empleado por el visitante para la navegación dentro del sitio con previo registro se le llama sesión. Con la aplicación de los servidores proxy en un sistema se puede establecer el límite de tiempo de una sesión en cada visita.

Dentro de los ficheros logs existe un registro llamado “Referrer”. Este registro almacena las URLs de las páginas antes vistas por el visitante a partir de las cuales se siguió el enlace. Un ejemplo sencillo de esté registro es que si un usuario se encuentra en una página llamada /productos/ y por medio de esta se manda a llamar a una nueva página llamada /productos/oferta/ es entonces cuando se crea una nueva entrada en el registro Referrer con el nombre de la página referida, en este caso /productos/.

Otra manera de encontrar e identificar transacciones es por medio de lo que se conoce como longitud de la referencia. Este método supone que los usuarios gastan mayor cantidad de tiempo en páginas de contenido y por el contrario

gastan menor cantidad de tiempo en páginas de solo navegación. La extracción de transacciones se lleva a cabo por medio de las referencias delanteras máximas.

Las referencias delanteras máximas son aquellas páginas que son accedidas antes de retornar a una página vista previamente. Un ejemplo para entender este concepto es la suposición de que un usuario durante una sesión visita las siguientes páginas S-T-U-V-S-W-X-U-Y-Z-S es entonces que las páginas V, X y Z son las referencias delanteras máximas.

2.3.3.1 Identificación de usuarios

Durante este proceso el dato principal es la dirección IP. Ya que ésta identifica al usuario que realiza una petición. Pero en muchas ocasiones este dato no es suficiente, ya que diversos usuarios pueden figurar con la misma dirección IP, esto debido a mecanismos intermedios como firewalls ó servidores proxies.

Otra forma consiste en emplear la cadena de identificación enviada por el navegador, la cual contiene información acerca del software usado por el programa cliente.

2.3.3.2 Identificación de sesiones

Este proceso consiste en separar los accesos de un usuario en sesiones de navegación independientes. Agrupando los accesos pertenecientes a una misma sesión y que han sido de manera ininterrumpida por el usuario. Por lo regular los usuarios en la red visitan diversos sitios web. Además pueden hacer varias visitas al mismo sitio web en el transcurso del día. Así mismo un usuario puede terminar su sesión y otro usuario diferente puede comenzar su sesión en el mismo sitio web desde la misma computadora.

El método más sencillo para realizar esta identificación es el tiempo límite o umbral (timeout) que consiste en medir el tiempo transcurrido entre dos peticiones consecutivas del usuario. Esto se hace para saber si se ha iniciado una nueva sesión. Una aproximación empírica de este tiempo se considera en 30 minutos.

2.3.4 Extraer características

Durante este proceso se trata de identificar las principales características de las transacciones previamente seleccionadas. Así mismo se busca disminuir la dimensión en los datos desechando aquellas características con menor relevancia y que se consideran inaplicables en el proceso de selección. En la extracción de características el objetivo es transformar aquellas transacciones que tienen una longitud variable en cadenas de longitud estática de la característica en cuestión.

Cuando se logra tener cadenas de una longitud “n” para cada característica, estas cadenas se transforman en un formato que necesita la herramienta para preparar la información (datos). Los visitantes de los sitios web pasan un lapso de tiempo mayor en todas aquellas páginas que son de su interés. La velocidad de conexión en la red y la extensión de cada página son dos variables que determinan el tiempo estimado de navegación entre páginas web. Un problema que surge al momento de querer obtener la situación exacta de acceso al sitio se da por los servidores proxies y por los navegadores web.

Por un lado, los servidores proxies almacenan aquellas páginas que tienen una mayor demanda en la intranet con la finalidad de disminuir el tráfico de la red y evitar la sobrecarga del servidor. Por otra parte los navegadores web registran las páginas visitadas junto con el número de veces que se han solicitado dichas páginas. La solución ante dicho problema se puede resolver a través de agentes remotos o bien a través del uso de cookies.

2.4 Procesamiento de los datos

En el transcurso de análisis de minería de uso web se considera la etapa de procesamiento y preparación de los datos. La finalidad de este proceso es generar datos que sean de alta calidad. Los cuales conduzcan a descubrir patrones de comportamiento o a inferir tendencias.

Partiendo de los datos de los ficheros logs pre procesados se necesita encontrar las inconsistencias y resolver las mismas. La integración y preparación de los datos se lleva a cabo por medio de herramientas que pueden trabajar del lado del servidor web o bien del lado del cliente. El fin de la preparación de los datos es brindar datos que se puedan usar de entrada en la etapa de descubrimiento de patrones.

Durante el procesamiento de los datos se evalúan los mismos para poder determinar de qué tipo son, para que pueden llegar a servir y que beneficios se pueden obtener a partir de ellos. Es así que en la etapa de procesamiento existen sub etapas que permiten integrar técnicas de análisis de datos con el objetivo de mejorar la calidad de los datos.

A continuación se presentan las sub etapas que abarca la etapa de procesamiento.

2.4.1 Establecimiento de objetivos

Esta tarea consiste en el establecimiento de los objetivos que se pretenden alcanzar, vistos desde el punto de la lógica del negocio. Simultáneamente se definen las estrategias de validación relacionadas con los objetivos descritos previamente.

2.4.2 Limpieza de datos

En esta tarea se trata de eliminar de los ficheros log del servidor web todos aquellos accesos que se consideren irrelevantes. Todos estos accesos a descartar dependerán del objetivo primario del sitio web en concreto.

2.4.3 Completación de caminos

El fin de esta tarea es detectar la navegación del usuario a través de aquellas páginas que no quedan registradas en los ficheros log debido a la memoria de cache de los sitios web. Los inconvenientes de esta son la complejidad junto con la poca fiabilidad de la información.

2.4.4 Formateo

Se considera como el último paso del procesamiento y consiste en dejar los datos preparados en un formato que sea de utilidad para una técnica de minería empleada en el descubrimiento de patrones dentro de la minería de uso web. Terminada la aplicación de este proceso en los ficheros logs se puede garantizar que se tienen datos limpios y de calidad.

2.5 Descubrimiento de patrones

Para comenzar este proceso se deben de tener previamente identificados, procesados y disponibles todos los datos de los ficheros logs a utilizar en el análisis de minería web. Con estos ficheros logs transformados ya es posible comenzar a buscar patrones de acceso y de comportamiento relacionados con los usuarios en la web. Para encontrar patrones en los datos la minería web hace uso de técnicas conocidas de la minería de datos por ejemplo: análisis estadístico, técnicas de agrupación, clasificación y reglas de asociación entre otras.

Las técnicas antes mencionadas emplean para su funcionamiento la información registrada en los ficheros log. Además hacen uso del comportamiento de los usuarios por medio de los patrones de navegación, los perfiles de usuario, las transacciones realizadas y la localización geográfica entre algunos otros. Dependiendo de la situación del problema que se trate de resolver y de los datos que se dispongan unas técnicas serán más adecuadas que otras.

A continuación se presenta las técnicas más empleadas al realizar un análisis de minería de uso web.

2.5.1 Análisis estadístico

El análisis estadístico es la forma más común de extraer información de un sitio web. Este análisis permite obtener valores estadísticos de variables como tiempos por sesión, páginas consultadas, productos buscados, descargas populares, etcétera. Esta técnica proporciona informes poco elaborados pero que son de gran utilidad al momento de analizar el rendimiento, la disponibilidad, la usabilidad y la detección de errores entre otras características del sitio web. El análisis estadístico trabaja con una porción relativa de datos que sirven de

muestra para estudiar las relaciones existentes entre ellos y a la vez permite encontrar errores poco visibles.

El análisis estadístico hace uso de dos técnicas para analizar un sitio web, estas técnicas son: las técnicas basadas en inferencias y las técnicas descriptivas. Por un lado las técnicas basadas en inferencias realizan suposiciones acerca de la información oculta o que no se conoce a fondo. Estos supuestos se llevan a cabo empleando el teorema de Bayes o bien estableciendo probabilidades de frecuencias. Por otra parte las técnicas descriptivas se usan con mayor frecuencia para resumir datos e información empleando cálculos como: la media, la varianza, la desviación estándar, la moda, la frecuencia relativa, la frecuencia absoluta, promedios, correlaciones, regresiones, la covarianza, etcétera.

También existen algoritmos estadísticos entre los cuales se puede mencionar el algoritmo *PageGather* y el algoritmo de los *k-medios* entre otros. El algoritmo de *k-medios* efectúa un análisis estadístico sobre los datos con el objetivo de realizar agrupaciones en los datos y así poder medir sus distancias; otra función de este algoritmo es que efectúa cálculos medios y promedios de los datos empleados de la muestra parcialmente tomada.

Un problema del análisis estadístico se da cuando el tamaño de las características posibles es demasiado grande en comparación al tamaño tomado de la muestra. De igual forma otro problema del análisis estadístico ó de los métodos probabilísticos es que en su mayoría estas técnicas asumen una total independencia de atributos que puede llegar a convertirse en una suposición demasiado restrictiva en ciertos dominios. Debido a estos problemas el análisis estadístico se queda muy atrás en comparación a otras técnicas que son capaces de encontrar asociaciones entre los datos y que funcionan con tamaños de dimensiones mayores.

2.5.2 Clasificación

Las técnicas de clasificación permiten elaborar un perfil para usuarios que acceden a ficheros específicos del servidor en función de sus patrones de acceso. La técnica de clasificación pretende encontrar todos aquellos patrones de navegación por parte de los usuarios, con el fin de poder crear categorías de clasificación conforme su comportamiento regular.

La clasificación prácticamente permite crear perfiles de los usuarios que pertenecen a cierto grupo particular. El perfil creado posteriormente se puede emplear para agregar valor de información en la base de datos. Los perfiles creados se pueden basar en los patrones de acceso por parte de los usuarios a ciertas páginas web o archivos de cualquier tipo con alguna información demográfica en su contenido.

Toda la información para elaborar un perfil se obtiene a partir del comportamiento del usuario, de los requerimientos del cliente y de toda aquella información enviada por medio de los navegadores web. Otra forma de obtener información es por medio de formularios on-line, por medio de suscripciones y por medio de registraciones en sitios web.

Para poder construir perfiles de clasificación se hace uso de técnicas tales como: sistemas de lógica difusa, arboles de decisión, algoritmos genéticos, redes neuronales, modelos vectoriales y teorema de Bayes. Por medio de las técnicas de clasificación se puede llegar a generar resultados con un gran valor dentro del sitio web, por ejemplo:

- El 20 % de los clientes que realizan una compra en línea a través de la página /Productos/mayoreo/losmasofertados.jsp se encuentran en

un rango de edad entre 40 – 48 años y además su ubicación demográfica esta en el área metropolitana.

- El 35 % de los usuarios que pasan un promedio de 2 horas durante una sola sesión son mujeres por arriba de los 30 años.
- El 10 % de las compras que se realizan los fines de semana en horas de la tarde se hacen desde el área occidental.

La clasificación puede facilitar el desarrollo de estrategias de mercado futuras tanto on-line como off-line; por ejemplo envío de correos automáticos a clientes de un grupo, encuestas a un sector especifico de usuarios, personalización del sitio web conforme el tipo de cliente, etcétera. En ciertos casos la información que se extrae se puede combinar con los datos de la organización con la finalidad de poder identificar a un usuario en diferentes sesiones y aún así poder registrar todas sus transacciones.

2.5.3 Reglas de asociación

Las reglas de asociación se aplicaron inicialmente dentro de las bases de datos relacionales transaccionales en donde una transacción representa un conjunto de ítems ó procesos. Dentro de este esquema la problemática consiste en encontrar las relaciones entre ítems de datos. Donde un número de ítems en una transacción significa el surgimiento de nuevos ítems. Ubicados dentro del entorno de minería web las reglas de asociación se encargan de encontrar correlaciones existentes entre las páginas vistas por un usuario en un sitio web.

En otras palabras, las reglas de asociación tratan de descubrir las relaciones entre cada página visitada durante una misma sesión por un usuario. Basándose por medio de los patrones de ocurrencia en cada transacción. Cada transacción u operación registrada de un usuario se compone de un conjunto de direcciones URL las cuales fueron accedidas por medio del servidor. Con la información

obtenida de las reglas de asociación se puede definir de mejor manera la estructura del sitio web. Además se puede anticipar que páginas web visitara un usuario con mayor probabilidad y así agilizar el proceso de carga de cada página.

Por medio de las reglas de asociación se puede llegar a encontrar y descubrir relaciones que generen valor dentro del sitio web, algunos ejemplos son:

- El 23 % de los clientes que realizaron una compra en línea a través de la página /Productos/verano/rebajas/productoX.asp también accedieron a la página /Productos/verano/rebajas/productoY.asp aunque no compraron este último.
- El 10 % de los clientes que visitaron la página /Accesorios/navideños/juguetes.xml visito seguidamente la página /Accesorios/navideños/cocina.xml y finalmente visitaron la página /Accesorios/navideños/adornoscasa.xml.
- El 30 % de los clientes que compra semanalmente en la página /Comida/boquitas.php también compra en la página /Bebidas/licor.php y visita la página /Video/películas/estreno.php.

Las reglas de asociación también toman en cuenta un soporte para las reglas encontradas. Un soporte es una medición que se basa en el número registrado de ocurrencias dentro de una transacción. Con las reglas de asociación se puede llegar a descubrir las relaciones existentes sin necesidad de que intervenga un operador. El estudio de las reglas de asociación se enfoca generalmente en dos partes:

- La extracción de ítems que cumplan con la cobertura requerida, a partir de
- los datos.
- La generación de las inferencias, a partir de los documentos disponibles.

La temprana aplicación de las reglas de asociación en un entorno de comercio electrónico puede ser de gran ventaja competitiva para el desarrollo de estrategias de marketing dentro del sitio web. Los sitios web comúnmente se encuentran organizados de forma jerárquica. Es aquí que las reglas de asociación también ayudan a reestructurar de mejor forma la organización de un sitio web.

Por ejemplo si se obtiene el dato de que un 90 % de los clientes que acceden desde la página principal de un sitio web a la página A y que posteriormente pasan de la página A a la página B se puede inferir de que existe algún tipo de información en la página A que hace que los clientes vayan luego a la página B. Por lo cual esta relación da la pauta de que se debería cambiar la información de la página A á la pagina principal con el objetivo de aumentar el acceso de la página B.

Las medidas para conocer la calidad de una regla de asociación son dos: la cobertura ó soporte y la confianza ó precisión. Por un lado la cobertura/soporte se puede definir como el número de veces que la regla de asociación predice de manera correcta. Por otra parte la confianza/precisión se encarga de medir el porcentaje de veces que la regla de asociación se cumple cuando se lleva a cabo.

Dentro de un sitio web por ejemplo si se tiene un soporte llamado /Peliculas/dvd/estrenos que es de baja incidencia automáticamente se puede decir que las búsquedas por reglas de asociación en las páginas web descendientes jerárquicamente del soporte en cuestión no tendrán un soporte necesario válido.

Una forma de inferir y denotar las reglas de asociación se pueden dar de la siguiente manera: si P y Q son un conjunto de características del conjunto de datos se puede expresar una regla como P => Q que significa que cualquier operación que contiene a P también contiene a Q.

Generalmente las reglas de asociación emplean y hacen uso del algoritmo "*Apriori*" el cual fue desarrollado por los autores Srikant y Agrawal en el año de 1994. Este algoritmo de aprendizaje de reglas de asociación se basa en la búsqueda de grupos de elementos con un específico soporte. Este algoritmo se desarrollo para reglas de minería de datos donde exista una numerosa cantidad de transacciones sobre las bases de datos.

El algoritmo Apriori separa el problema en dos partes: primeramente se encuentran todas las combinaciones de elementos que tengan validez de transacción. Las combinaciones encontradas se llaman frecuencias de aparición del conjunto de elementos. Luego se emplean las frecuencias halladas con el objetivo de generar las reglas de asociación.

2.5.4 Patrones secuenciales

La técnica de patrones secuenciales permite determinar el tiempo de las secuencias ordenadas de todas las URLs que los usuarios han visitado y así poder predecir las mismas en un tiempo futuro con más certeza. El descubrimiento de patrones de secuencia en los ficheros log sirve para intuir futuras visitas con lo que se mejoran los servicios y las personalizaciones durante ciertos intervalos de tiempo.

En los ficheros históricos del servidor web quedan registradas las visitas de los usuarios durante un período de tiempo relacionado. En los archivos log de transacciones quedan almacenados los datos de fechas y horas en que un respectivo usuario efectúo alguna operación. A través del análisis de estos conjuntos de datos es que se descubre el comportamiento de un usuario respecto a la línea del tiempo en un sitio web.

Además los patrones secuenciales permiten a las organizaciones ofrecer publicidad y propaganda a diversos usuarios en relación con los patrones hallados. Por medio de los patrones secuenciales se puede llegar a determinar las relaciones entre ítems de datos del sitio web, algunos ejemplos son:

- El 53 % de los clientes que realizan una compra en línea por medio de la página /Productos/ofertas/productoZ.aspx también realizan una compra en la página /Productos/ofertas/productoP.aspx en un período no mayor de 10 días.
- El 40 % de los clientes que accedieron al sitio /Productos/ofertas/ lo hicieron por medio de una consulta en el buscador Google el día anterior al acceso con las palabras claves: oferta de productos.
- El 60 % de los usuarios de la cuenta de correo gmail revisan su bandeja de entrada por lo menos 12 veces en un período de tiempo de 24 horas.

Una dependencia de datos es cada una de las secuencias de transacciones registradas en un lapso de tiempo. Es decir que se podría estar interesado en hallar aquellas características comunes entre los usuarios que visitan una página X en particular en un período de tiempo dado $[t_o, t_1]$. Por otro lado, se podría estar también interesado en un período de tiempo específico (día, semana, mes, año, etcétera) con la finalidad de saber cuál fue la página web o el fichero más visitado.

Los patrones secuenciales se pueden emplear para encontrar también tendencias, características del sitio, secuencias de eventos, etcétera. Con la información resultante se puede mejorar los campos comerciales y técnicos dentro del sitio web.

En concreto los patrones secuenciales se centran en el análisis de secuencias de tiempo. Ya que cada una de las transacciones registradas en el

servidor web tiene una estrecha relación con el período de tiempo en que se produjo.

Por ejemplo, si se sabe que los fines de semana entre las 11:00 a.m. y las 14:00 p.m. varias personas ordenan comida en línea. Entonces se debe facilitar el acceso en cada uno de los sitios web que sean de comida en este lapso de tiempo. Además ofrecer promociones para llamar la atención del consumidor.

2.5.5 Agrupación o clustering

Las técnicas de agrupamiento también se conocen como *clustering.* Consisten en identificar patrones de comportamiento similar dentro de grupos homogéneos. Es decir que ítems con características similares pertenecen a un mismo grupo y las características de un grupo serán diferentes a las de otro grupo.

En otras palabras, el clustering consiste en las formaciones automáticas de grupos o bloques de datos con características parecidas sin contar con una previa clasificación. Una vez descubiertos los perfiles de cada grupo se pueden usar las características de los mismos para implementar técnicas de clasificación apropiadamente. Como ya se había detallado anteriormente, las técnicas de clasificación permiten elaborar un perfil para usuarios que acceden a ficheros específicos del servidor en función de sus patrones de acceso.

El empleo y uso de técnicas de clustering para analizar ficheros log en un servidor web pueden brindar conocimiento útil del comportamiento de los usuarios. Con lo cual se pueden identificar diferentes niveles de grupos de clientes y por tanto implementar estrategias de mercadeo en base a los sectores identificados.

Con base al análisis de los datos registrados en los ficheros logs se puede llegar a identificar grupos de clientes. Algunos ejemplos son:

- Grupo de clientes que solo realizan compras de abarrotes los fines de semana.
- Grupos de clientes que visitan una cantidad grande de páginas con un lapso de tiempo parecido entre cada página.
- Grupos de usuarios que pasan períodos de tiempo largos dentro del sitio web en una sola sesión.

- Grupos de usuarios que comparten las mismas preferencias en las compras en línea.
- Grupos de usuarios que comparten el mismo gusto decorativo en la personalización de su perfil en una red social.

Regularmente el clustering emplea un previo conocimiento acerca de la organización de los archivos con el objetivo de establecer la distancia entre cada archivo. También el clustering hace uso de métodos probabilísticos para calcular dicha distancia. Las técnicas de clustering basadas en distancia emplean como dimensión un grupo de palabras que se encuentran en diversos archivos. Cada grupo o conjunto representado como vector identifica al archivo y se le puede ver como un punto de partida dentro del espacio dimensional.

