PROTOTIPO PARA LA DETECCIÓN AUTOMÁTICA DE PERFILES DE NAVEGACIÓN DE USUARIOS DE UNA EMPRESA APLICANDO TÉCNICAS DE MINERÍA DE DATOS Y ALGORITMOS DE APRENDIZAJE.

Fredy Andrés Aponte Novoa

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
MAESTRÍA EN SOFTWARE LIBRE
BUCARAMANGA
2011

PROTOTIPO PARA LA DETECCIÓN AUTOMÁTICA DE PERFILES DE NAVEGACIÓN DE USUARIOS DE UNA EMPRESA APLICANDO TÉCNICAS DE MINERÍA DE DATOS Y ALGORITMOS DE APRENDIZAJE.

FREDY ANDRÉS APONTE NOVOA

Tesis

Director
Javier Hernández Cáceres
Magister

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
MAESTRÍA EN SOFTWARE LIBRE
BUCARAMANGA
2011

Nota de Aceptación:

_____

_____

_____

_____

_____

_____

Presidente del Jurado

_____

Firma del Jurado

_____

Firma del Jurado

Bucaramanga, Julio 18 de 2011

A mi madre, quien con su trabajo diario, hace de mí una persona íntegra, responsable pero ante todo un ser humano sensible al dolor y a las alegrías. Gracias por todos tus sacrificios y esfuerzos; Dios te bendiga y me permita compartir muchos años más de tu compañía y algún día poder recompensar todo lo que haces por mí.

A la memoria de mi padre, que desde el cielo me ha protegido y guiado y para no perder el horizonte. Siempre te llevare en mi corazón.

A mis tres lindas hermanas, Nubia, Paola y Mónica, con quienes he contado en distintas situaciones de mi vida y quienes cada día me dan un nuevo motivo para sentirme orgulloso de ellas.

A Alejita, quien ha sabido comprenderme durante mucho tiempo y siempre tiene una sonrisa para mí, su amor y confianza, me hacen cada día más fuerte y me alientan para seguir trabajando

A William el hermano que la vida me regalo. El cómplice de muchos proyectos profesionales y familiares.

A mis amigos y compañeros de oficina Nubia, Sandrita, Teresa, Yuri, Esneda, Leidy, Oscar, Héctor, Alirio. Por cada uno de ellos guardó un sentimiento muy profundo de cariño y aprecio y espero nunca perder su amistad.

Agradezco a Dios por darme las cualidades y habilidades para llevar a cabo este trabajo.

A la Universidad Autónoma De Bucaramanga por permitirme formar como Magíster en Software Libre.

A mi director, Magister Javier Hernández Cáceres, quien me asesoro durante el desarrollo del trabajo y me brindó su apoyo y colaboración

A Julián Monsalve, quien con su apoyo contribuyo a dar feliz término a este trabajo. Un sentimiento de verdadera gratitud para Julián quien me colaboró de forma incondicional y sin esperar reconocimiento alguno por su trabajo.

CONTENIDO

Pág.

LISTA DE TABLAS

Pág.

LISTA DE FIGURAS

Pág.

RESUMEN

En este trabajo se describe el desarrollo de un prototipo para la detección automática de perfiles de usuarios, el cual realiza recomendaciones a el administrador web teniendo en cuenta la navegación de los usuarios, aplicando técnicas de minería de datos.

El prototipo se compone de 5 etapas: carga de datos, preprocesamiento, sesionalizacion, identificación de perfiles y recomendaciones.

En la carga de datos, los registros de acceso son leídos desde el archivo Log generado por el servidor Web. Este archivo puede encontrarse en diferentes formatos pero el prototipo trabaja con archivos en formato NCSA generado por servidores Web Apache.

La etapa de preprocesamiento de datos, se divide en 3 tareas:-Selección. –Limpieza. –Transformación. En la primera tarea se analizan los atributos que conforman el archivo de log y se identifican cuales son relevantes en la identificación de perfiles de usuario.
La tarea de limpieza está orientada a eliminar todos los datos innecesarios para las siguientes etapas, como son los registros de robots, solicitudes no exitosas, objetos de páginas y registro de páginas índice. Como último paso en la transformación de los datos se realiza la identificación de perfiles de usuario.