Un problema que sucede a menudo en la web es que el tamaño de la distancia es muy grande en comparación al tamaño de los archivos en cuestión. Este problema hace que los algoritmos tradicionales de clustering no funcionen correctamente bajo estas condiciones. Una solución en los algoritmos de clustering para la función de distancia es medir la frecuencia de aparición de cada palabra. Más sin embargo esta solución no es del todo apropiada ya que en la web no todos los documentos tienen el mismo tamaño y por tal razón algunas palabras pueden tener mayor frecuencia de aparición que otras.

Aparte de esto los esquemas que se basan en las distancias necesitan calcular los promedios de los clúster. Si la medida arroja un valor alto no se difiere tanto de un clúster y de otro. Pero por el contrario si la medida es baja entonces el resultado no es nada confiable. Ante esta situación de desconfianza se ve la necesidad de crear y generar algoritmos de clustering más sofisticados que no requieran de un conocimiento previo para poder encontrar una función de distancia. Estos algoritmos deben tener la capacidad de encontrar comparaciones

y crear agrupaciones entre los archivos. Además deben funcionar sin problemas en espacios de gran tamaño y dimensión.

2.5.5.1 Técnicas mejoradas de clustering

2.5.5.1.1 Análisis de componentes principales

Para este algoritmo cada archivo se representa en un vector normalizado que contiene las frecuencias de cada palabra. Luego el algoritmo suprime los espacios de los archivos por medio de un hiperplano. Este hiperplano atraviesa por la mitad todos los vectores. El hiperplano es perpendicular a la varianza límite del conjunto de archivos.

Los archivos se separan en dos bloques divididos por el hiperplano. A cada grupo identificado nuevamente se le aplican los pasos anteriores la cantidad de veces que se considere necesario. Al final del procedimiento se pretende obtener una estructura jerárquica en forma de árbol. Donde las hojas del mismo son los clúster identificados.

2.5.5.1.2 Clustering sintáctico

En esta técnica se puede definir una medida de comparación y una medida de inclusión entre los archivos. Para declarar estas medidas cada archivo se transforma en una secuencia de solicitudes (tokens). La secuencia contiene únicamente las palabras del archivo suprimiendo los comandos html, xml y el formateo de los mismos.

Con esta secuencia limpia de tokens se define el concepto de *shingle;* que es una subsecuencia continua de tamaño "*k*" de palabras. Posteriormente se

declara un conjunto k-shingling [S (D, k)] como el conjunto de todos los shingles de tamaño k.

Un ejemplo de shingle sería que con la secuencia (yo, estudio, en, la, ciudad, de, Guatemala) y el tamaño de la subsecuencia igual a k=3. Se tiene que el conjunto 3-shingle es {(yo, estudio, en), (estudio, en, la), (en, la, ciudad), (la, ciudad, de), (ciudad, de, Guatemala)}.

Si se cuenta con dos archivos independientes llamados X y Y se puede definir el grado de parentesco entre los archivos de la siguiente manera:

$$p(X, Y) = ½S(X)\ S(Y)^{1/2}\ \ ½S(X)\ S(Y)^{1/2}$$

De igual manera se puede definir el grado de contención entre los archivos de la siguiente manera:

$$c(X, Y) = ½S(X)\ S(Y)^{1/2}\ \ ½S(X)^{1/2}$$

Para optimizar el costo transaccional del cálculo surge un nuevo concepto denominado "*sketch*". Este concepto define que a partir de un conjunto S (D, k) se toma un subconjunto más pequeño que contiene bastante información acerca del archivo. De esta manera se puede emplear el sketch en comparaciones de parentesco entre los archivos.

Este algoritmo calcula para cada archivo el conjunto S (D, k) para un número k cualquiera. Seguidamente se calcula el sketch del archivo, luego se igualan los sketchs de cada par de archivos entre ellos. Finalmente si el grado de parentesco o similitud es más grande que el máximo establecido se colocan los archivos en el mismo clúster.

2.5.5.1.3 Reglas de asociación de particionamiento de hipergrafos

En la agrupación o clustering de archivos cada característica representa una transacción u operación y cada archivo representa un elemento. En este caso se emplean reglas de asociación para identificar grupos de archivos que tengan las mismas características. Cada uno de los conjuntos identificados debe satisfacer un soporte respectivo.

Esta técnica inicialmente realiza búsquedas por medio de las reglas de asociación para encontrar un conjunto de elementos que con frecuencia se encuentren juntos. Seguidamente cada conjunto de elementos se usa para concentrar los elementos en un arco de un hipergrafo. Para encontrar los clúster deseados se emplea un algoritmo de particionamiento de hipergrafos.

El hipergrafo HG = (V, I) se encuentra formado por distintos vértices (V), donde los vértices representan los archivos y los hiperarcos (conexiones de más de dos vértices). La variable (I) representa los elementos que son más frecuentes en el archivo. Al hiperarco formado se le asigna un peso que se calcula a través del promedio de confianza de las reglas de asociación que están relacionadas con los archivos del hiperarco. La confianza de cada regla de asociación es la probabilidad existente de que una característica se halle contenida en un archivo. Sabiendo que dicha característica se encuentra en los demás archivos que conforman el hiperarco.

Posteriormente se divide el hipergrafo de modo de que los pesos de los hiperarcos que son particionados sean menores. La principal ventaja de esta técnica es la filtración de archivos irrelevantes durante el proceso de clustering de archivos. Esta filtración se debe a la característica de soporte en las reglas de

asociación empleando un valor máximo de soporte. En otras palabras todos aquellos archivos que tengan un soporte menor al valor máximo son desechados.

2.5.6 Análisis de caminos

Esta técnica proporciona una generación de algunas formas de grafos con el fin de representar las relaciones entre páginas web. Estos grafos pueden ser esquemas físicos donde las páginas son los nodos y los enlaces entre páginas son las flechas entre los nodos.

Pueden existir otros tipos de grafos a partir de las páginas web que creen arcos mostrando el número de usuarios que van de una página a otra ó que creen arcos mostrando la similitud del contenido entre las páginas web. La técnica de análisis de caminos también se puede emplear para encontrar las rutas que tienen una mayor popularidad dentro del sitio web. Algunos ejemplos de la técnica de análisis de caminos son:

- El 50 % de los usuarios que consultaron el sitio web abandonaron el mismo luego de visitar un promedio de dos páginas.
- El 49 % de los usuarios que visitaron la página /Guatemala/lugares/turísticos/peten/tikal.jsp empezaron el recorrido inicialmente en /Guatemala/ luego pasaron por /Guatemala/atracciones/ seguidamente visitaron la página /Guatemala/lugares/turísticos/top10/ y finalmente llegaron a /Guatemala/lugares/turísticos/peten/tikal.jsp.
- El 92 % de los usuarios que visitaron el sitio web visitaron primero la página /Universidad/Facultades/ y seguidamente el 70% de esta muestra visitaron la página /Universidad/Facultades/Ingenieria.aspx.

La primera inferencia indica que algo está sucediendo en el sitio web ya que las visitas son relativamente cortas en el sitio web. Esto puede indicar que la información publicada no es de importancia para los usuarios. O bien que la estructura del sitio web no se entiende o que la usabilidad es muy pobre. Todos estos factores indican que se debe de reorganizar el sitio web con el objetivo de aumentar las visitas y facilitar su uso.

La segunda inferencia indica que la información publicada en cada página web es del interés y aceptación dentro de los usuarios ya que para llegar a /Guatemala/lugares/turísticos/peten/tikal.jsp los usuarios tuvieron que llegar hasta un nivel de profundidad en el servidor de cuatro enlaces.

La tercera inferencia indica que la mayoría de usuarios que acceden al sitio web tienen una mayor preferencia por obtener información acerca de una unidad académica en especial en este caso /Universidad/Facultades/Ingenieria.aspx. Por tal razón se debería brindar un acceso más rápido a este enlace con el fin de evitar saturaciones en el servidor. Además se debería brindar información actualizada y de interés que esté relacionada con el área.

2.6 Análisis de patrones

El análisis de patrones es el último paso en el proceso de minería de uso de la web. Consiste en la selección, validación e interpretación de los patrones más sobresalientes detectados durante el proceso de descubrimiento. Esta es una de las etapas de mayor importancia en todo el proceso de la minería web.

Los resultados de esta etapa deben ser los más efectivos posibles en el descubrimiento de información que se halla oculta. Los resultados se deben presentar de manera que las transacciones, tendencias, cambios, operaciones, preferencias y comportamientos de los usuarios sean lo más claro posible y de fácil identificación. Con esto se pretende mejorar los sistemas, personalizar y reestructurar los sitios web.

El análisis de patrones permite extraer conocimiento útil con respecto al sitio web estudiado. Dentro del análisis de patrones se pueden emplear técnicas de visualización, técnicas de filtración de datos y herramientas OLAP entre otras. Algunos aspectos que se pueden mejorar del lado del servidor son:

- Reorganización de la estructura de las páginas en el servidor web.
- Personalización del servicio a usuarios nuevos.
- Seguridad dentro del sitio web.
- Reorganización de publicidad en el servidor web.

La personalización dinámica de los sitios web es una tendencia que da como resultado la satisfacción del usuario y una correcta administración de perfiles dentro del sitio. Esto se puede llegar a implementar luego del análisis del proceso terminado de minería web. Para tener creada la base de datos de conocimientos que se empleara en el proceso de personalización es necesario tener la

recopilación de técnicas, métodos y algoritmos con el objetivo de obtener información relevante de las necesidades de cada usuario.

Conforme el análisis realizado de minería web y con la finalidad de realizar personalizaciones en los sitios web se recomienda el empleo de componentes de tipo cliente/servidor. Dentro de esta arquitectura el usuario puede participar en el proceso de minería de uso web. Mientras que el servidor ejecuta procesos y se encarga de la administración de la base de conocimientos junto con sus respectivos componentes.

3 PERSONALIZACIÓN WEB

3.1 Minería web en servidores web

Los servidores web despachan miles de peticiones y por tal razón generan una gran cantidad de datos proveniente de los registros de las transacciones que se realizan dentro de ellos mismos. Cada vez que un cliente o usuario lleva a cabo una transacción esta queda almacenada dentro de los ficheros log de manera automática. Todo este volumen de datos que se genera lleva dentro de sí una valiosa información no visible de forma sencilla. Anteriormente los datos se empleaban en su potencia mínima. Es decir solo para obtener estadísticas, detectar accesos inválidos y problemas dentro del servidor web.

Con el surgimiento de la minería de datos y su orientación a la web se dio paso a una nueva tecnología denominada minería web. Por medio de una herramienta de minería de datos se puede llegar a descubrir cosas interesantes acerca de los usuarios como por ejemplo tendencias, preferencias, comportamientos, etcétera. Dada la gran variedad de negocios y transacciones que se manejan a través de internet sumado a la diversidad de competencia existente y a la creciente necesidad de mejoras personalizadas en los servicios es importante tomar en cuenta el análisis de los datos generados.

El conocimiento que se genera en el análisis de la información se emplea entre otras cosas para clasificar clientes usuales y clientes potenciales. Con el análisis también se puede reestructurar el sitio web. El rendimiento de los servidores también se puede mejorar sabiendo cuales son las horas pico del servidor y qué tipo de operaciones requieren de mayor atención.

Existe una gran cantidad de herramientas que llevan a cabo minería del lado del servidor. Estas herramientas son de nivel comercial y de nivel académico. En el análisis de herramientas de minería que trabajan del lado del servidor se observa un común denominador entre todas el cuál es la aplicación del proceso de KDD en los ficheros log del servidor. La forma de almacenamiento de los datos en el fichero log sigue el estándar dado por NCSA y CERN.

Algunas veces los datos almacenados en los ficheros log son insuficientes para realizar la minería. Sin embargo con la correcta aplicación de técnicas de minería de datos se llega a obtener resultados de interés. En ocasiones no se refleja del todo el comportamiento de los usuarios debido a que algunos requerimientos se cachean por el tipo de browser usado ó en ocasiones el servidor proxy no permite el almacenamiento correcto en el servidor. La pérdida de información puede llegar a dar resultados equivocados.

Para solucionar este tipo de problemas se pide a los usuarios que llenen un formulario de registro inicial. De igual forma se implementa algún tipo de logueo o cookies entre el servidor web y el navegador del cliente. Con esta solución se pueden registrar las diferentes transacciones que realizan los clientes. Pero se debe tener en cuenta que en muchas ocasiones los usuarios o clientes prefieren mantenerse en el anonimato. Es por esto que para llevar a cabo el análisis de los datos se toma más en cuenta todas las entradas que poseen los ficheros log.

3.1.1 Beneficios

La aplicación correcta de minería web en los ficheros log de los servidores web proporciona ventajas entre las cuales se puede mencionar:

- Reestructura del sitio web.
- Mejoras en el rendimiento del servidor web.
- Identificación de tipos de clientes o usuarios.
- Mejoras en publicidad y propaganda.
- Personalización de servicios.
- Visualización del comportamiento de los clientes.

3.1.2 Problemas

Algunos problemas que se puede mencionar que afectan la exactitud en los resultados obtenidos del análisis de la información son:

- Datos ambiguos en el fichero log debido a proxies o firewalls.
- Inadecuada estructura de los ficheros log para realizar minería.
- Problemas de delimitación de sesiones o transacciones del usuario.
- Problemas en registros de accesos de páginas cacheadas.
- Uso de backtracking del navegador web.

3.2 Minería web en clientes web

La web contiene una inmensa cantidad de información que cada vez se hace más grande. Existen sitios web que se dedican a la búsqueda de información por medio de palabras clave. Estas herramientas realizan búsquedas por categoría o por contenido.

Cada una de estas consultas se realiza internamente por medio de índices de archivos que se encuentran en lo largo de la red. Los buscadores como Google y Yahoo entre otros encuentran las direcciones URLs relacionadas con la petición del usuario. En ocasiones el resultado de la búsqueda presenta inconsistencias en la información. Otras veces el resultado presenta archivos que satisfacen la consulta pero no cumplen con lo buscado por el usuario.

3.2.1 Agentes inteligentes

Los agentes inteligentes son los encargados de la extracción de características semánticas. Estas características se pueden encontrar en la estructura de un archivo o bien en las palabras en concreto. Las características encontradas se emplean en la clasificación y categorización de archivos. Los agentes inteligentes hacen uso de clustering con lo cual tienen la ventaja de no necesitar conocimiento de las categorías. Es por eso que durante el proceso de categorización no se necesita necesariamente un control. Todos los resultados devueltos por el clustering sirven para plantear consultas automatizadas. Con esto se pueden realizar búsquedas de archivos parecidos y realizar construcciones de perfiles de usuarios.

Los agentes inteligentes se pueden encontrar clasificados en cualquiera de las siguientes categorías:

- *Filtradores y categorizadores de información*: Estos agentes utilizan técnicas de recuperación de información. Además hacen uso de las características de los archivos con la finalidad de recuperar, filtrar y categorizar los archivos. Algunos de estos elementos usan clustering de palabras empleando diversos algoritmos por ejemplo *k-means*. Un ejemplo de estos agentes es *HyPursuit*.

- *Agentes web personalizados*: Estos agentes registran las preferencias del usuario. Encuentran orígenes de información en internet que respondan a las mismas preferencias de otros usuarios con interés iguales. Un ejemplo de este agente es *WebWatcher*.

3.2.2 Diferencias entre un documento web y un hipertexto clásico

Una gran parte de las herramientas de búsqueda en la web únicamente se concentran en información de texto de los archivos. En estos casos se hace caso omiso de información implícita que se encuentre en los enlaces de página. Tratando de manera similar un documento web y un hipertexto cuando aún estos no son del todo iguales. Algunas diferencias entre ambos documentos son las siguientes:

- Links entre documentos y sitios: Los hipertextos contienen enlaces que son diferentes de los encontrados en la web. Los enlaces entre documentos hacen referencia a partes del mismo documento y nunca apuntan a otros documentos que se encuentren en otros sitios web. En la web hay enlaces que apuntan siempre al mismo documento, también hay enlaces que apuntan a un documento en el mismo sitio. También existen otros enlaces que apuntan a documentos de otros sitios web. Para determinar a qué tipo de enlace pertenece cada uno es necesario analizarlos para así clasificarlos detalladamente.

- Información repetida o falta de información: El hipertexto clásico regularmente es completo y sin redundancia. Todo lo redundante en la web puede llegar a convertirse en ventaja si es usado correctamente. Por otro lado la falta de un enlace entre dos páginas no significa que no guarden relación entre ellas. Cosa que se infiere desde un principio en un hipertexto común y corriente.

- Cambio constante: La web se encuentra en un proceso de cambio constante lo que da auge a problemas de consultas de información que aún no se encuentren relacionadas. El problema de cambios constantes no se da en los hipertextos clásicos.

Para buscar y recuperar información en la web es necesario emplear herramientas que consideren las diferencias antes explicadas. Las herramientas que generan índices de búsqueda con mayor razón deberían considerar los diferentes tipos de enlaces para su análisis.

3.2.3 Beneficios

La gran parte de operaciones y desarrollos que se realizan del lado del cliente en la minería web emplean clustering. Esto se hace debido a que en este nivel el propósito principal es el mejoramiento de la calidad de información y mejorar las técnicas de búsqueda de información.

Se puede determinar que los algoritmos de clustering no son los más adecuados para el manejo de información documental. Los algoritmos de clustering empleados en la web no se basan en una función de distancia. Los algoritmos de clustering en su lugar usan el particionamiento de hipergrafos siendo más eficientes por las siguientes razones:

- No dependen de la elección de la función de distancia.

- No son sensitivos a las dimensiones de los datos.
- Son linealmente escalables.

La principal ventaja de aplicar minería de datos del lado del cliente es por la simplificación en el proceso de obtener información. Por medio del empleo de técnicas de minería de datos se pude detectar información irrelevante. Además se obtiene información en mayor cantidad sobre el tema buscado.

3.3 Cookies

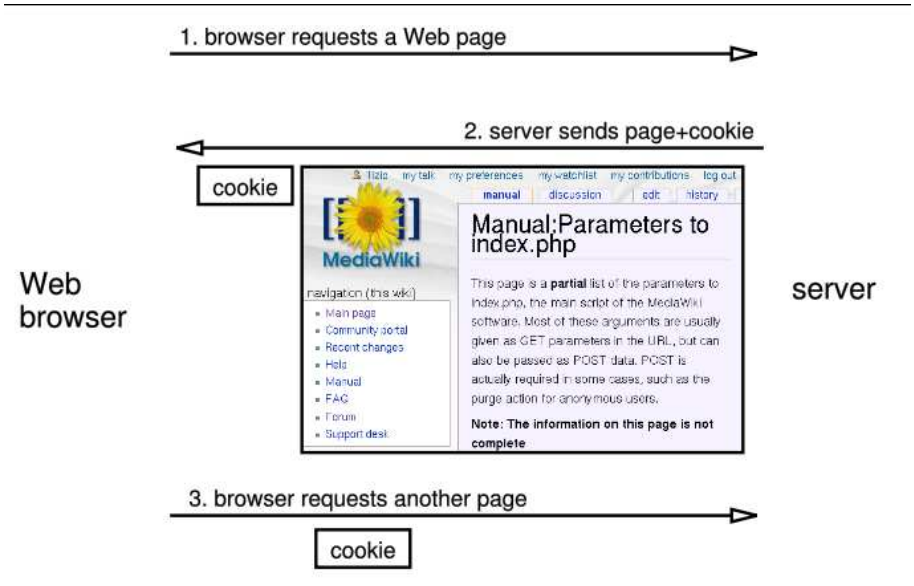
Las cookies ó galletas es una técnica popular que facilita la identificación de usuarios y de sesiones. La principal ventaja de las cookies es que no necesitan que se registre algún tipo de información por parte de los usuarios. Con el uso y empleo de cookies se pueden realizar operaciones como rastreos, autenticaciones, preservación de información, etcétera. Del lado del cliente las cookies se encuentran como una opción en los navegadores web. Es decir que para permitir la recepción de las mismas se necesita que el usuario configure el navegador web. Una desventaja del uso de cookies es que cuando se borran las mismas por medio del navegador web ya sea de forma accidental ó apropósito, como consecuencia se pierde todo el historial del usuario.

Hay varias implicaciones acerca del anonimato y privacidad de los usuarios cuando navegan en la red. Aunque las cookies son enviadas únicamente entre el servidor y el cliente, el uso de cookies de terceras personas se puede dar por medio de anuncios ó imágenes de otros dominios cuando se carga una página web. Aún con el uso de cookies la identificación de usuarios puede ser algo muy complejo ya que las cookies no identifican de forma directa a una persona sino más bien identifican parte de la cuenta del usuario, del navegador y de la computadora empleada. Siendo de esta forma que si un usuario usa varias computadoras, varios navegadores o varias cuentas de usuario, las cookies

registraran la información como si fueran diferentes personas y no la misma. Cuando se tiene una situación en que diferentes usuarios usan la misma computadora, con el mismo navegador web y la misma cuenta de usuario, las cookies identificaran a todas estas personas como si fueran la misma.

Un problema de seguridad es la vulnerabilidad que hay de ciertos navegadores web que dejan a un tercero colocar una cookie inexistente en el espacio de otro dominio. Cada vez que se genera una cookie se debe tomar en cuenta la configuración que realiza el servidor sobre algunos parámetros de la cookie (nombre del dominio, nombre de la cookie, fecha de expiración). Si la fecha de una cookie no se encuentra detallada, la cookie desaparece cuando se cierra el navegador web. Por tal motivo, es mejor detallar una fecha de expiración, a las cookies que poseen fecha límite de vencimiento se les llama cookies persistentes.

Figura 6. Funcionamiento de una cookie



Fuente: *http://es.wikipedia.org/wiki/Cookie*

3.4 Personalización web

La personalización web es la habilidad que tiene un sitio web para mantener la atención de los usuarios. Con la personalización se puede guiar a los usuarios de forma satisfactoria para obtener información útil. La personalización web considera el modelado de objetos web (como páginas y tópicos), la búsqueda de coincidencias, así como el conjunto de acciones que se debe tomar para realizar la personalización.

Las acciones a tomar dependen del tipo de personalización a realizar. La personalización puede ser basada en sistemas de reglas de decisión manuales, en agentes de filtrado basados en el contenido o en sistemas de filtrado colaborativo.

Al inicio el usuario se encuentra navegando y el sistema determina cuales son las preferencias del usuario. Seguidamente se traslada la información detectada al módulo de personalización. Este modulo toma las preferencias del usuario y sus necesidades de información. El modulo se encarga de procesar la información y crea una consulta para la base de conocimientos.

En la base de conocimientos se lleva a cabo un proceso con la finalidad de extraer información. Para extraer la información se usan diferentes mecanismos implementados con anterioridad. Luego se retorna la información que se considera de mayor relevancia para el usuario al módulo de personalización. Entonces el sistema de personalización toma estos datos y efectúa los cambios más convenientes. El sistema de personalización le informa al usuario de estos cambios por medio de sugerencias y solicitudes.

La personalización por medio de sugerencias se puede llevar a cabo de manera tanto grupal como individual. Cuando se realiza una personalización a

nivel grupal se trata de encontrar la relación de un usuario con cierto grupo por medio de sus preferencias y comportamientos. Cuando se encuentra un grupo de usuarios el proceso de personalización no necesita de procesos intensivos para efectuar los cambios pedidos por el usuario. En cambio si las sugerencias son proporcionadas a los usuarios de forma individual el proceso en cuanto refiere a extracción y procesamiento es más intensivo. Esto se debe a que los procesos de búsqueda resultan más complicados de llevar a cabo.

Por otra parte, si el análisis del usuario se lleva a cabo en tiempo real el proceso de personalización también se realiza en línea. Cuando se emplea un enfoque estático el análisis y la personalización se pueden efectuar con anticipación para que se le pueda brindar al usuario tiempos de respuesta más efectivos. Cuando se realizan enfoques híbridos es decir combinaciones de métodos la mejor manera de realizar el proceso es en línea.

El proceso de análisis de comportamiento del usuario y de personalización en tiempo real se lleva a cabo cuando se emplean técnicas que están basadas en las visitas y en las transacciones. De igual forma en aquellos métodos que hacen uso de cookies. Estos tipos de sistemas de análisis necesitan de capacidades de cómputo y de ancho de banda más grandes de lo común ya que son procesos que consumen más recursos. Siempre que se comienza cualquier tipo de personalización se debe tener previo conocimiento de los gustos, comportamientos y elecciones de los usuarios.

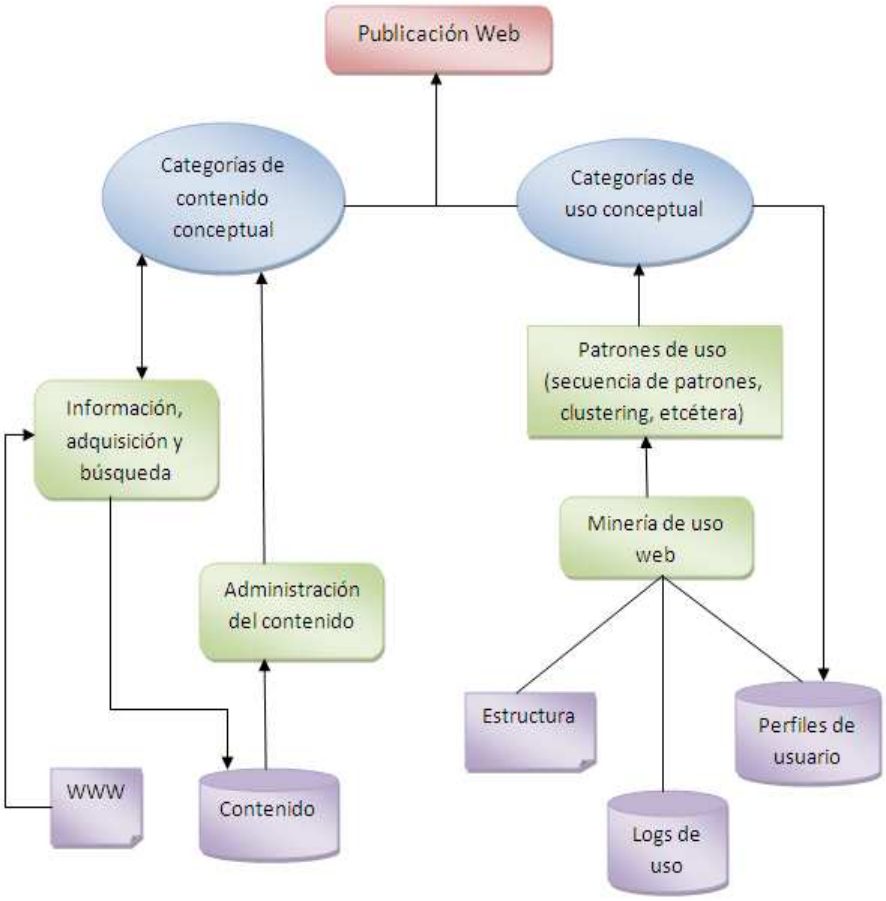
Las personalizaciones individuales y grupales se basan en los perfiles creados de los usuarios. Pero una desventaja de basarse en perfiles de usuario es que estos contienen información subjetiva que en poco tiempo se convierte en información obsoleta. Por esta razón, es que se necesita de la implementación de mecanismos por aparte que sean capaces de aprender del comportamiento del usuario y tengan la capacidad de modificar el perfil de manera automática.

Cuando se realizan personalizaciones en grupo con base a los perfiles de usuario se pretende buscar perfiles que tengan un común denominador entre ellos. Una vez que se realice una sugerencia al grupo de usuarios es recomendable que los usuarios evalúen los elementos con el fin de que el sistema cambie los perfiles con base a las ponderaciones obtenidas.

La desventaja de las sugerencias grupales es la ausencia de escalabilidad en los sistemas de personalización. La formación de un conjunto de perfiles parecidos se debe efectuar en línea y los conjuntos de información grandes provocan retrasos en los tiempos de respuesta. Cuando los datos se encuentran alejados el proceso se vuelve más difícil. Cuando el conjunto de usuarios que tiene preferencias similares es demasiado grande la ponderación de cada usuario se convierte en un valor significativo en relación al tamaño del conjunto.

Cuando se realizan recomendaciones o sugerencias grupales el filtrado colaborativo necesita tener concordancia con la ponderación de un usuario y con la de usuarios similares para crear recomendaciones en elementos que aún no se procesan. La personalización se puede llevar a cabo sin necesidad de conocer el comportamiento previo de los usuarios. Esto se logra por medio del estudio del comportamiento de los usuarios mientras estos navegan en el sitio web. Para realizar este estudio se analiza las secuencias de páginas visitadas con el patrón de secuencia de páginas visitadas para completar una transacción.

Figura 7. Módulos de un sistema de personalización web



Fuente: *http://www.engr.sjsu.edu/meirinaki/papers/EV03_TOIT.pdf*

3.5 Clasificación de herramientas de minería web

Debido a la expansión de las fuentes de información en Internet, es de gran utilidad que los investigadores y analistas empleen herramientas para poder conocer y determinar los patrones en su uso. Estas herramientas son sistemas de aprendizaje autónomo que trabajan en dos lugares: del lado del servidor y del lado del cliente para poder "*minar*" la información que se genere al momento de utilizar Internet.

3.5.1 Herramientas incorporadas en el servidor

Estas corresponden a programas de software que son capaces de procesar en tiempo real los datos que se van almacenando en los archivos log. Se ejecutan del lado del servidor y el acceso a los datos se lleva a cabo por medio de un interfaz en línea. Por lo regular estas herramientas ya se incluyen en servidores dedicados o compartidos.

3.5.2 Herramientas incorporadas en maquinas personales

Estas soluciones son de igual forma programas de software que se instalan de manera separada en las máquinas de los usuarios. El objetivo es de igual forma analizar los archivos log con la diferencia que estas herramientas no procesan en tiempo real. Estas herramientas consisten en descargar los archivos log de un servidor para luego realizar su procesamiento.

Este tipo de solución es una manera atractiva de realizar el análisis por parte de los investigadores ya que no se necesita conexión a Internet y se llevan a cabo informes estadísticos en un tiempo mínimo.

3.6 Herramientas comerciales

Dentro de las herramientas de software comerciales para el desarrollo de minería de web se encuentran:

3.6.1 Clementine

Es una herramienta de minería de datos bajo la firma *SPSS*. Emplea algoritmos innovadores de inferencia para construir los caminos transversales e identificar las sesiones de usuario. Los avanzados algoritmos de minería descubren los movimientos de los usuarios dentro del sitio web.

El resultado final es una recopilación de valiosos patrones de navegación los cuales ayudan a los administradores del sitio web a comprender mejor la conducta de los usuarios.

3.6.2 Commerce trends

Es una herramienta de minería de datos bajo la firma *Web Trends*. Proporciona una poderosa herramienta de reportes de e-Business intelligence. Permite la localización de clientes, el manejo y optimización de estrategias e-Business.

Esta herramienta incluye funcionalidades poderosas tales como la escalabilidad de la empresa en el análisis de tráfico web, el manejo de la campaña, el rédito del e-Commerce y el e-Marketing. Todo esto para permitir a los usuarios aplicar principios de data warehouse en la correlación de datos del tráfico web con otra información corporativa como CRM, ERP y sistemas de personalización.

3.6.3 DB miner

Esta es una herramienta de minería de datos desarrollada por *Fraser Simon* de la *Universidad de Canadá*. Es una herramienta poderosa y económica para el desarrollo de data warehouse y bases de datos relacionales de una manera rápida y eficiente empleando múltiples funciones de minería.

Esta herramienta emplea Microsoft SQL Server 7.0 para elaborar los cubos de datos en los cuales realizara las tareas de minería. Esta característica aumenta considerablemente la eficiencia y versatilidad.

3.6.4 Funnel web pro

Es una herramienta de análisis de ficheros log desarrollada por *Active Concepts*. Es una de las más recientes versiones del clásico análisis inteligente y software de reportes de Internet.

Esta herramienta ofrece una gama de nuevas capacidades tal como una entera administración remota basada en web.

3.6.5 Knowledge studio

Es una herramienta de minería desarrollada por *Angoss*. Es una herramienta de nueva generación de minería de datos. Integra técnicas de minería de datos avanzadas en ambientes corporativos para que las empresas puedan alcanzar beneficios máximos a partir de su inversión en los datos.

Incluye herramientas eficientes para la visualización de datos que sirven de apoyo y explican el conocimiento encontrado.

3.6.6 Net analysis

Es una herramienta de minería de datos desarrollada por *Net Genesis*. Es una solución del análisis de la conducta en línea. Proporciona una escalabilidad superior y una poderosa extensibilidad requerida para empresas e-Businnes. Con lo que logra una ventaja dinámica aumentando la competencia del ambiente en línea.

Con una alta flexibilidad y funcionalidad esta herramienta puede personalizarse para satisfacer las necesidades específicas en la inteligencia del e-Customer de cualquier compañía.

3.6.7 Sawmill

Es una solución de análisis de ficheros log desarrollada por *Flowerfire*. Es una herramienta poderosa de análisis jerárquico de logs para plataformas MacOS, UNIX, OS/2, BeOS y Windows 95/98/NT/2000.

Es particularmente empleada en el acceso al servidor web, pero puede procesar cualquier tipo de fichero log. Los reportes que genera son jerárquicos y atractivos para una fácil navegación. La documentación completa se construye directamente en el programa.

3.6.8 Speed tracer

Es una herramienta de minería de datos desarrollada por *IBM*. Es una solución de análisis y de minería de uso web que identifica patrones de usuarios, genera reportes para ayudar a los administradores del sitio web a mejorar la estructura y la navegación del sitio.

El resultado final es una colección de patrones valiosos del navegador que ayudan a comprender el comportamiento de los usuarios. Speed tracer genera tres tipos de estadísticas: basadas en el usuario, basadas en la ruta y basadas en el grupo.

Las estadísticas basadas en el usuario hacen referencia en las cuentas de usuario y el tiempo de duración de sus accesos. Las estadísticas basadas en la ruta identifican los frecuentes caminos que se toman en la web. Las estadísticas basadas en grupo proporcionan información sobre los grupos de páginas de sitios web más visitados.

3.6.9 WUM

Es una herramienta de minería de datos desarrollada por la *Universidad de Berlín*. Es una solución de secuencia de minería. Su principal objetivo es analizar la navegación de los usuarios en un sitio web. Esta herramienta es apropiada para el descubrimiento de patrones secuenciales en cualquier tipo de ficheros log. Además descubre patrones de eventos que no necesariamente se encuentren relacionados.

WUM es un ambiente integrado para la preparación de ficheros log, consultas y visualizaciones. Su lenguaje de minería de consultas es "*MINT*" que soporta la especificación de criterios describiendo patrones dominantes ó patrones estadísticamente extraños.

Su mecanismo de visualización muestra los nodos comprometidos de los patrones deseados y los diferentes caminos poco frecuentes de las rutas localizadas. Esto es de vital importancia al momento de examinar cómo está siendo realmente navegado el sitio web.

3.7 Herramientas públicas

Dentro de las herramientas de software público para el desarrollo de minería web se encuentran:

3.7.1 Analog

Es una herramienta basada en el análisis de ficheros log. Fue desarrollada por el laboratorio de estadísticas de la Universidad de Cambridge. Es un programa para analizar los ficheros log de un servidor web.

Esta herramienta indica que páginas son las más populares, de que países son las personas que están visitando el sitio web, de que sitios los usuarios están tratando de seguir los enlaces incompletos, etcétera.

3.7.2 STStat

Es una herramienta de reportes y estadísticas, fue desarrollada por *ST Software*. La herramienta se encuentra formada por un juego de scripts CGI que produce reportes HTML en base al acceso de logs que los servidores HTTP mantienen.

Esta solución es conveniente para casi todos los servidores de software HTTP (UNIX y Windows) soportando tres formatos de logs los cuales son: Common, IIS y Extended.

3.7.3 WebLog

Es una herramienta basada en el análisis de ficheros log, fue desarrollada por *Darryl C. Burgdorf*. Es una herramienta que permite almacenar huellas de las actividades en el sitio por mes, semana, día y hora.

Monitorea el total de hits, los bytes transferidos, las páginas vistas y las preferencias de los usuarios.

3.7.4 WebLog parse

Es una herramienta para el procesamiento de ficheros log. Fue desarrollada por los laboratorios de software *ACME* y permite la extracción específica de campos de un fichero log.

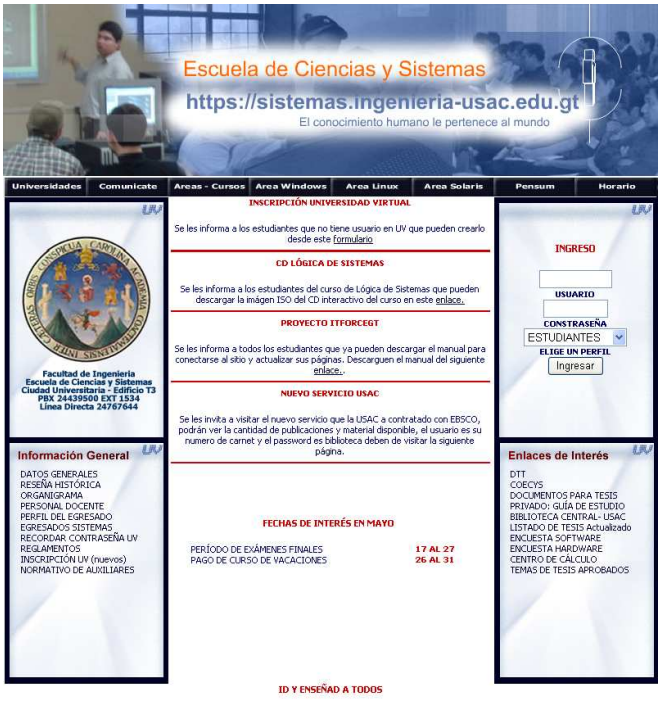
Puede leer un fichero de log en cualquier formato de fichero log *Common* o en cualquier formato de fichero log *Combined*. Parsea los archivos log y escribe en la salida solo los campos especificados por el usuario, separados por etiquetas para un manejo más fácil.

4 CASO DE ESTUDIO

4.1 Descripción

En el caso de estudio se analizaron los archivos log del servidor de la Escuela de Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala. Esto con la finalidad de observar el comportamiento de los estudiantes de dicha institución y así poder determinar patrones de conducta dentro del sitio web. Con la información procesada y analizada se puede extraer conocimiento útil que sirva para personalizaciones, reestructuraciones o bien como apoyo en la toma de decisiones. Para el estudio se procesaron y analizaron los ficheros log del servidor, comprendidos en el periodo del 1 de enero del año 2006 al 31 de diciembre del año 2006.

Figura 8. Sitio web de estudio



Fuente: *http://sistemas.ingenieria-usac.edu.gt*

4.2 Tipos de usuario

El sitio web de la escuela de ciencias y sistemas comúnmente llamado universidad virtual tiene como finalidad la interacción entre el personal docente y el estudiantado. El acceso al sitio web se lleva a cabo por medio del ingreso de un nombre de usuario, de una contraseña de usuario y la selección del tipo de usuario. El portal posee un alto valor intangible ya que este medio sirve como puente de comunicación entre ambas partes. Los usuarios que tienen acceso al sitio web son los siguientes:

4.2.1 Estudiante

El tipo de usuario "*estudiante*" representa a todos los alumnos de la escuela de ciencias y sistemas. Las operaciones que este tipo de usuario puede realizar son: modificación de datos personales, consulta de cursos asignados, consulta de actividades, consulta de horario de clases, asignación de cursos, eliminación de cursos asignados, envío de mensajes a profesores y consulta de enlaces públicos.

Cuando un alumno se asigna cursos de la carrera de ciencias y sistemas se le presenta una lista de enlaces de los cursos asignados actualmente y de los cursos asignados en pasados ciclos académicos. En donde en cada curso se presentan mensajes respecto al desarrollo del mismo, se presenta contenido del curso previamente subido por los profesores ó auxiliares (notas, tareas, enunciados de proyecto, etcétera).

4.2.2 Catedrático

El tipo de usuario "*catedrático*" representa a todo el personal docente que imparte cursos de la escuela de ciencias y sistemas. Las operaciones que este

tipo de usuario puede realizar son: modificación de datos personales, consulta de cursos que imparten, consulta de actividades, consulta de horarios, consulta de listado general de profesores, consulta de alumnos asignados en el curso impartido, ingreso de material relacionado al curso y envío de mensajes a alumnos.

4.2.3 Auxiliar

Los auxiliares son las personas que sirven de apoyo al profesor del curso y que manejan el contenido del laboratorio de la clase. Estas personas se registran bajo el tipo de usuario “catedrático” y pueden realizar las mismas operaciones de este tipo de usuario.

4.2.4 Administrador

El administrador es la persona responsable del manejo del sitio web. Entre algunas operaciones que puede realizar están: mantenimiento del sitio, realización de cambios en la estructura, registro de nuevos usuarios, evaluación y control del sitio web.

4.3 Herramienta empleada

Para el caso de estudio se utilizo la herramienta Sawmill en su versión Sawmill7.2.9_x86_win32 (Demo). Esta herramienta puede emplearse en servidores de navegación de tipo ISA SERVER Proxy. Lo cual significa que se puede usar en servidores que generan ficheros log con una estructura diferente a los que genera un servidor Internet Information Server (IIS).