La penúltima etapa, la sesionalizacion se encarga de identificar las sesiones de navegación de los usuarios, para esta tarea se maneja un umbral de tiempo para identificar las visitas correspondientes a una misma sesión.

La última etapa llamada etapa de recomendación, el sistema realiza una serie de informes usados por el administrador web para realizar mejorar al sitio web.

INTRODUCCION

El uso de tics y la automatización de procesos han marcado el éxito de muchas empresas, pero hoy en día esto no es suficiente, ahora este éxito radica en la toma de decisiones empresariales apropiadas, que se soportan con la información propia y externa, principalmente de empresas con las cuales se compite en el mercado.

Es por esto que muchas organizaciones han adoptado el internet como herramienta que permite el acceso a estos datos, independientemente de la situación geográfica en que se requieran, contribuyendo de forma significativa en las medidas adoptadas por la empresa y en consecuencia en el futuro de la organización.

Así mismo el internet se ha convertido en el medio de comunicación que permite a las organizaciones estar en contacto con todo el mundo en cualquier momento y a un bajo costo y por medio de este ofrecer sus productos y/o servicios. Es por esto y por servicios como correo electrónico, mensajería instantánea, foros, buscadores especializados, acceso remoto a quipos, entre otros, que Internet se ha masificado y hasta el punto de convertirse en la red de información más importante del mundo, la cual reúne millones de personas son importar la actividad que realicen.

Pero todo son beneficios, estos avances informáticos han ocasionado inconvenientes en relación a la cantidad de información publicada y a la variedad de fuentes que la proporcionan. En muchas ocasiones, esta gran cantidad de información contribuye a que los usuarios web pierdan mucho tiempo cuando el resultado de sus búsquedas no corresponde con sus necesidades.

Para un usuario no es fácil encontrar lo que realmente requiere, pues se presenta diversidad en cuanto a contenidos de los sitios sino de los usuarios, dado que estos no presentan el mismo comportamiento de navegación.

Para dar solución a estos inconvenientes, surge el concepto de web inteligentes cuya finalidad es dotar de inteligencia a internet, es decir, que los sitios web puedan aprender a partir del comportamiento de sus usuarios. El conocimiento adquirido en usado con el objetivo de reacondicionar el sitio a las necesidades de los usuarios, lo que se representa en un mejor servicio intuitivo a las necesidades del usuario.

El comportamiento de los usuarios en los sitios web es la base para el diseño y creación de sitios web inteligentes, ya que este comportamiento determina como se usa el sitio web, cuales son las necesidades y servicios utilizados por los usuarios. Es aquí donde surge la importancia de la minería de datos, concretamente la minería web, como herramienta para la extracción de conocimiento a partir de los datos que se tienen de la empresa e interacción con sus clientes.

La minería de datos se ha desarrollado como un área de conocimiento influenciada de forma directa por la probabilidad, estadística, el aprendizaje computacional y la inteligencia artificial, que busca dar solución a las necesidades de las organizaciones de procesar y analizar grandes volúmenes de información almacenados en bases de datos y archivos planos, para la extracción de conocimiento útil para la organización.

La minería web conserva todos los fines de la minería de datos, pero se concentra únicamente en el entorno Web y obtiene las fuentes de bases de datos, servidores proxy o de archivos del sitio o de los usuarios. Debido a esta variedad de fuentes, los datos en la web a utilizar en minería se pueden categorizar en datos de uso, de estructura y de contenido

Debido a esta diversidad de fuentes de datos en la Web, se pueden establecer tres categorías de datos a utilizar en la minería como son datos de uso, de estructura y de contenido. Los datos de uso hacen referencia a la forma como es utilizado el sitio web y se registran en archivos de log de acceso; los datos de estructura describen la organización del contenido

en las páginas, la jerarquía del sitio y los enlaces internos y externos, y los datos de contenido son los datos reales que se entregan al usuario como texto y elementos multimedia. De acuerdo a estos tipos de datos y con objetivo de mejorar su desempeño, la minería web divide su estudio en minería de la estructura, del contenido y de uso web.

Hoy en día los sitios web encuentran en la minería de uso web una opción para entender el comportamiento de navegación de sus usuarios representado en los archivos de log de acceso, y a partir de los registros allí alojados, adoptar estrategias para el mejoramiento del sitio web.

1. PLANTEAMIENTO DEL PROBLEMA

Los servidores web registran las visitas de sus usuarios en archivos de log de acceso, almacenan información relacionada con el acceso como la dirección IP del servidor, IP del cliente, fecha y hora de acceso, enlaces visitados, bytes transferidos, bytes recibidos, navegadores utilizados, sistemas operativos del cliente, entre otros.

Existen programas que analizan los archivos de log de los servidores web y presentan información resumida sobre el uso de un sitio web.

Esta información puede ayudar a los administradores para la conformación de alertas, la presentación de informes oportunos y, en general, a prestar un servicio de calidad.

Al desarrollo de este proyecto se pretende dar respuesta a la siguiente pregunta:

¿El análisis de la información registrada en los archivos de log de acceso del servidor web de una empresa y el desarrollo de una herramienta que estructure dicha información facilitará la toma de decisiones y la conformación de alertas e informes oportunos al administrador web?

2. JUSTIFICACIÓN

La minería de datos o Data Mining (DM) es una etapa del proceso de descubrimiento de información en bases de datos (Knowledge Discovery in Databases o KDD). Esta etapa integra las ventajas de áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo de información, usando como materia prima grandes volúmenes de datos.

El prototipo desarrollado servirá de ayuda a los administradores de sitios web que no dispongan de herramientas para analizar la información de los archivos log que les permita evaluar la calidad del servicio y en base a esto realizar mejoras de usabilidad en su sitio web.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Desarrollar un prototipo para crear perfiles de usuarios automáticamente, utilizando algoritmos de Minería de Datos y algoritmos de aprendizaje.

3.2 OBJETIVOS ESPECÍFICOS

- Evaluar los algoritmos que utiliza la minería de datos y de aprendizaje en la creación de perfiles de usuarios.

- Apoyar el área de automatización de los procesos de búsqueda y diseminación de información, en diferentes áreas, por medio del resultado de esta investigación.

- Desarrollar el prototipo, con sus respectivas pruebas, para evaluar el funcionamiento del mismo por medio de expertos.

4. ESTADO DEL ARTE

4.1 MINERÍA WEB

Sushmita Mitra y Tinku Achaya [MIT03] definen la minería Web como el uso de técnicas de minería de datos para recuperar, extraer y evaluar (generalizar o analizar) automáticamente información para descubrir conocimiento de documentos y servicios web. Según Gerd Stumme, Andreas Hotho y Bettina Berendt [STU02] la minería Web es la aplicación de técnicas de minería de datos al contenido, la estructura y el uso de recursos Web.

4.2 FASES DE LA MINERÍA WEB

[ROM05] La minería web tiene cuatro fases para la selección y transformación de los resultados.

- Localización documentos y servicios web: Recuperación de datos desde las fichas textuales de la web usando índices.

- Extracción automática de la información específica desde las fuentes: Incluye cualquier proceso de selección y de transformación de los datos originales.

- Descubrimiento de patrones generales desde los sitios web (proceso de generalización): se usan técnicas de agrupamiento (cluster), reglas de asociación, recuperación de información para extraer secuencia de patrones de navegación desde archivos log.