Sawmill es una poderosa herramienta de análisis de logs que puede procesar casi cualquier tipo de logs. Los reportes que genera Sawmill son jerárquicos, atractivos y de fácil navegación. Sawmill se puede ejecutar sobre las siguientes versiones de plataformas:

- x86/Pentium system running Windows (95, 98, ME, NT, 2000, XP, or 2003)
- x86/Pentium system running Linux
- x86/Pentium system running FreeBSD
- x86/Pentium system running OpenBSD
- x86/Pentium system running BSD/OS
- x86/Pentium system running Solaris
- Macintosh running MacOS X
- Sun workstation running Solaris
- Sun workstation running Linux
- Alpha workstation running Digital UNIX
- Alpha workstation running Linux
- IBM workstation running AIX
- HP workstation running HP/UX

Sawmill puede soportar hasta 828 formatos de logs diferentes. Además ofrece una gran cantidad de opciones para el procesamiento de los ficheros. Sawmill

incluye una base de datos persistente junto con una variedad de opciones de filtrado sobre los ficheros. Algunas características sobresalientes de Sawmill son:

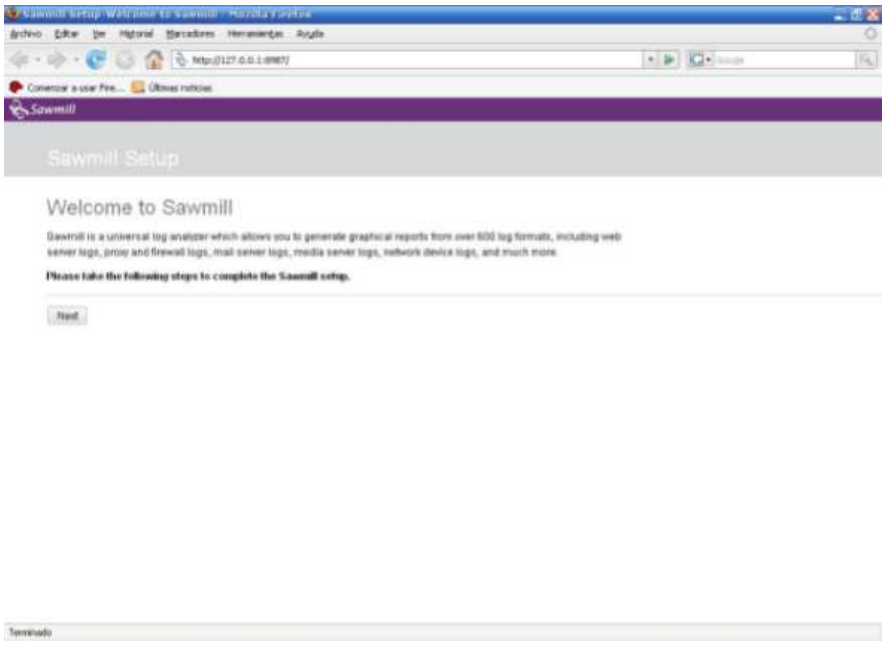
- Facilidad de uso.
- Documentación extensiva.
- Reportes y graficas en tiempo real.
- Paquete de herramientas de análisis.
- Estadísticas atractivas.
- Manejo de la base de datos.
- Rapidez.
- Multiplataforma.
- Altamente configurable.

4.4 Instalación de la herramienta

A continuación se detallan los pasos a seguir para instalar Sawmill de forma correcta.

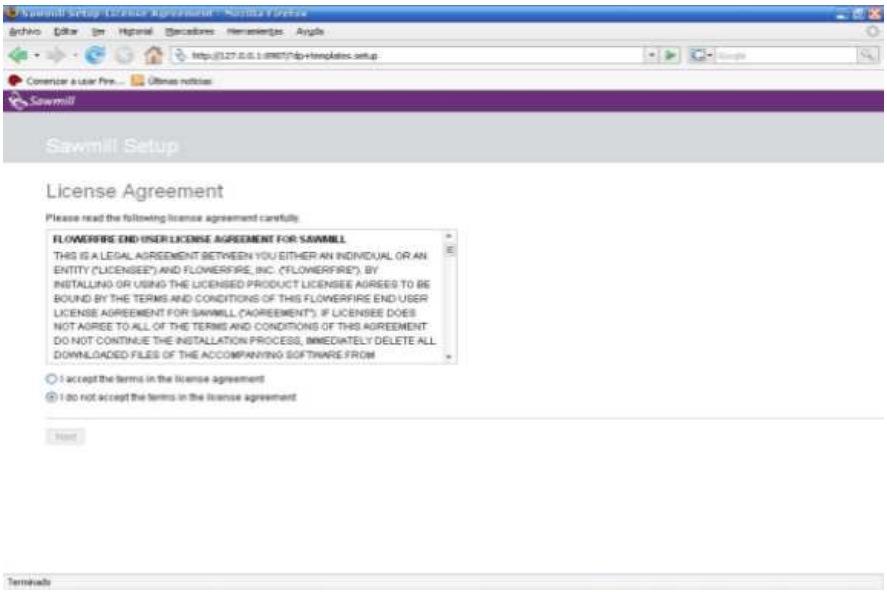
- Al ejecutar el instalador de la herramienta se muestra la pantalla de bienvenida en la cual aparece una pequeña descripción de la herramienta. Presionando el botón “siguiente” continua el próximo paso.

Figura 9. Pantalla de bienvenida de Sawmill



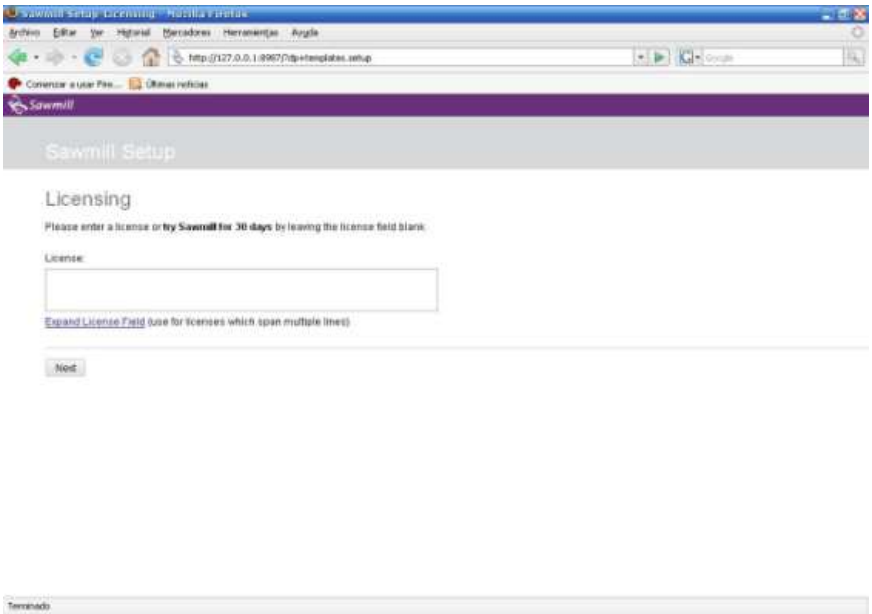
- Seguidamente se presenta la pantalla del acuerdo de licencia de la herramienta, donde se debe aceptar la misma. Presionando el botón “siguiente” continua el próximo paso.

Figura 10. Acuerdo de licencia de Sawmill



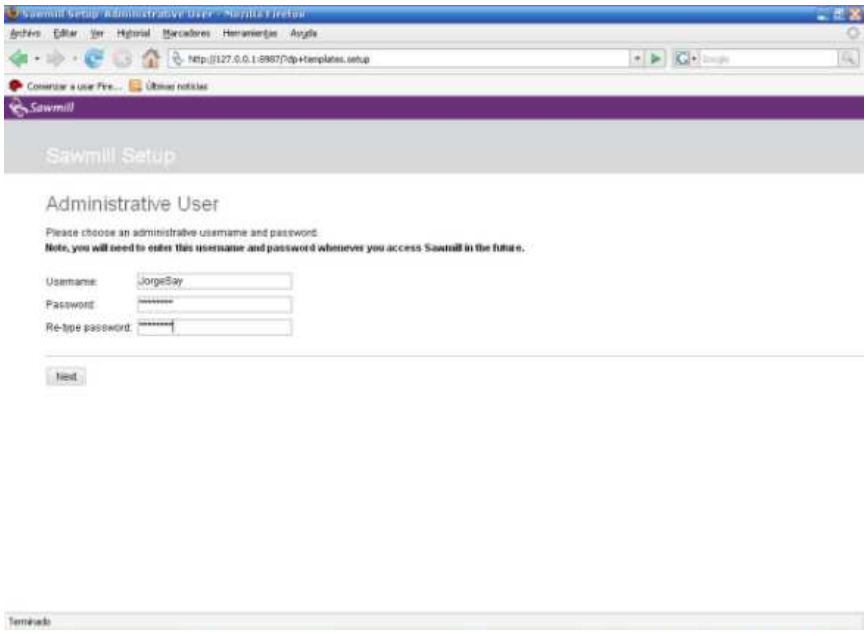
- Luego se muestra una pantalla donde se debe ingresar la clave de registro de licencia o bien se deja este espacio en blanco para acceder a una versión de prueba de 30 días.

Figura 11. Registro de licencia de Sawmill



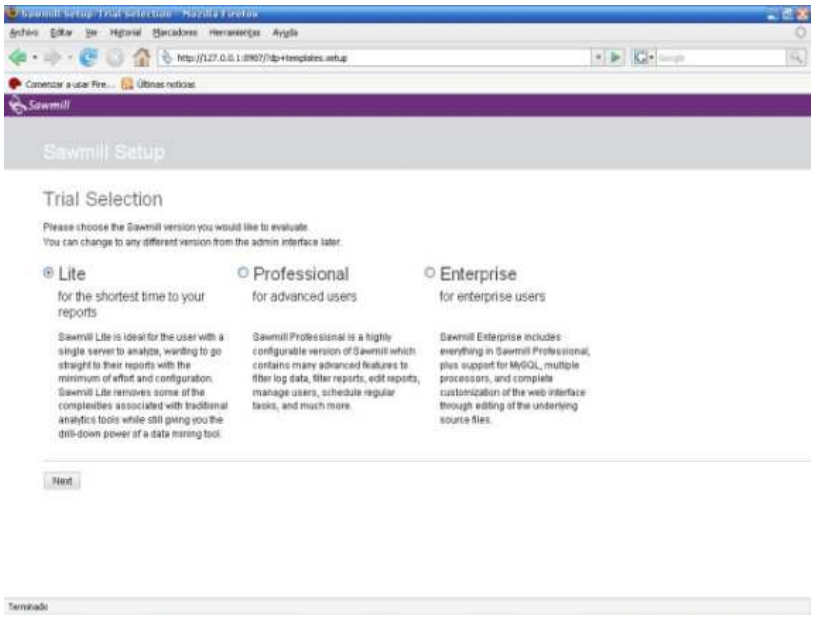
- En la siguiente pantalla se solicita el ingreso de los datos del administrador. Se debe ingresar el nombre y la contraseña del usuario. Presionando el botón “siguiente” continua el próximo paso.

Figura 12. Ingreso de datos del administrador de Sawmill



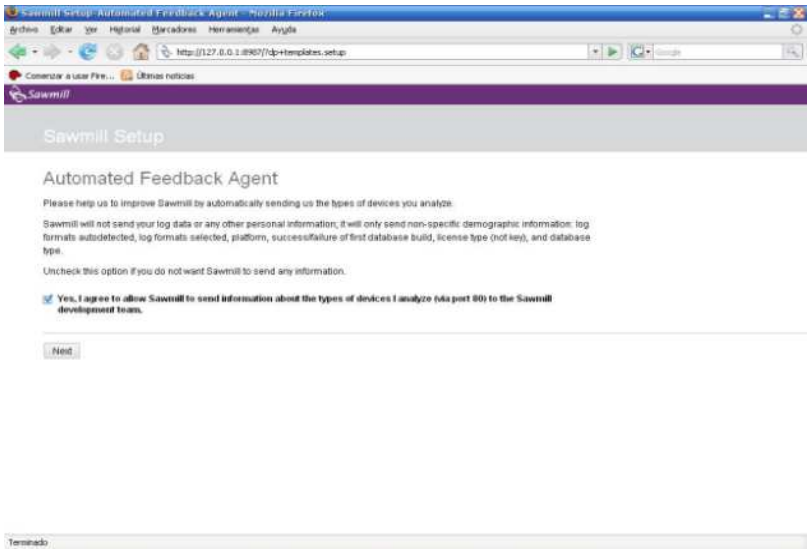
- A continuación se solicita la selección de la versión de Sawmill que se desea instalar. Las versiones con las que cuenta Sawmill son ligera, profesional y empresarial. Luego de la selección se presiona el botón “siguiente” para continuar con el próximo paso.

Figura 13. Versiones de Sawmill



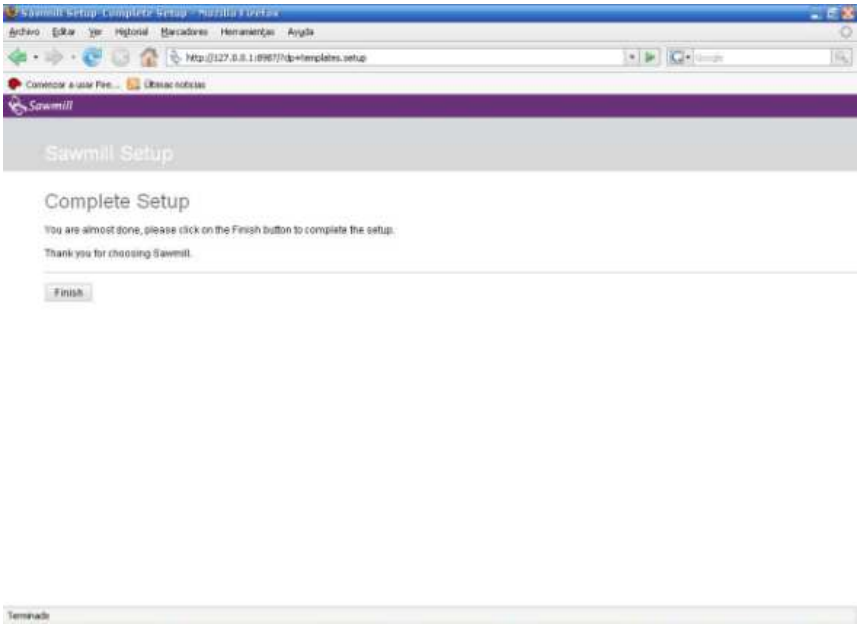
- Luego se presenta una pantalla donde se muestra la opción de recibir y enviar información del equipo de desarrollo de Sawmill. Presionando el botón “siguiente” continua el próximo paso.

Figura 14. Retroalimentación de Sawmill



- Como último paso se presenta una pantalla donde se indica el estado de finalización de la instalación.

Figura 15. Finalización de la instalación de Sawmill

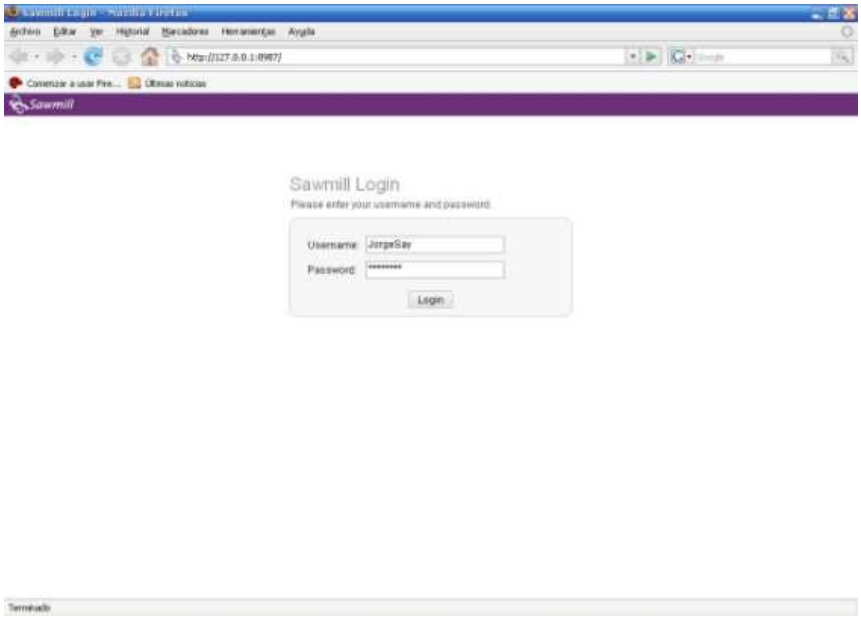


4.5 Creación de un nuevo perfil en la herramienta

A continuación se detallan los pasos a seguir para crear un nuevo perfil en Sawmill de forma correcta. Este perfil sirve para el ingreso, procesamiento y análisis de ficheros log.

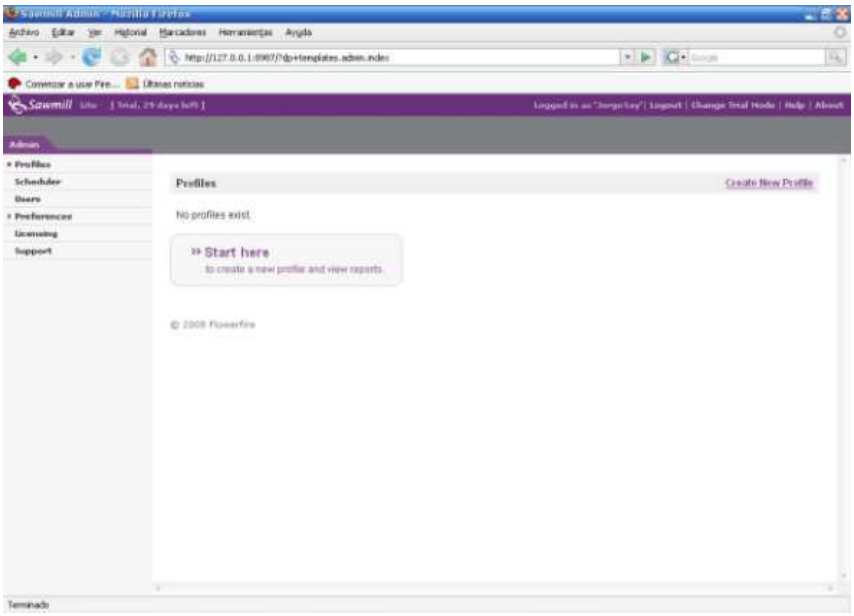
- Luego de instalada la aplicación se procede a ingresar a la misma por medio del nombre y contraseña del usuario.

Figura 16. Ingreso a la aplicación Sawmill



- Seguidamente del logueo correcto en la aplicación se presenta la pantalla principal de Sawmill. En este momento no existe ningún perfil creado, por lo cual se procede a crear un nuevo perfil seleccionando la opción “crear nuevo perfil” que se encuentra en la barra superior central de la aplicación.

Figura 17. Creación de un perfil en Sawmill



- Después de seleccionar la opción para crear un nuevo perfil se muestra una ventana en la cual se pide el ingreso de la ruta dentro de la computadora donde se encuentran los ficheros log.