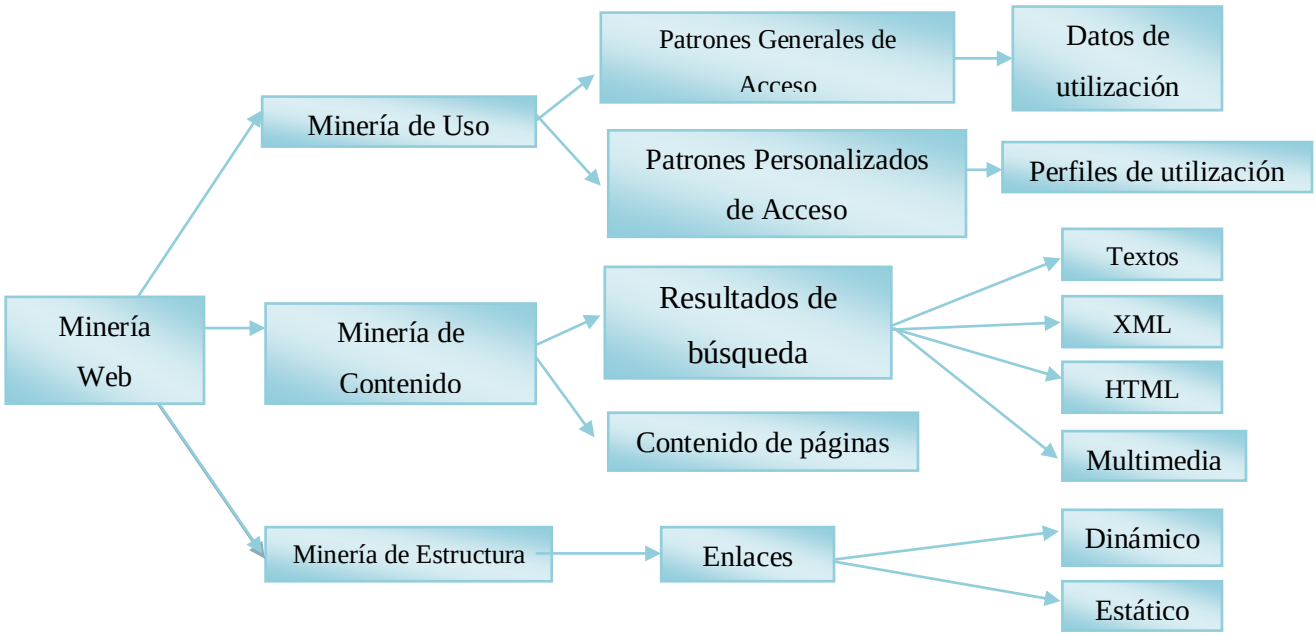
- Interpretación y validación de los patrones minados (análisis): se desarrollan técnicas y herramientas que permitan la interpretación del conocimiento minado.

4.3 CATEGORÍAS DE LA MINERÍA WEB

Con base en los estudios realizados, la minería web se categoriza de acuerdo con el área de la web a ser explorada: La minería del Contenido, minería de la estructura y minería del uso. La minería del contenido se encarga de la extracción de modelos de contenidos de objetos Web (texto plano, documentos semiestructurados (HTML o XML)), documentos estructurados (librerías digitales), documentos dinámicos y documentos multimedia.

Los modelos extraídos son usados para clasificar objetos Web, extraer palabras claves útiles en los procesos de recuperación de información e inferir la estructura de otros objetos. La minería de la estructura se encarga de encontrar la topología de las interconexiones entre objetos Web. El modelo que se construye puede ser utilizado para valorar y categorizar sitios y a la vez encontrar similitud entre estos. La minería de uso se enfoca en el descubrimiento de patrones de uso de datos Web. Los datos son usualmente tomados de la interacción de los usuarios con la web (archivos de log del servidor, logs de servidores Proxy, logs de navegación, datos de registro, sesiones de usuario, cookies, bookmarks o favoritos, entre otros) [HAN0] [KAN03]. Las herramientas de minería de uso ayudan a mejorar la calidad y entrega de servicios de información al usuario final, identificar clientes potenciales en sitios de comercio electrónico y mejorar el desempeño del servidor Web.

Figura 1. Categorías de la Minería Web



Fuente: El autor

4.4 TÉCNICAS EMPLEADAS EN LA MINERÍA DE USO DE WEB

Tres de las técnicas más usadas en la minería web son [ROM05]:

- Agrupamiento y clasificación: Las técnicas de agrupamiento o clustering distribuyen comportamientos de individuos similares en grupos homogéneos. Esto es, elementos con características parecidas pertenecerán al mismo grupo y las características de un grupo serán diferentes a los de otro grupo. Una vez descubiertos los prototipos o perfiles de cada grupo, se pueden usar las características de cada uno de ellos para realizar la clasificación.

- Reglas de Asociación: Las reglas de asociación permiten capturar patrones referentes a los ítems sets sin distinción, en los que ocurre una transacción de datos.

- Secuencias Frecuentes: La minería de secuencias puede ser considerada como asociación minando sobre los datasets temporales y una secuencia de lista ordenada (con el paso del tiempo) de ítem sets no vacíos. El objetivo de esta técnica es descubrir el tiempo de las secuencias ordenadas de URLs que ha sido seguido por usuarios, para predecir a futuros.

Hoy en día debido al gran volumen de información y servicios que ofrece internet a tomado gran importancia el perfilamiento de usuarios por el objetivo de ofrecer sitios web adaptativos en los se le ofrezcan a los usuarios información y servicios de acuerdo a su perfil (preferencias).

Bajo el área de la minería de uso, encontramos los sistemas de detección de perfiles de usuario basados en la información particular de casa usuario construyendo patrones de comportamiento de navegación.

4.5 ARCHIVO LOG

Los archivos de log de acceso generados por los servidores Web son archivos planos donde se registran los recursos solicitados (páginas, imágenes, sonidos, etc), la fecha y hora en la cual fueron accedidos, la dirección Ip que realizó la solicitud y el estado de la solicitud, entre otros datos. Estos archivos son fuente de información acerca de la interacción de los usuarios con un sitio web.

4.6 TRABAJOS REALIZADOS

Basado en el centro: Sarwar, Karypis, Konstan y Ried [SAR00] usan este método para crear un grupo de sesiones de tamaño N para un usuario específico. Las sesiones agrupadas corresponden a aquellas que son más similares a la sesión del usuario actual. La medida del coseno y la medida de correlación se utilizan en el cálculo de la similitud entre sesiones.

Vecindario agregado: Se crea un grupo de sesiones de tamaño N para un usuario U en particular donde la sesión del usuario U es considerada el centro del grupo [SAR00]. Luego, la sesión más cercana al centro es agregada al grupo y el centro es recalculado. Este proceso se repite hasta alcanzar el tamaño del grupo. Este esquema es usualmente usado en conjunto de datos dispersos, puesto que las sesiones más cercanas afectan la formación de los vecindarios.

El Algoritmo de Aglomeración competitiva utilizado por Nasraoui, Frigui, Joshi y Krishnapuram, divide el conjunto de datos en pequeños y numerosos clusters o grupos [NAS99]. Cada unos de los grupos compite por obtener una sesión de usuario. En el transcurso de este proceso es posible que el tamaño de algunos grupos se vaya reduciendo y por ende puedan desaparecer.

Algoritmo k-Means, utilizado por Herr y Chi para la generación de perfiles de usuario [HEE02]. Seleccionan de forma aleatoria el centro de lo K grupos. Cada uno de los elementos o sesiones es agregada al cluster más cercano. Una vez todas las sesiones son asignadas a un grupo se recalcula el centro de los grupos y nuevamente los elementos son reasignados. El algoritmo finaliza cuando los centros de los grupos permanecen estables. Otros autores como Han y Zhao han complementado este algoritmo mediante la búsqueda del K óptimo maximizando la similitud intercluster y disimilitud extracluster y minimizando la disimilitud intercluster y la similitud extracluster, lo que han denominado la teoría Auto-K [HAN05].

Clustering sustractivo difuso relacional: este método de agrupamiento implementado por Bhushan, Nematollaah y Mudar [SHA05], toma cada una de las sesiones para calcularles el potencial mediante la fórmula

$$P_i = \sum_{j=1}^{Nu} e^{-\alpha R_{ij}^2}$$

donde P_i es el potencial para la sesión i, Nu es el número total de sesiones a ser agrupadas, Rij es la similitud entre la sesión i y la sesión j, $\alpha = \frac{4}{\gamma^2}$ con γ igual al valor medio de las similitudes de la sesión i con todas las otras sesiones.