Figura 18. Solicitud de la ruta de los ficheros log

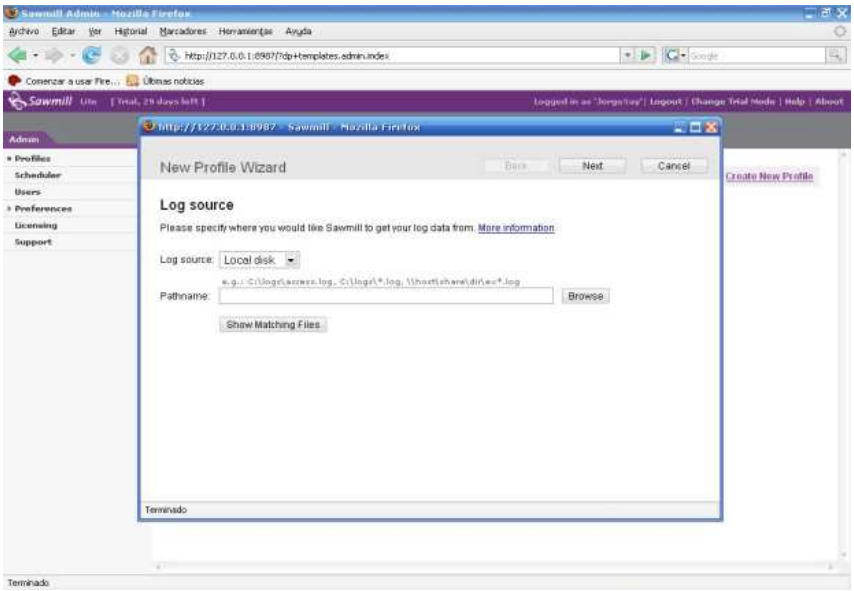
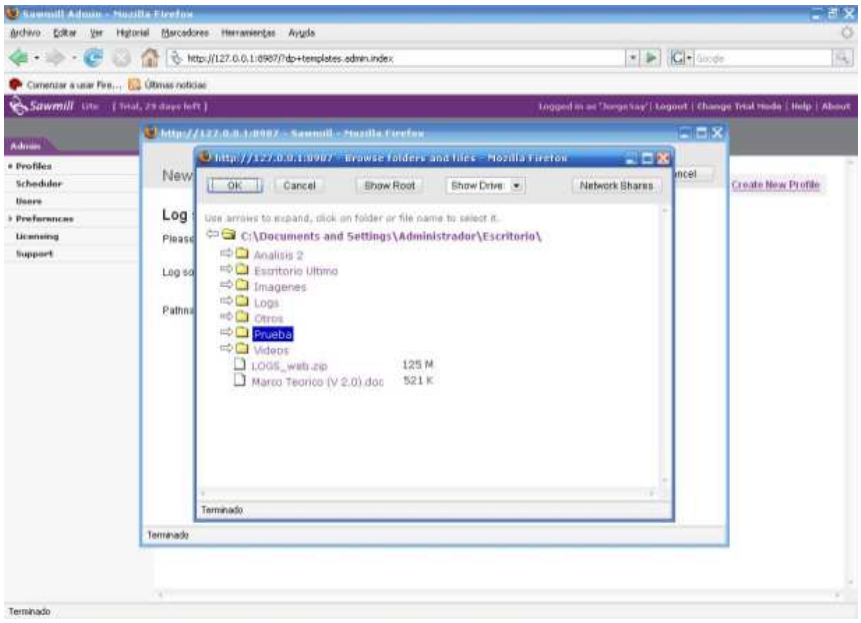


Figura 19. Búsqueda y selección de la ruta de los ficheros log



- Seleccionada la ruta de los ficheros la herramienta procede a la detección de los ficheros log seleccionados.

Figura 20. Carga de los ficheros log

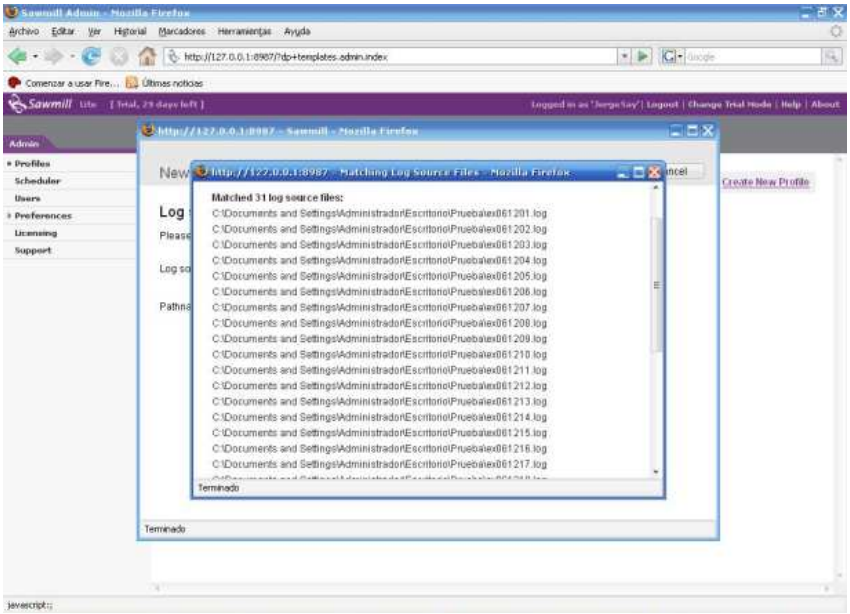
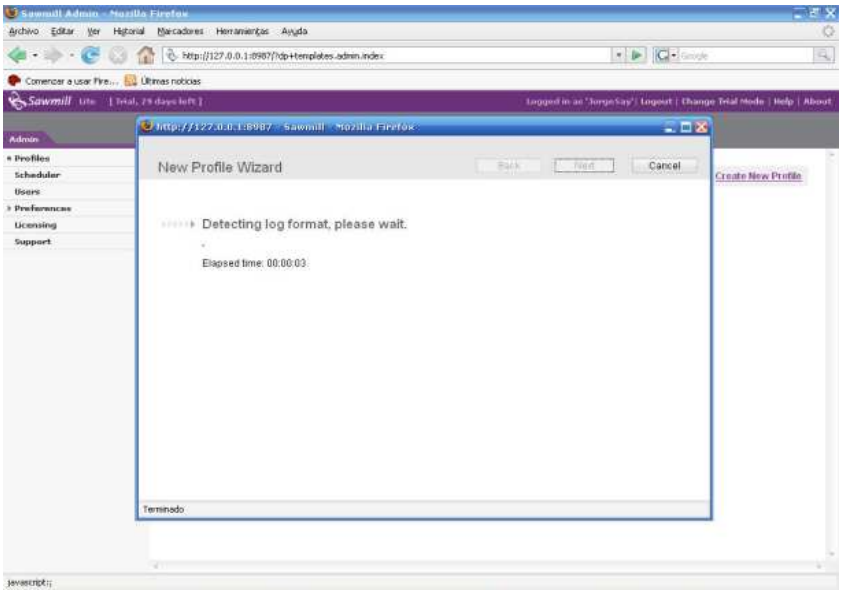
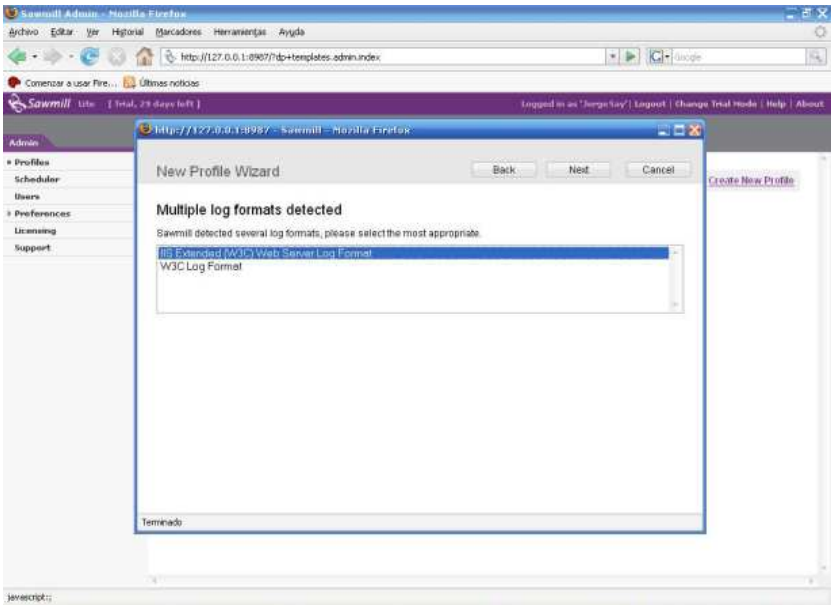


Figura 21. Detección de los ficheros log



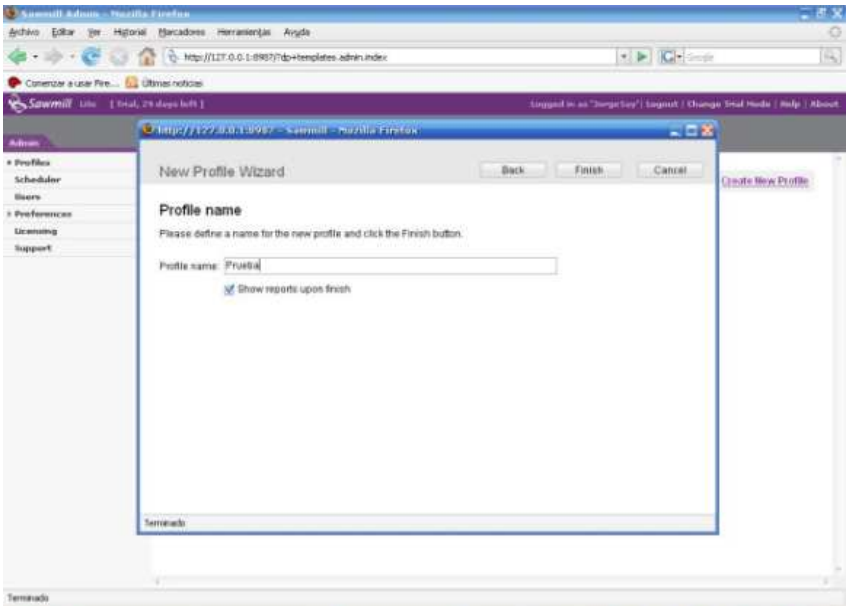
- Luego de la detección de los ficheros log, se procede a seleccionar la estructura de los ficheros logs encontrados.

Figura 22. Selección de estructura de los ficheros log



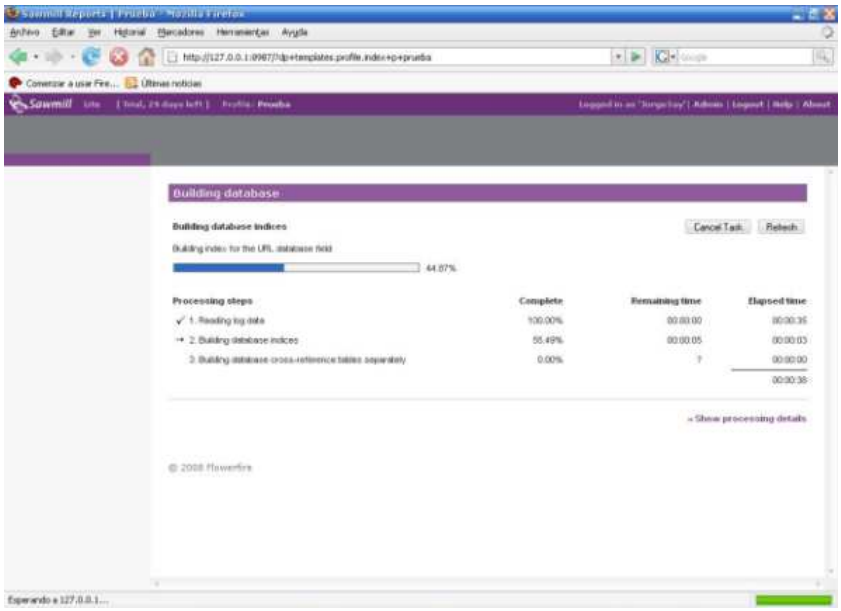
- Seguidamente la herramienta solicita la entrada del nombre que se le dará al nuevo perfil.

Figura 23. Ingreso de nombre del nuevo perfil



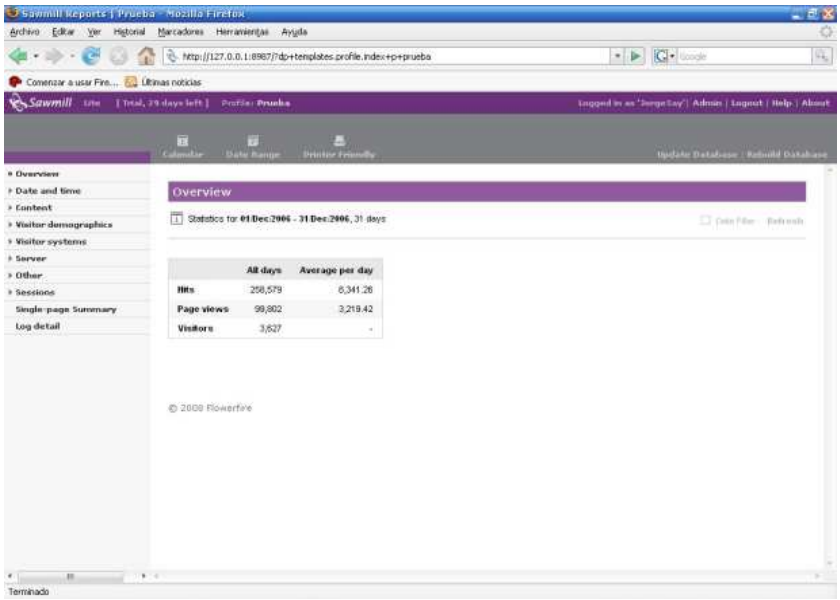
- A continuación la herramienta comienza a procesar los ficheros log. En este paso se genera una nueva base de datos en base a la información previamente detectada y registrada en los ficheros log.

Figura 24. Procesamiento de ficheros log



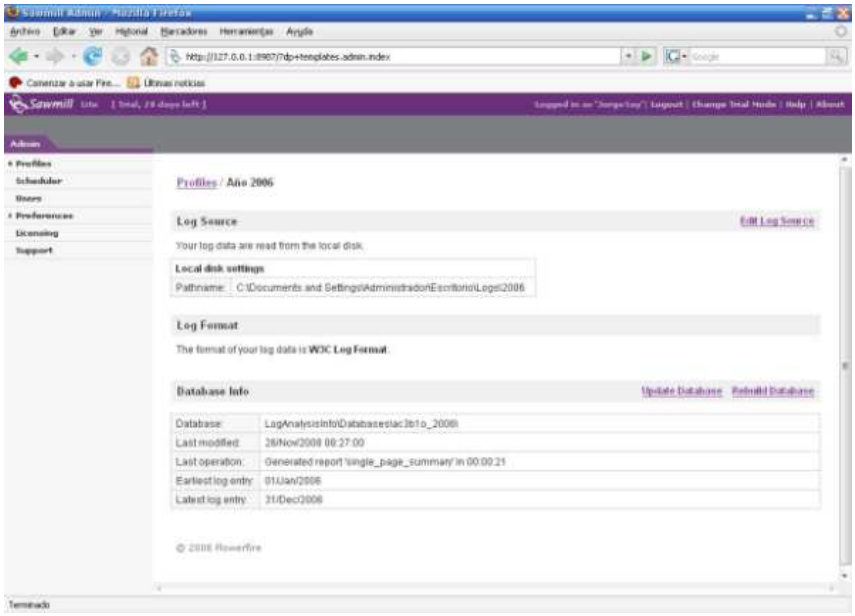
- Terminado el procesamiento de los ficheros log se presenta una pantalla en la cual se indica la finalización del proceso. En esta pantalla se presentan los resultados generales del procesamiento.

Figura 25. Finalización del procesamiento



- En la herramienta existe la opción de verificar información acerca del nuevo perfil creado. Se cuenta con información de la ruta fuente de los ficheros log, la estructura de los ficheros log y la información del procesamiento de la nueva base de datos. Además se pueden consultar los usuarios registrados.

Figura 26. Información general del nuevo perfil



4.6 Resultados obtenidos

4.6.1 Informe de uso por año y mes

A continuación se presenta un informe de uso del sitio web por parte de los usuarios. Este informe se encuentra detallado por año y por mes.

Tabla II. Informe de uso por año

Año	Peticiones	Páginas Vistas	Visitantes
2006	8,281,839	3,498,124	59,469

Figura 27. Informe de uso por año

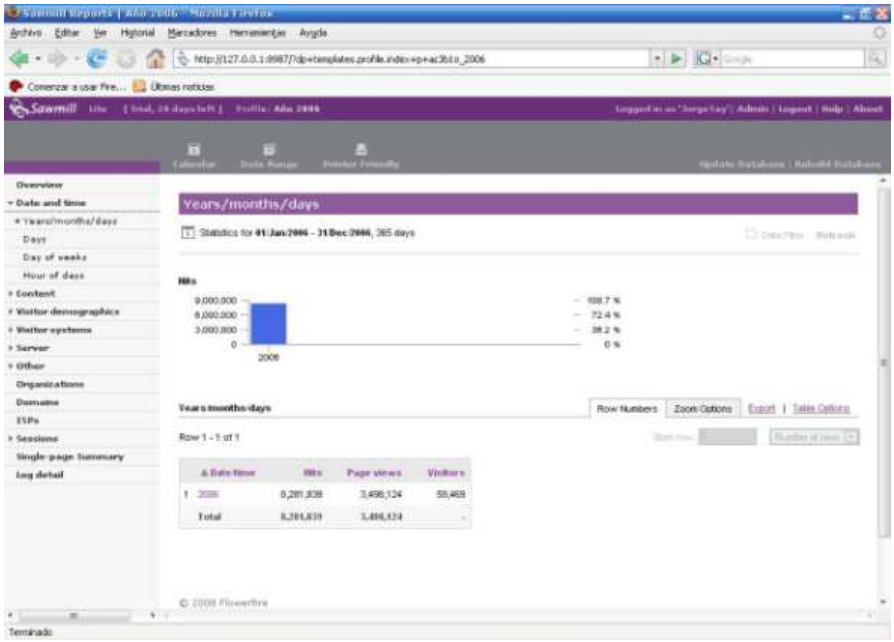
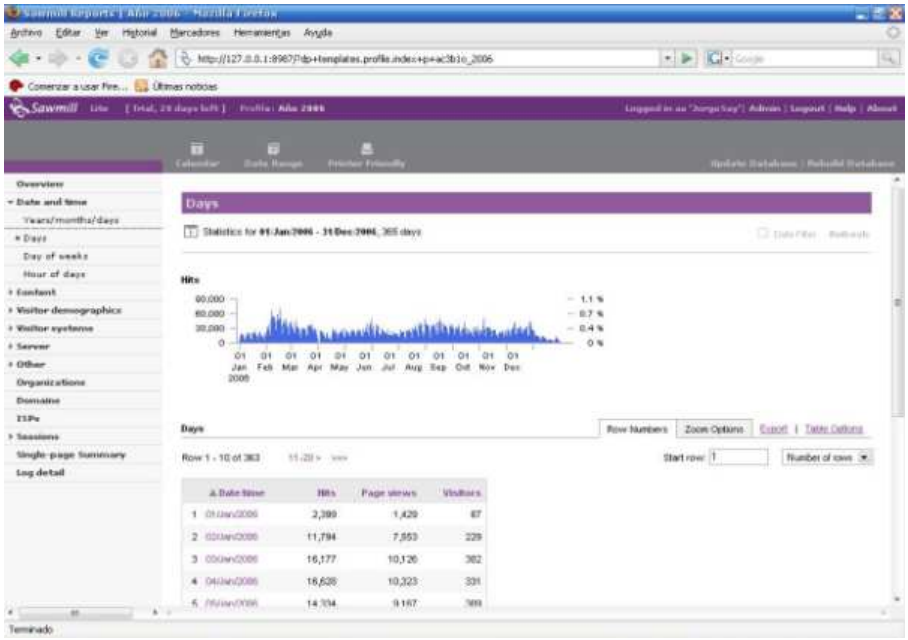


Tabla III. Informe de uso por mes

Mes	Peticiones	Páginas Vistas	Visitantes
Enero	419,978	260,698	10,180
Febrero	1,103,794	476,919	12,974
Marzo	828,842	329,245	10,607
Abril	474,299	187,623	7,076
Mayo	703,582	296,380	9,912
Junio	613,246	254,801	8,830
Julio	717,325	290,322	9,263
Agosto	950,536	400,240	10,973
Septiembre	745,909	304,299	11,147
Octubre	689,134	276,075	10,234
Noviembre	776,615	321,720	10,938
Diciembre	258,579	99,802	5,098

Figura 28. Informe de uso por mes



4.6.2 Informe de uso por día de la semana y por hora

A continuación se presenta un informe de uso del sitio web por parte de los usuarios. Este informe se encuentra detallado por día de la semana y por hora del día.

Tabla IV. Informe de uso por día de la semana

Día	Peticiones	Páginas Vistas	Visitantes
Domingo	585,269	244,886	8,230
Lunes	1,388,787	569,618	13,233
Martes	1,388,557	592,352	14,311
Miércoles	1,322,003	571,432	13,938
Jueves	1,393,448	605,106	13,845
Viernes	1,300,004	544,480	12,279
Sábado	903,771	370,250	9,335

Figura 29. Informe de uso por día de la semana

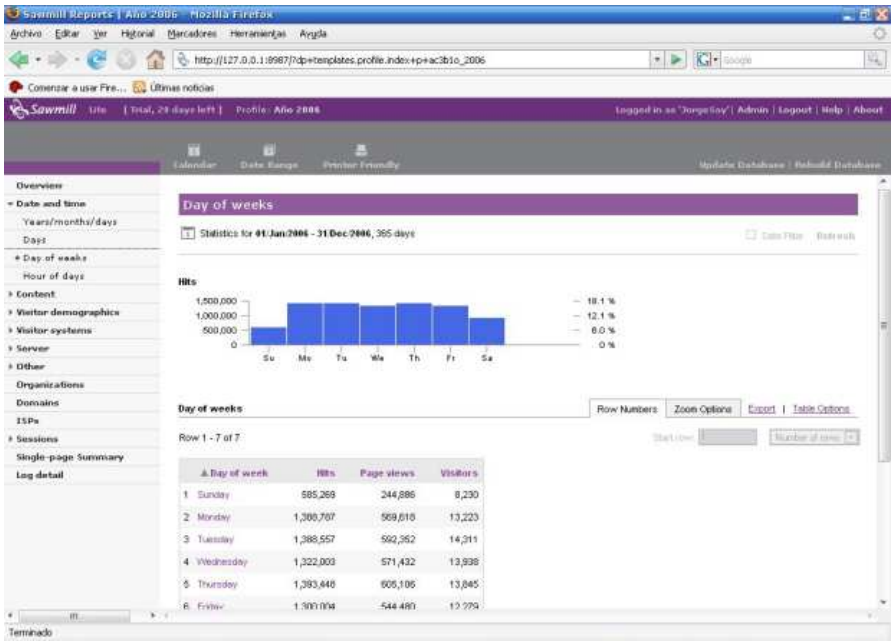
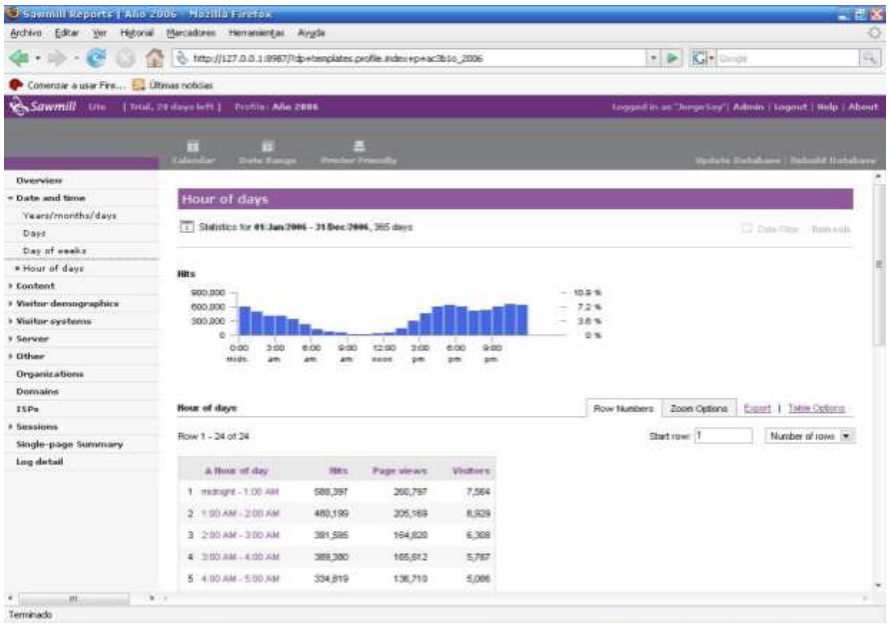


Tabla V. Informe de uso por hora del día

Hora	Peticiones	Páginas Vistas	Visitantes
12:00 a.m – 1:00 a.m.	588,397	260,797	7,564
1:00 a.m – 2:00 a.m.	480,199	205,169	6,929
2:00 a.m – 3:00 a.m.	391,585	164,820	6,308
3:00 a.m – 4:00 a.m.	389,380	165,612	5,787
4:00 a.m – 5:00 a.m.	334,819	136,710	5,086
5:00 a.m – 6:00 a.m.	221,950	86,750	3,784
6:00 a.m – 7:00 a.m.	109,103	39,842	2,308
7:00 a.m – 8:00 a.m.	53,877	19,771	1,324
8:00 a.m – 9:00 a.m.	37,798	17,638	885
9:00 a.m – 10:00 a.m.	20,063	6,923	722
10:00 a.m – 11:00 a.m.	18,824	6,850	994
11:00 a.m – 12:00 p.m.	25,863	9,642	966
12:00 p.m – 1:00 p.m.	50,347	18,173	1,342
1:00 p.m – 2:00 p.m.	125,416	48,555	2,219
2:00 p.m – 3:00 p.m.	289,878	110,719	3,537
3:00 p.m – 4:00 p.m.	444,481	178,814	4,836
4:00 p.m – 5:00 p.m.	590,006	254,442	5,781
5:00 p.m – 6:00 p.m.	632,584	270,854	6,355
6:00 p.m – 7:00 p.m.	592,736	258,936	6,227
7:00 p.m – 8:00 p.m.	506,392	214,063	6,111
8:00 p.m – 9:00 p.m.	521,045	217,437	6,361
9:00 p.m – 10:00 p.m.	584,868	248,244	7,063
10:00 p.m – 11:00 p.m.	639,490	280,126	7,459
11:00 p.m – 12:00 a.m.	632,738	277,237	7,700

Figura 30. Informe de uso por hora del día



4.6.3 Informe de directorios consultados

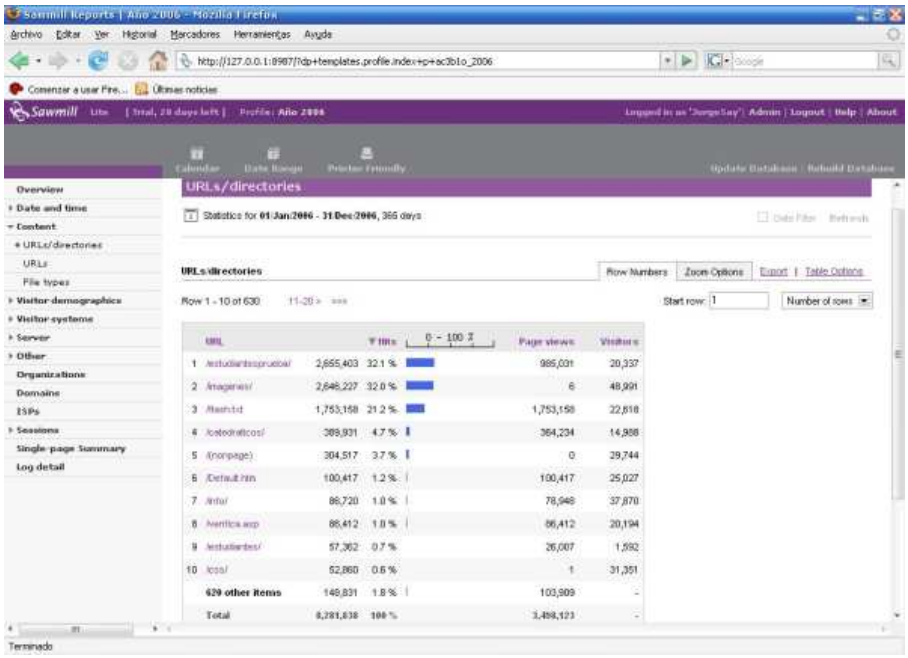
A continuación se presenta una parte de los directorios consultados dentro del servidor. A continuación unos ejemplos.

Tabla VI. Informe de directorios consultados

Directorio	Peticiones	Porcentaje	Páginas Vistas	Visitantes
/estudiantesprueba/	2,655,403	32.1	985,031	20,337
/imágenes/	2,646,227	32.0	6	48,991
/flash.txt	1,753,158	21.2	1,753,158	22,618
/catedráticos/	389,931	4.7	364,234	14,988
/(nonpage)	304,517	3.7	0	29,744
/Default.htm	100,417	1.2	100,417	25,027

/info/	86,720	1.0	78,948	37,870
/verifica.asp	86,412	1.0	86,412	20,194
/estudiantes/	57,362	0.7	26,007	1,592
/css/	52,860	0.6	1	31,351

Figura 31. Informe de directorios consultados



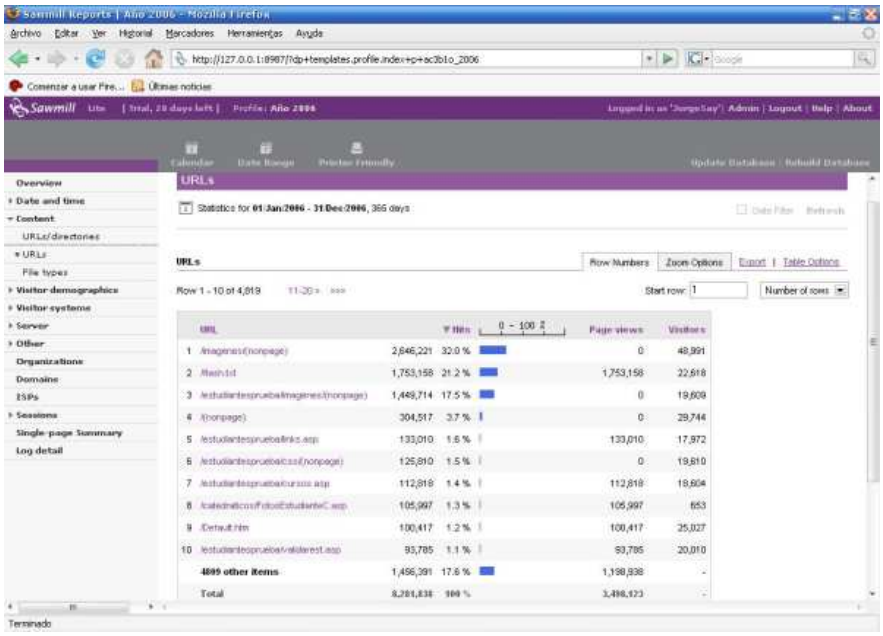
4.6.4 Informe de direcciones URL consultadas

A continuación se presenta una parte de las direcciones URL consultadas dentro del servidor. A continuación unos ejemplos:

Tabla VII. Informe de direcciones URL consultadas

URL	Peticiones	Porcentaje	Páginas Vistas	Visitantes
/imagenes/(nonpage)	2,646,221	32.0	0	48,991
/flash.txt	1,753,158	21.2	1,753,158	22,618
/estudiantesprueba/imagenes/(nonpage)	1,449,714	17.5	0	19,609
/(nonpage)	304,517	3.7	0	29,744
/estudiantesprueba/links.asp	133,010	1.6	133,010	17,972
/estudiantesprueba/css/(nonpage)	125,810	1.5	0	19,610
/estudiantesprueba/cursos.asp	112,818	1.4	112,818	18,604
/catedraticos/FotosEstudianteC.asp	105,997	1.3	105,997	653
/Default.htm	100,417	1.2	100,417	25,027
/estudiantesprueba/validarest.asp	93,785	1.1	93,785	20,010

Figura 32. Informe de direcciones URL consultadas



4.6.5 Informe de archivos consultados

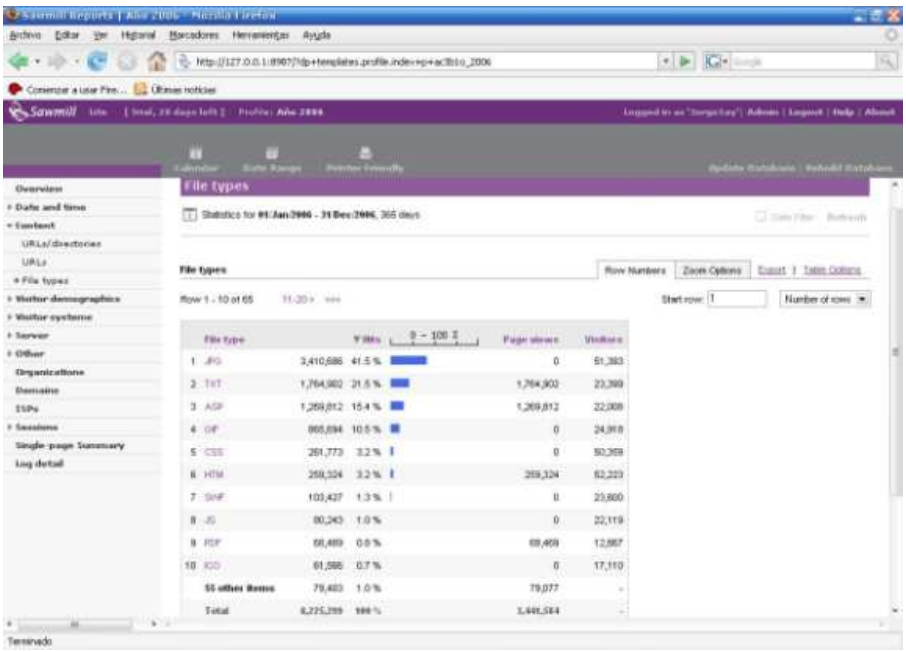
A continuación se presenta una parte de los tipos de archivo consultados dentro del servidor. A continuación unos ejemplos.

Tabla VIII. Informe de archivos consultados

Tipo Archivo	Peticiones	Porcentaje	Páginas Vistas	Visitantes
JPG	3,410,686	41.5	0	51,393
TXT	1,764,902	21.5	1,764,902	23,399
ASP	1,269,812	15.4	1,269,812	22,008
GIF	865,694	10.5	0	24,198
CSS	261,773	3.2	0	50,359
HTM	259,324	3.2	259,324	52,223
SWF	103,427	1.3	0	23,600
JS	80,243	1.0	0	22,119

PDF	68,469	0.8	68,469	12,667
ICO	61,566	0.7	0	17,110

Figura 33. Informe de archivos consultados



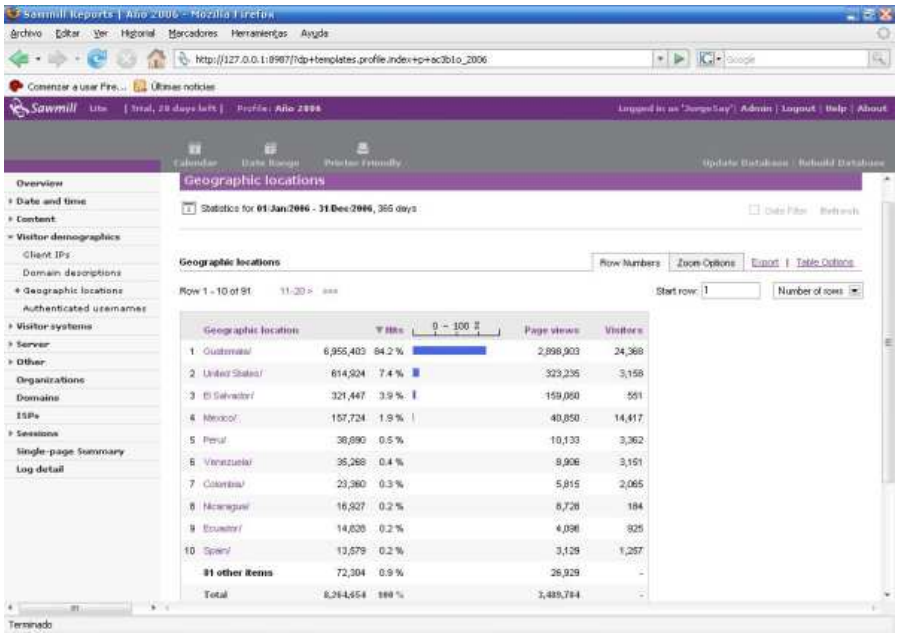
4.6.6 Informe de localización geográfica

A continuación se presenta un informe de localización geográfica que muestra los países que más han consultado el sitio web. A continuación unos ejemplos.

Tabla IX. Informe de localización geográfica

País	Peticiones	Porcentaje	Páginas Vistas	Visitantes
Guatemala	6,955,403	84.2	2,898,903	24,368
Estados Unidos	614,924	7.4	323,235	3,158
El Salvador	321,447	3.9	159,060	551
México	157,724	1.9	40,850	14,417
Perú	38,890	0.5	10,133	3,362
Venezuela	35,268	0.4	8,906	3,151
Colombia	23,360	0.3	5,815	2,065
Nicaragua	16,927	0.2	8,728	184
Ecuador	14,828	0.2	4,096	925
España	13,579	0.2	3,129	1,257

Figura 34. Informe de localización geográfica



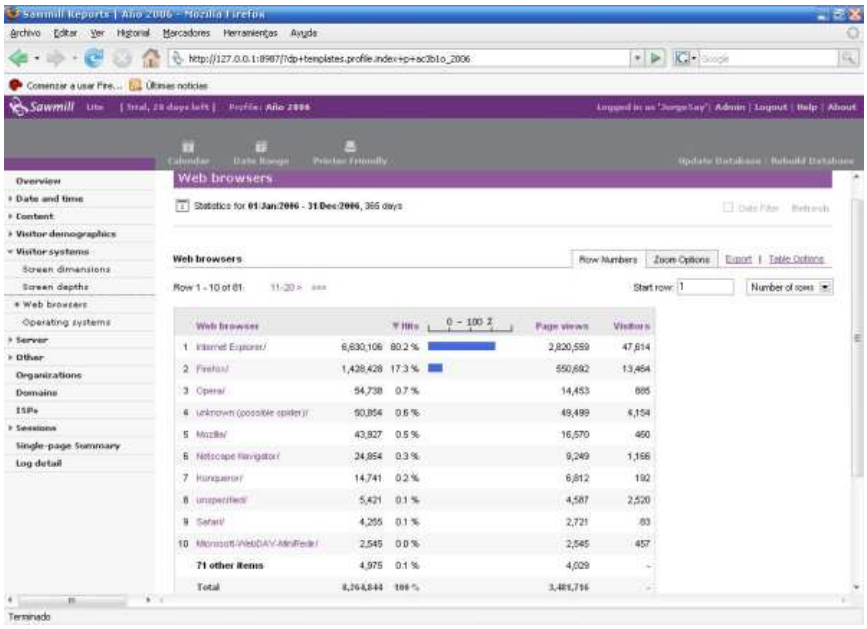
4.6.7 Informe de navegadores web

A continuación se presenta un informe que resume los navegadores web más usados por parte de los usuarios para acceder al sitio web.

Tabla X. Informe de navegadores web

Browser	Peticiones	Porcentaje	Páginas Vistas	Visitantes
Internet Explorer	6,632,998	80.3	2,822,730	47,636
Firefox	1,428,715	17.3	550,979	13,464
Opera	55,274	0.7	14,708	921
Mozilla	43,930	0.5	16,573	460
Netscape	24,976	0.3	9,351	1,181
Konqueror	14,745	0.2	6,816	192
Safari	4,255	0.1	2,721	83

Figura 35. Informe de navegadores web



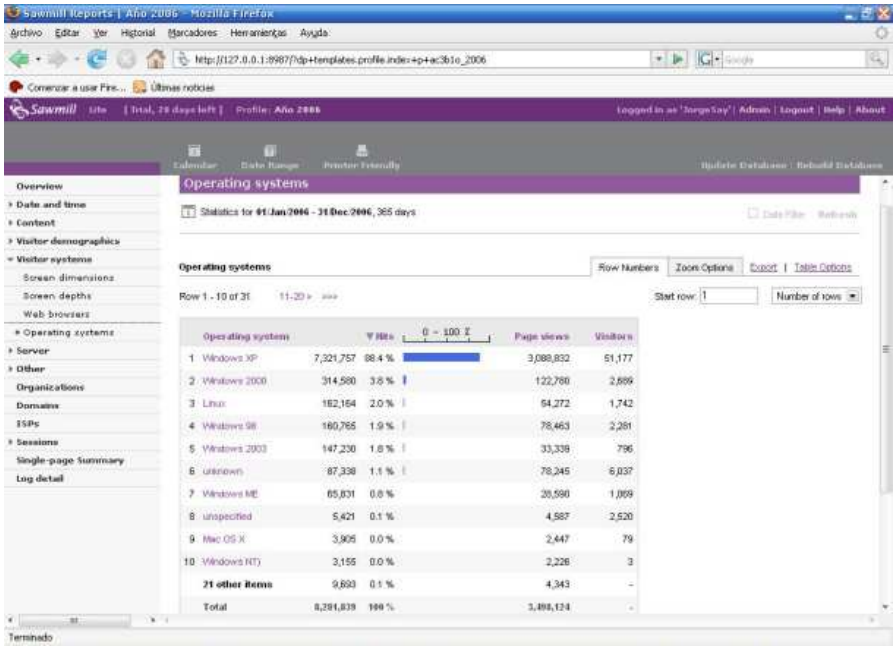
4.6.8 Informe de sistemas operativos

A continuación se presenta un informe que resume los sistemas operativos con mayor uso frecuente por parte de los usuarios para acceder al sitio web.