La sesión con el más alto potencial es seleccionada como el primer centro de cluster. El siguiente paso es calcular el potencial de las sesiones restantes con respecto al centro del cluster y nuevamente el potencial más alto es seleccionado como un nuevo centro de cluster. Este proceso se repite hasta alcanzar el número de centros deseados. Una vez obtenidos los N centros de cluster, se calcula la relación entre cada sesión y los clusters con la fórmula

$$u_{ij} = e^{-\alpha R_{c_i j}^2},\ i = [1..C]\ y\ \ j = [1..N_u]$$

para ser asignadas a un cluster en particular, donde $R_{c_i j}$ es la similitud entre el centro del i-esimo cluster con la j-esima sesión.

AntClust: El comportamiento que presentan las hormigas para adherirse a una colonia, ha inspirado el algoritmo AntClust usado en la etapa de agrupamiento de sesiones de usuario

propuesto por Labroche, Monmarch´E, y Venturini. A diferencia de otros algoritmos, AntClust [LAB03] es un algoritmo no supervisado que no requiere la especificación del número del clusters finales pero es necesario definir un valor límite de aceptación entre las hormigas (sesiones) y unas reglas de comportamiento para creación, unión y eliminación de colonias (clusters) así como para la adición o remoción de hormigas a una colonia. AntClust modela las principales características de las hormigas (sesiones) como su olor, su patrón de reconocimiento y su genoma, y asocia un objeto del conjunto de datos al genoma de una hormiga artificial. El olor de cada hormiga es representado con un número que identifica el grupo al que pertenece y va cambiando hasta que encuentre el grupo con el cual mejor se integre. La medida euclidiana es utilizada para calcular la similitud entre características de las sesiones.

5. DISEÑO METODOLOGICO PARA EL PROTOTIPO DE SISTEMA

Para el desarrollo de este trabajo se usara la metodología manejada en investigaciones en torno a la Minería de Datos. Este es un procedimiento general adaptado a problemas que pueden ser resueltos usando las técnicas de minado [KAN03] y que involucra los siguientes pasos: Primero que todo, se busca la documentación relacionada con investigaciones que involucren creación (detección) de perfiles de usuario. A continuación En base a estos documentos se identifican las fases en el proceso de minería y se trabaja en estas de la siguiente forma: Limpieza de datos, se analiza la estructura que tienen los archivos de trabajo, y se define que información de estos archivos se usara en el proyecto.

Luego se busca la forma de transformar los datos y se consulta que técnicas pueden ser aplicadas para la detección de los perfiles de usuario.

Paso seguido se realiza la codificación de cada una de las etapas en las cuales se realizaran las respectivas pruebas con el fin de verificar su funcionalidad.

5.1 PREPROCESAMIENTO DE LOS DATOS

El preprocesamiento de los datos es una tarea muy importante en el proceso de minería de datos, pues contribuye de manera significativa en la creación de patrones de calidad. El objetivo de este preprocesamiento es obtener datos minables de gran calidad. Técnicas de limpieza y transformación de datos son usadas en esta etapa.

Usualmente los archivos de log presentan datos duplicados y datos ruido. La etapa de limpieza se encarga de dar solución a estos inconvenientes, con el objetivo de contribuir al buen desempeño del proceso de transformación de los datos.

Limpieza de los Datos. El objetivo de esta etapa es dejar los datos listos para la etapa de transformación. En algunos casos en este proceso de limpieza incluye relleno de valores faltantes, resolución de inconsistencias y la identificación y borrado de valores anormales que producen ruido en los datos [HAN01]. Según el modelo de minería a implementar se selecciona el tipo de limpieza para los datos.

Los archivos log son fuentes de datos generados automáticamente por los servidores web, proceso en el cual no interviene las manos del hombre, debido a esto, la información tiene una probabilidad baja de presentar inconsistencias y datos erróneos.

En campos como cs-username, cs-uri-query, cookies y cs-referrer, se presenta ausencia de valores, pero esto no afecta el estudio que se va a realizar. En el archivo log se puede encontrar duplicidad en los datos debido al registro de objetos vinculados con cada una de las páginas. La etapa de limpieza está enfocada en el borrado de solicitudes sin éxito, registros de accesos de robots y de información duplicada.

Eliminación de accesos realizados por Robots. En el campo cs(user-Agent) se registra el nombre del explorador web por el cual se accede al sitio web (ej Mozilla/4.0/5.0, Internet Explorer 6.0, Netscape, entre otros), pero en los archivos de log además de registrarsen los accesos de los usuarios y administradores del sitio, se registran los accesos de los robots o crawlers.

Los crawlers son programas que viajan automáticamente a través de internet por los sitios web indexando las páginas con el objetivo de facilitar la tarea realizada por los buscadores.

En el archivo de log es registrado las visitas realizadas por los crawlers y podemos diferenciar estas visitas con las de un usuario normal por medio del campo cs(User-Agent) en donde es registrado el nombre del crawler.

Eliminación de registros basado en el código de estado http. No todas las peticiones realizan a los servidores web son exitosas, inconvenientes en el servidor, autenticaciones de usuario errónea, solicitudes de páginas no existente y demás problemas no permiten que los usuarios reciban una respuesta satisfactoria por parte de los servidores, y por ente no tengan acceso a la página que solicita el usuario.

En el campo sc-status del archivo log es guardado este estado de la solicitud, categorizado de la siguiente manera:

Códigos de información (1xx): Muestran una respuesta provisional a las peticiones realizadas. Es posible que se realicen varias respuestas de este tipo antes de recibir el recurso solicitado.

Códigos de Éxito (2xx): Muestran que el servidor acepto la petición del usuario de forma correcta.

Códigos de Redirección (3xx): Muestran que el explorador del cliente debe realizar acciones adicionales para poder acceder al recurso, por ejemplo, cuando una página ha sido movida y la solicitud debe ser redireccionada por medio de un servidor Proxy.

Códigos de error del cliente (4xx): Muestran errores causados de forma aparente por el usuario, un ejemplo de esto, es cuando se solicita una página que no existe o cuando se ingresa información de autenticación errónea.

Códigos de error del servidor (5xx): Es posible que en algunos casos el servidor no pueda resolver la solicitud del clientes pues se presenta algún error (por ejemplo, el servicio no está disponible, puerta de enlace caducada, versión del protocolo http no compatible, error interno del servidor).

Para el caso de estudio de este trabajo, serán eliminados los registros que no presenten código de estado satisfactorio, es decir se eliminaran del conjunto de datos todas las peticiones cuyo estado haga referencia a errores del cliente y del servidor

Eliminación de duplicidad. Cuando un usuario solicita un página, se registra en el archivo de log además de las url solicitada, las hojas de estilo, los frames y los elementos que componen la página (imágenes, videos y demás).

Para realizar esta limpieza es necesario conocer el tipo de servicio que presta el sitio web para asi establecer los registro y entradas q son importantes. por ejemplo, si el sitio está orientado a ofrecer descargas de imágenes y ringtones para celulares, entonces, deben ser eliminadas las entradas no relacionadas con objetos .gif, .jpeg, .jpg, .bmp, .png, u objetos .wav, .mid, entre otros.
En nuestro estudio sólo se tienen en cuenta los registros referentes a páginas html , htm, asp, jsp, y php.

Transformación de los datos. Uno de los pasos más importantes en el proceso de minado es la transformación de los datos. El principal objetivo de esta fase es consolidar los datos en formas apropiadas para hacer efectivas las técnicas de minería. Para este caso, la etapa de transformación está dirigida hacia la conversión de los datos obtenidos del archivo log, una vez realizada la selección y limpieza, en sesiones de usuario. Una sesión es un conjunto de páginas visitadas por un único usuario en un lapso de tiempo determinado.

Antes de la identificación de los usuarios se realiza la indexación de las páginas del sitio por medio de un vector con 2 campos, índice de la página y nombre de la misma.