Tabla XI. Informe de sistemas operativos

Sistema Operativo	Peticiones	Porcentaje	Páginas Vistas	Visitantes
Windows XP	7,325,035	88.4	3,091,313	51,222
Windows 2000	314,810	3.8	122,992	2,717
Linux	162,205	2.0	54,307	1,745
Windows 98	160,801	1.9	78,499	2,286
Windows 2003	147,440	1.8	33,369	796
Windows ME	65,854	0.8	28,609	1,070
Mac OS X	3,905	0.1	2,447	79
Windows NT	3,155	0.1	2,226	3
Power Macintosh	1,910	0.1	926	44
Windows Vista	730	0.2	251	16

Figura 36. Informe de sistemas operativos



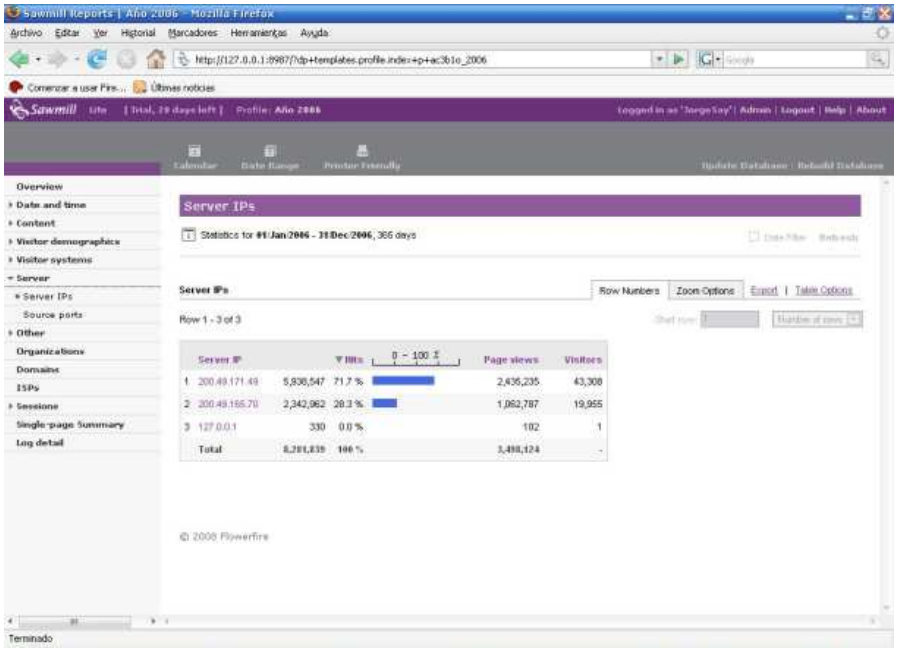
4.6.9 Informe de servidores IP

A continuación se presenta un informe de los servidores IP dentro del sitio web junto con las peticiones resueltas por los mismos.

Tabla XII. Informe de servidores IP

Servidor IP	Peticiones	Porcentaje	Páginas Vistas	Visitantes
200.49.171.49	5,938,547	71.7	2,435,235	43,308
200.49.165.70	2,342,962	28.2	1,062,787	19,955
127.0.0.1	330	0.1	102	1

Figura 37. Informe de servidores IP



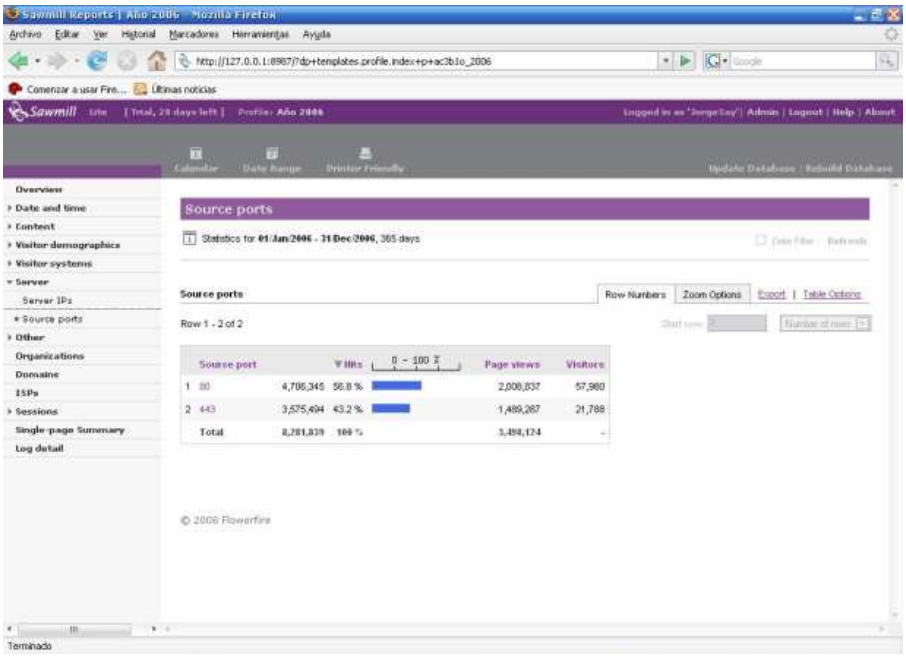
4.6.10 Informe de puertos

A continuación se presenta un informe que muestra los puertos por los cuales se ingreso al sitio web al momento de consultarlo.

Tabla XIII. Informe de puertos

Puerto	Peticiones	Porcentaje	Páginas Vistas	Visitantes
80	4,706,345	56.8	2,008,837	57,980
443	3,575,494	43.2	1,489,287	21,788

Figura 38. Informe de puertos



4.6.11 Informe de spider

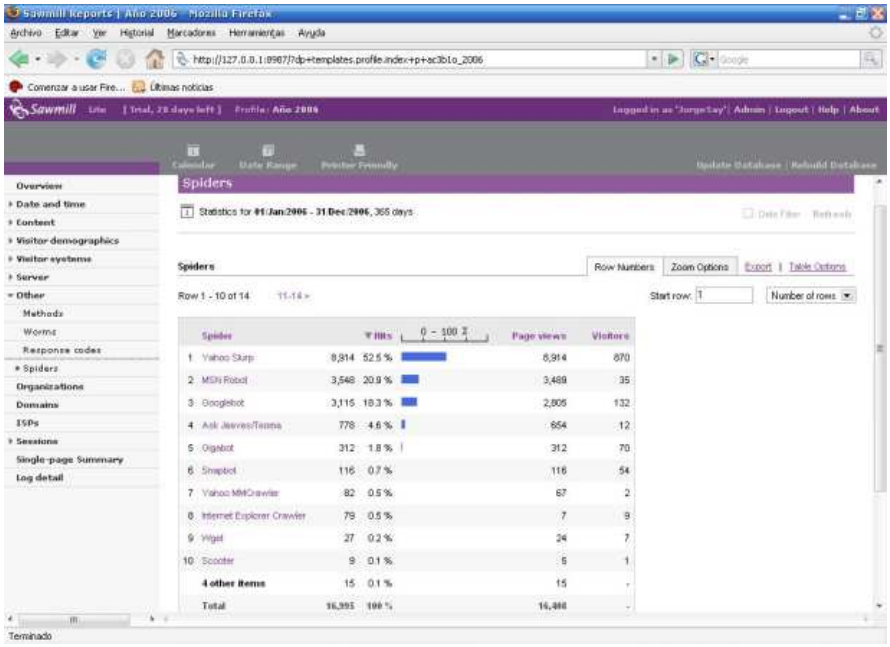
A continuación se presenta un informe de Spiders. Los Spiders representan una lista de los motores de búsqueda empleados y la relación de las páginas visitadas. A continuación algunos ejemplos.

Tabla XIV. Informe de Spiders

Spider	Peticiones	Porcentaje	Páginas Vistas	Visitantes
Yahoo Slurp	8,914	52.5	8,914	870
MSN Robot	3,548	20.9	3,489	35
Googlebot	3,115	18.3	2,805	132
As Jeeves/Teoma	778	4.6	654	12
Gigabot	312	1.8	312	70
Snapbot	116	0.7	116	54
Yahoo MMCrawler	82	0.5	67	2

Internet Explorer Crawler	79	0.5	7	9
Wget	27	0.2	24	7
Scooter	9	0.1	5	1

Figura 39. Informe de Spiders



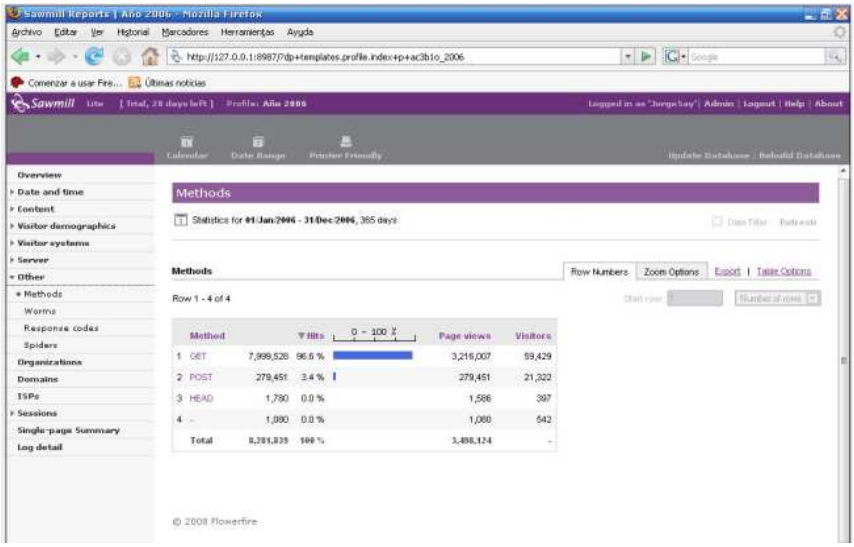
4.6.12 Informe de métodos

A continuación se presenta un informe de los métodos que son empleados en el sitio web.

Tabla XV. Informe de métodos

Servidor IP	Peticiones	Porcentaje	Páginas Vistas	Visitantes
GET	7,999,528	96.6	3,216,007	59,429
POST	279,451	3.4	279,451	21,322
HEAD	1,780	0.0	1,586	397

Figura 40. Informe de métodos



4.6.13 Informe de caminos a través de una página

A continuación se presenta un informe que muestra los caminos que puede seguir un usuario a través de una página dentro del sitio web.

Figura 41. Informe de caminos a través de una página

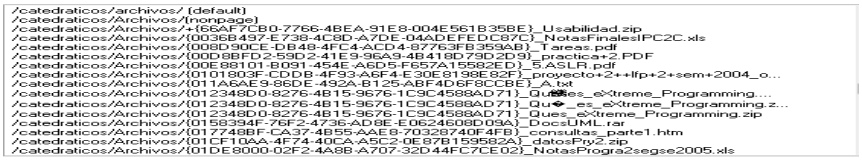
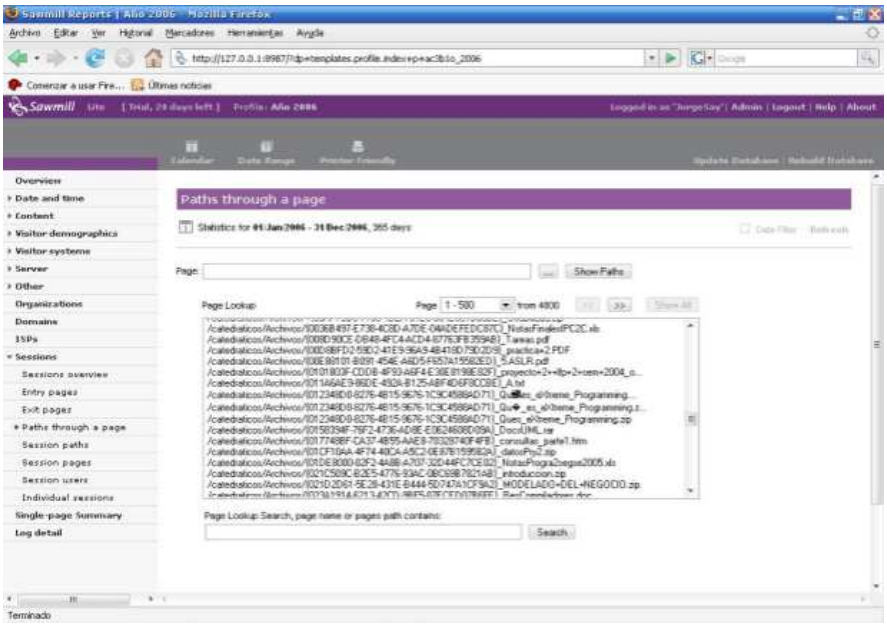


Figura 42. Vista del informe de caminos a través de una página



4.6.14 Informe de caminos en una sesión

A continuación se presenta un informe que muestra el recorrido que un usuario realiza durante una sesión. Este informe es útil al tratar de entender el comportamiento del usuario dentro del sistema.

Figura 43. Informe de caminos en una sesión

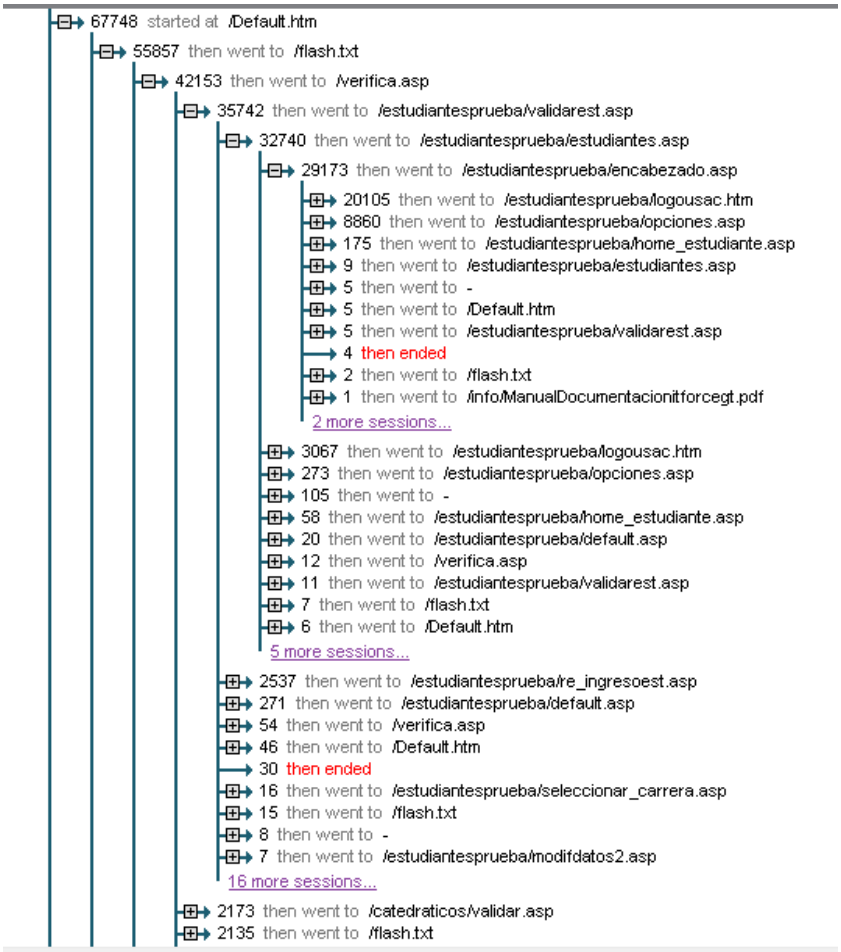
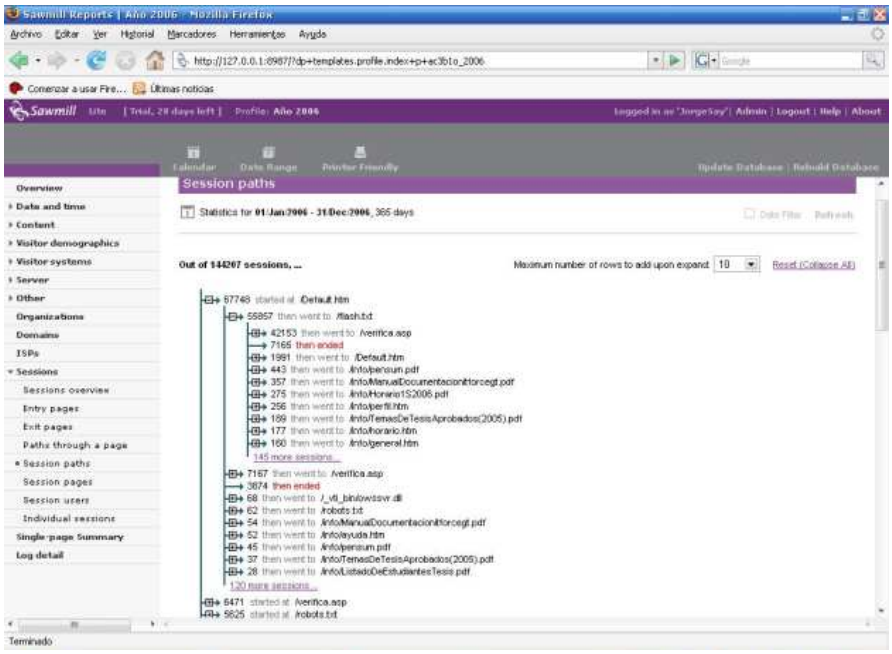


Figura 44. Vista del informe de caminos en una sesión



4.6.15 Informe de sesiones individuales

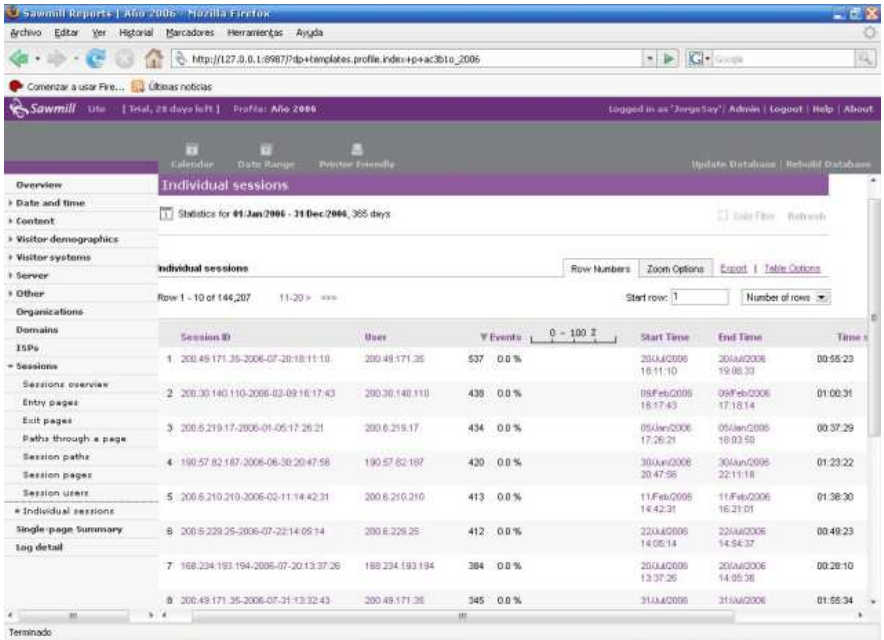
A continuación se presenta un informe que muestra todas las sesiones registradas que han tenido acceso al sistema. A continuación unos ejemplos.

Tabla XVI. Informe de sesiones individuales

ID de Sesión	Usuario	Hora Inicio	Hora Salida	Tiempo
200.49.171.35-2006-07-20:18:11:10	200.49.171.35	20/Jul/2006 18:11:10	20/Jul/2006 19:06:33	00:55:23
72.252.57.236-2006-08-11:13:40:10	72.252.57.236	11/Aug/2006 13:40:10	11/Aug/2006 15:23:39	01:43:29
69.79.114.206-2006-05-25:04:46:48	69.79.114.206	25/May/2006 04:46:48	25/May/2006 05:46:06	00:59:18

200.6.212.15-2006-02-13:17:59:47	200.6.212.15	13/Feb/2006 17:59:47	13/Feb/2006 19:41:26	01:41:39
72.252.46.200-2006-09-17:05:01:51	72.252.46.200	17/Sep/2006 05:01:51	17/Sep/2006 05:35:47	00:33:56
200.49.169.198-2006-03-17:16:53:24	200.49.169.198	17/Mar/2006 16:53:24	17/Mar/2006 17:33:20	00:39:56
190.57.92.39-2006-06-23:21:54:57	190.57.92.39	23/Jun/2006 21:54:57	23/Jun/2006 22:45:41	00:50:44

Figura 45. Informe de sesiones individuales



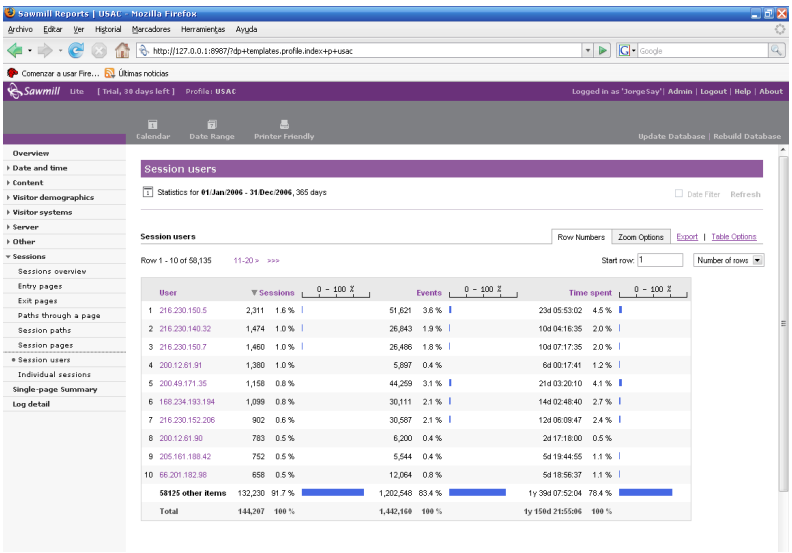
4.6.16 Informe de sesiones de usuario

A continuación se presenta un informe de sesiones de usuarios que han accedido al sistema. Se indican el tiempo total y el número de eventos realizados. A continuación unos ejemplos.

Tabla XVII. Informe de sesiones de usuario

Usuario	No Total Sesiones	No eventos	Tiempo Total Empleado
216.230.150.5	2,311	51,621	23d 05:53:02
216.230.140.32	1,474	26,843	10d 04:16:35
168.234.193.194	1,099	30,111	14d 02:48:40
66.201.182.98	658	12,064	5d 18:56:37
168.234.194.125	628	13,502	4d 23:37:41
72.30.252.140	290	388	01:07:25
200.12.61.91	1,380	5,897	6d 00:17:41

Figura 46. Informe de sesiones de usuario



4.6.17 Informe general de sesiones

A continuación se presenta un informe que resume todas las sesiones registradas en el sitio web.

Figura 47. Informe general de sesiones

	All days	Average per day
Total accesses	1,442,160	3,951.12
Total sessions	144,207	395.09
Sessions by one-time users	42,145	-
Sessions by repeat users	102,062	-
Total session users	58,135	159.27
One-time users	42,145	-
Repeat users	15,990	-
Two-time users	7,625	-
Three-time users	3,244	-
Four-time users	1,687	-
Five-time users	922	-
Six+-time users	2,512	-
Total duration of all sessions	1y 150d 21:55:06	-
Average accesses per session	10.00	-
Average sessions per user	2.48	-
Median sessions per user	1.00	-
Maximum concurrent sessions	17	-
Average session duration	00:05:09	-

CONCLUSIONES

1. La minería web es una rama de la minería de datos que consiste en aplicar técnicas de análisis sobre los ficheros históricos de un sitio web. Los conceptos y temas desarrollados en el trabajo forman parte de la teoría de la minería web. El trabajo está orientado a la minería de uso web, que es una clasificación dentro de la minería web.

2. Los sitios web cada día se encuentran con un constante aumento en sus volúmenes de información. Este panorama conduce a que el control, manejo, análisis y uso de la información sea más difícil. Es aquí que la minería web ayuda a procesar y analizar más rápido los datos. Los resultados de la minería web sirven de apoyo para tomar decisiones de cambios dentro del sitio web.

3. El proceso de minería de uso web consiste de una serie de etapas que van desde la recolección de datos hasta el análisis de patrones. Para descubrir patrones se utilizan algoritmos que son los encargados de encontrar comportamientos, tendencias y preferencias de los usuarios. Los principales algoritmos empleados por la minería de uso web se detallan en la etapa de descubrimiento de patrones.

4. En la parte práctica del trabajo se aplicó la minería web a un sitio de estudio por medio del análisis de los ficheros históricos. Para procesar la información se empleo una herramienta que facilita el análisis de los datos por medio de reportes gráficos. Los resultados obtenidos del análisis del sitio dan una pauta de la situación del mismo y que aspectos requieren atención para mejorar la calidad del sitio web.

RECOMENDACIONES

1. El trabajo presentado tiene conceptos y definiciones un tanto abstractas para gente que no está muy relacionada con análisis de información. Para comprender de una mejor manera la minería web se recomienda un previo estudio acerca de la minería de datos. Ya que a partir de la minería de datos se desprende la rama del tema estudiado en el trabajo.

2. La minería web es un área que aplicada correctamente en los sitios ofrece beneficios de análisis, reestructuración y personalización. Es por esto que se recomienda un estudio acerca de la tecnología con que se cuenta antes de su aplicación. Además se debe tener muy en cuenta los pros y contras que conlleva la minería web. Esto con el fin de evitar problemas y situaciones inesperadas durante el proceso.

3. Actualmente en el mercado existe una gran cantidad de soluciones personalizadas que llevan a cabo el proceso de minería web. Para aplicar una de estas soluciones se recomienda realizar un análisis del entorno del sitio web. Esto se sugiere con el fin de seleccionar la herramienta que más se apegue a las necesidades del sitio para así evitar el uso de herramientas complicadas y poco manejables.

BIBLIOGRAFÍA

1. Berson Alex, Smith Stephen J., McGraw Hill. "Data Warehouse, Data Mining and OLAP" Estados Unidos (1997)

2. Ramírez Quintana, María José; Hernández Orallo, José. "Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software". España (2003)

3. Kantardzic M., "Data Mining: Concepts, Models, Methods and Algorithms", John Wiley & Sons, (2003). ISBN: 0471228524

4. Calderón Méndez, Neftalí de Jesús (2006), "Minería de datos una herramienta para la toma de decisiones". Escuela de Ingeniería en Ciencias y Sistemas, Universidad de San Carlos de Guatemala. Guatemala, Guatemala. Pág. 6 -18, 27 - 38, 42 - 44, 49 -51.

5. Cuello, Gabriel (2006). "Técnicas de minería de datos dentro de contextos metodológicos y de empresa". Escuela de postgrado, Instituto tecnológico de Buenos Aires. Buenos Aires, Argentina.

6. Vallejos, Sofia J., (2006). "Minería de datos". Facultad de ciencias exactas, naturales y agrimensura, Universidad Nacional del Nordeste. Argentina.

7. Román, Ulises; Alarcón, Luis. (2005). "Minería de uso web para predicción de usuarios en la universidad". ISBN: 1816-3823 (Versión electrónica). http://sisbib.unmsm.edu.pe/BibVirtualData/publicaciones/risi/n3_2005/a01.pdf

8. “Data Mining Techniques”. http://www.statsoftinc.com/textbook/stdatmin.html (2 de octubre 2008)

9. “Cookie*”.* http://es.wikipedia.org/wiki/Cookie (15 de diciembre 2008)

10. DAEDALUS – Data, Decisions and Language, S.A. (Noviembre 2002). “Web Mining”. http://www.daedalus.es (5 de octubre 2008)

11. Fuentes Reyes, Sady C.; Ruiz Lovaina, Marina (2007). “Minería Web: un recurso insoslayable para el profesional de la información”. http://bvs.sld.cu/revistas/aci/vol16_4_07/aci111007.htm (28 de agosto 2008)

12. de Gyves Camacho, F. M., “Web Mining: Fundamentos básicos”. http://zarza.usal.es/~fgarcia/doctorado/iweb/05-07/Trabajos/WMINING.pdf (10 de septiembre 2008)

13. Montes y Gómez M., “Minería de texto: Un nuevo reto computacional”. http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf (29 de agosto 2008)

14. Dürsteler, Juan C. (2005). “Minería Web”. http://www.infovis.net/printMag.php?num=172&lang=1 (29 de agosto 2008)

15. Sánchez Sánchez, Jorge. “*Análisis de accesos a un servidor web de contenidos dinámico* ”. http://jordisan.net/pfc/Memoria_PFC_web_mining_Jordi.pdf (10 de octubre 2008)

16. Galeas, Patricio. “*Web Mining*”. http://galeas.de/webmining.html (20 de septiembre 2008)

17. Eirinaki, Magdalini; Vazirgiannis, Michalis. "*Web Mining for Web Personalization".* http://www.engr.sjsu.edu/meirinaki/papers/EV03_TOIT.pdf (20 de diciembre 2008)

18. González Torres, Antonio. "*Minería web y personalización: Revisión bibliográfica y propuesta de un marco de referencia".* http://zarza.usal.es/~fgarcia/doctorado/iweb/05-07/Trabajos/MineriaWeb%20y%20Personalizacion.pdf (4 de enero 2009)

19. "*Minería de Texto: Recuperación y organización de la información".* http://mineriatexto.50webs.com/recuperacion-informacion.html (18 de febrero 2009)

20. Koblinc, D. Gustavo; Vazirgiannis, Michalis. "*Web Mining: Estado Actual de Investigación".*http://www2.ing.puc.cl/gescopp/Sergio_Maturana/SAG/Webmining.html (25 de febrero 2009)

ANEXOS

ANEXO A – Minería de texto

Lingüística computacional

La lingüística computacional se encarga del estudio y aplicación de métodos y técnicas de las ciencias de la computación para la comprensión del lenguaje. El objetivo de la lingüística computacional es el entendimiento automático de contenidos de texto. La lingüística computacional es una ciencia de la cual se desprenden muchas ramas y la minería de texto es una de sus principales derivaciones.

Minería de texto

La Minería de texto surge como una iniciativa en la búsqueda de regularidades, conocimientos y patrones que puedan existir en una colección texto. Esta ciencia es una de las más recientes innovaciones en lo que respecta a la investigación de procesamiento de caracteres de texto. La minería de texto es la encargada del proceso de descubrimiento de conocimiento útil dentro de bloques de información específicamente de texto. Para realizar esta labor la minería de texto emplea técnicas de aprendizaje automático.

En diversas ocasiones se tiene una idea equivocada ya que muchas personas tienden a confundir la minería de texto con el concepto de recuperación de información. La recuperación de información se enfoca en la recuperación automática de archivos de texto sobresalientes a través de clasificaciones, enlaces de textos, agrupaciones, etcétera. Por lo regular la recuperación de información emplea el uso de palabras clave con la finalidad de hallar páginas relevantes en

su contenido. Mientras que por otra parte totalmente diferente la minería de texto se dedica a examinar detalladamente una colección de archivos de texto con el objetivo de descubrir información y conocimiento que no se encuentra contenido en algún archivo individual.

Proceso de minería de texto

El proceso que la minería de texto sigue para descubrir conocimiento en los archivos de texto es:

A. Adquisición y recopilación de archivos de texto.

B. Normalización de los archivos de texto.
   - Normalizar en formato XML.
   - Extraer metadatos.

C. Proceso de filtración: Registrar archivos con textos relevantes por medio de un análisis con palabras clave.

D. Proceso de análisis: Consiste en el establecimiento de todas las relaciones existentes entre los textos a partir de categorías y términos.

E. Visualización de resultados: A través de esquemas, diagramas, visualización gráfica, tablas, etcétera.

Técnicas de minería de texto

Técnicas clásicas

Las técnicas clásicas de minería de texto se dividen en tres etapas que se describen a continuación.

- Etapa de pre procesamiento: Esta etapa consiste en procesos mediante los cuales los bloques de textos se convierten en un tipo de representación jerárquica ó estructurada que ayude a facilitar su análisis.

- Etapa de representación: La etapa de representación depende muchas veces de la técnica que se halla empleado previamente durante la etapa de pre procesamiento. Las técnicas empleadas dan la pauta de cuáles serán los algoritmos que se emplearan para el descubrimiento de conocimiento.

- Etapa de descubrimiento: Esta etapa consiste en algoritmos que partiendo de la representación organizada y estructurada proporcionan información. Estos algoritmos tienen la capacidad de encontrar patrones de regularidad en los archivos de texto.

Como se puede notar las diferentes etapas se encuentran ligadas entre sí. Las técnicas clásicas más empleadas en la minería de texto son la secuencia de palabras que encuentra patrones en los bloques de texto, las tablas resumidas de datos que encuentran relaciones entre textos y los vectores de temas que presentan el nivel temático del texto.

Grafos conceptuales

Las técnicas antecesoras de los grafos conceptuales no podían resolver preguntas referentes a consensos, desviaciones y tendencias. Un grafo conceptual se pude considerar un grafo bipartito que se encuentra formado por dos tipos de nodo que son conceptos y relaciones conceptuales. Los grafos conceptuales se consideran dentro de la minería de texto una de las técnica de representación más avanzadas.

Los grafos conceptuales se pueden comparar empleando conocimiento acerca del dominio como por ejemplo niveles de jerarquías de los conceptos y diccionarios de sinónimos entre otros. La agrupación de más de dos grafos permite encontrar una estructura no visible de los bloques de textos. Se lleva a cabo una intersección entre dos grafos con el objetivo de brindar una síntesis de ambos grafos. A esta síntesis se le asigna un valor de puntuación que determina el grado de comparación y similitud entre los dos textos.

Para agrupar varios grafos se emplean técnicas de agrupación entre las cuales se puede mencionar comweb, agrupamiento de k-medias y estrategias colaborativas. Las técnicas antes mencionadas proporcionan a los algoritmos una colección de ejemplos con los cuales se llegan a generar las respectivas agrupaciones. A continuación se muestra cómo se observaría la frase “Bush crítica a Zapatero” representada por medio de un grafo conceptual.

Figura 48. Grafo conceptual

Fuente: http://mineriatexto.50webs.com/recuperacion-informacion.html

Programación lógica inductiva

La técnica de programación lógica inductiva permite introducir un temprano conocimiento a priori sobre el dominio representado mediante definiciones por medio de predicados que estén relacionados. La programación lógica inductiva tiene la ventaja de poseer una capacidad de representación que se basa en la lógica de segundo orden. Esta ventaja permite una generalización de conceptos y el descubrimiento automático de definiciones de conceptos.

Regularmente se emplea el lenguaje "*Prolog*" para programar las herramientas. La programación lógica inductiva necesita aparte de un conjunto de entrenamiento relaciones anteriormente encontradas por el diseñador que se basen en las clausulas de Horn.

Un ejemplo de la programación lógica inductiva es que si se quisiera encontrar la definición del término abuelo(a, b) partiendo de los conceptos de padre(a, b) y madre (a, b) antes descritos de manera extensiva. Daría como resultado una definición parecida a: abuelo(a, b) = padre(a, b) and madre(a, b).

Programación genética

La programación genética es una técnica que consiste en generar programas de manera automática con adaptación evolutiva para un ordenador. La programación genética empieza a través de programas sencillos que por medio del cruzamiento con otros más y de una mutación aleatoria genera programas mejor adaptados en la ejecución de la tarea que se le presente.

La aptitud de todos los programas se evalúa de manera numérica a través de una función llamada "fitness". La programación genética cuenta con

extensiones donde algunas de estas permiten explicar nuevas primitivas a partir de las primitivas iníciales. En la programación genética surgen problemas cuando las definiciones son recursivas. Esto se atribuye a que la programación genética carece de eficiencia en las primitivas recursivas.

La idea central de la programación genética es introducir como primitivas aquellas relaciones expresadas del tipo de clausula de Horn y emplear sistemas que permitan usar evoluciones paralelas de otras primitivas con la finalidad de generar definiciones lo más resumidas posibles sobre los conceptos. La tarea de la función fitness es la medición del número de ejemplos que se abarcan con la definición de cada individuo generado en cada nueva población (generación). Cada individuo se pondera con el tamaño de la definición para que el algoritmo pueda dar con soluciones más cortar y sencillas.

ANEXO B – Métodos, modelos y algoritmos de minería de uso web

A continuación se presentan los métodos, los modelos, las técnicas y los algoritmos que son empleados en los diferentes procesos de minería de uso web y de personalización de sistemas web.

Tabla XVIII. Métodos, modelos y algoritmos de minería de uso web

Métodos	Modelos	Técnicas ó Algoritmos
Modelos Básicos	Modelos para texto	➢ Modelo del espacio vectorial ➢ Modelo binario ➢ Modelo polinómico
	Modelos para hipertexto	Creación de grafos dirigidos a partir de los enlaces web
		AWA – Adaptive Window Algorithm
		Búsqueda primero en anchura
		Búsqueda primero en profundidad
		➢ Minería de subárbol mediante el algoritmo Freqt ➢ Induced y variantes
		Algoritmo TreeMinerV y variantes
		Algoritmo Induced UNOrdered
		FFSM
		gSpan
		AcGM
		FSG
		FTM
	Modelos para datos semiestructurados	Directorios de tópicos
		Modelos probabilísticos
		Descubrimiento de asociaciones estructurales
		Descubrimiento de patrones
		Técnicas basadas en grafos
Aprendizaje Supervisado	Modelos probabilísticos para aprendizaje de texto	Clasificación de Bayes
		Relajación de parámetros
		Redes bayesianas
		Máxima entropía
		Árboles de decisión
		Perceptrones
		SVM (Support Vector Machine)

	Métodos para relaciones de aprendizaje	Conexionistas
		Simbólicos
Aprendizaje no supervisado	Técnicas básicas de agrupación	k-means
		Aglomerativa
	Técnicas de álgebra lineal	Indexado semántico potencial
		Proyecciones aleatorias
Aprendizaje semisupervisado	Modelos generativos	Identificabilidad de los datos
		Correctitud del modelo
		EM – Expectation Maximization
		Clasificación y etiquetado
		Aprendizaje discriminativo mediante el uso del Kernel de Fisher
	Auto entrenamiento	
	Co-entrenamiento	
	Algoritmos para evitar cambios en regiones densas	TSVM –Transductive Support Vector Machine
		Procesos Gaussianos paralelos para TSVM
		Regularización de la información
		Minimización de la entropía
	Métodos basados en grafos	Construcción de grafos
		Regularización medio de grafos
		Inducción basada en grafos
		Consistencia de algoritmos basados en grafos
		Entrenamiento semisupervisado en datos estructurados
		Grafos dirigidos
		Conexión a modelos gráficos estándar
Análisis de redes sociales	Análisis de redes sociales aplicados a la web	PageRank
		HITS – Hyperlink Induced Topic Search: ➢ ARC – Automatic Resource Compilation ➢ Outlier Filtering

Fuente: http://zarza.usal.es/~fgarcia/doctorado/iweb/05-07/Trabajos/MineriaWeb%20y%20Personalizacion.pdf

ANEXO C - Métodos, modelos y algoritmos de minería de contenido web

A continuación se presentan los métodos, las técnicas, los procesos y los algoritmos que son empleados en los diferentes procesos de minería de contenido web.

Tabla XIX. Métodos, modelos y algoritmos de minería de contenido web

Método de Minería	Modelos	Proceso ó Algoritmos
Recuperación de Información	Análisis Automático del texto	Remover palabras que ocurren con una frecuencia muy alta
		Remover sufijos
		Detectar palabras con raíces equivalentes
		Realizar indexado
		Representación y discriminación de Documentos
		Clasificación automática de palabras claves y uso de tesauros
		Normalización
	Clasificación automática de palabras claves y documentos	K-medios
		Clasificación con el algoritmo QT
		Clasificación con el algoritmo difuso C-medios
	Estructuras de archivos	Archivos secuenciales
		Archivos secuenciales indexados
		Archivos invertidos
		Archivos multilistas
		Archivos celulares multilistas
		Archivos con estructuras de anillo
		Listas con múltiples procesos
		Archivos con direccionamiento asociativo
		Representaciones jerárquicas
		Representaciones de red
		Bases de datos relacionales

	Estrategias de búsqueda	Búsquedas booleanas
		Funciones de búsqueda de coincidencias
		Búsquedas secuenciales
		Búsquedas basadas en agrupaciones
		Formulación de búsquedas interactivas
		Búsquedas con mecanismos de retroalimentación
	Medidas de desempeño	Relevancia: ➢ Relevancia algorítmica ➢ Relevancia temática ➢ Pertinencia de la información ➢ Relevancia situacional ➢ Relevancia motivacional
		Precisión de la búsqueda de información
Extracción de Información	Preprocesamiento del documento	Creación de zonas de texto
		Separadores de unidades de texto
		Filtrado de texto
		Separación de unidades léxicas
		Analizadores léxicos
		Algoritmos para la resolución de ambigüedades
		Identificadores de las raíces de las palabras
		Lematizadores
	Análisis sintáctico completo o parcial	Coincidencia de patrones
	Interpretación semántica	Relaciones gramaticales
	Análisis verbal	
	Generación de plantillas	Relación de las piezas de información extraídas con el formato de salida deseado
Minería de la Estructura web	Generación de grafos dirigidos, asignación de pesos a los nodos y recorrido de los grafos	PageRank
		TrustRank
		HITS y sus variantes

Minería de Texto	Categorización de texto	SVM - Support Vector Machine
		Redes neuronales
		Algoritmos genéticos
		Sistemas de lógica difusa
		Clasificación bayesiana
		Árboles de decisión
	Agrupación de elementos	k-medios
	Aprendizaje inductivo	Reglas de asociación
		Aprendizaje simbólico
	Análisis de información	Análisis de la secuencia temporal
		Análisis estadístico

Fuente: http://zarza.usal.es/~fgarcia/doctorado/iweb/05-07/Trabajos/MineriaWeb%20y%20Personalizacion.pdf