Figura 2. Representación Indexación Páginas web

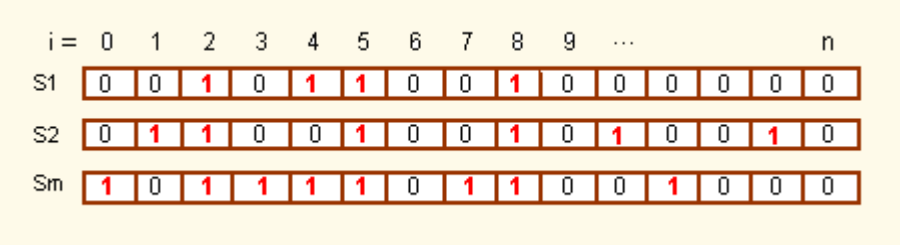
Índice	1	2	3	4	5	...
Pagina	a.Php	b.Php	c.Php	d.Php	d.Php	

Fuente: El autor

El primer paso para la creación de sesiones es la identificación del usuario registro a registro por medio de la dirección IP. Seguidamente, es necesario determinar que páginas corresponden a las visitadas por un usuario en una sesión mediante los atributos fecha y hora de los accesos.

Las sesiones de usuario son representadas como vectores binarios donde cada posición hace referencia a las páginas del sitio. El valor presente en estas posiciones indica si la página fue visitada (1) o no (0) por el usuario más no la frecuencia con que fue visitada dentro de la sesión.

Figura 3. Representación de la sesión de usuario como vector binario



Fuente: El autor

La figura 2 es un ejemplo de cómo se representan las sesiones de usuarios, donde i indica la posición dentro de una sesión S y hace referencia al índice de una página del sitio, n es el tamaño de la sesión o número total de páginas y m el número de sesiones. Se observan 3 sesiones o usuario distintos: el usuario 1 visitó las páginas identificadas con los índices 2, 4, 5 y 8, el usuario 2 las paginas 1, 2, 5, 8, 10 y 13 y el usuario m accedió a las páginas 0, 2, 3, 4, 5, 7, 8 y 11.

5.2 ALGORITMO APRIORI

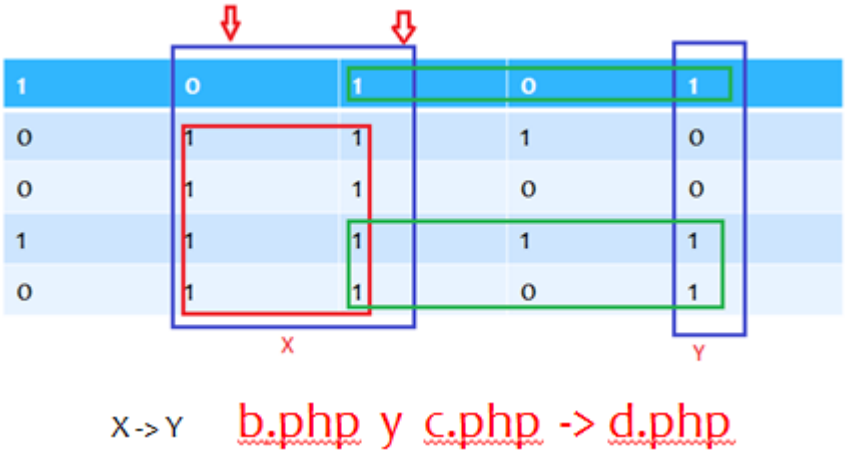
Luego de realizada la indexación de las páginas del sitio y la representación binaria de las sesiones de usuario, se realiza la identificación de las reglas de asociación y el cálculo del soporte y la confianza de estas reglas.

Como primer paso ee identifican las páginas que presentan mayor visita, con las 2 primeras páginas de mayor frecuencia se crea el conjunto X, y la tercera página más visitada es Y para la conformación de la regla de asociación. X-> Y.

Para el cálculo del soporte de X se toman el número de veces que se presenta este conjunto de ítems dividido sobre el número total de sesiones de usuario.
Para realizar el cálculo de la confianza es necesario calcular y soporte de (X u Y), este se realiza tomando el número de veces que se presenta el conjunto de ítems (X -> Y) dividido el número total de sesiones de usuario.

Por último se calcula la Confianza de (X -> Y), dividendo el soporte de (X u Y) sobre el soporte de (X).

Figura 4. Identificación Ítems para Calculo soporte y confianza Regla de asociación

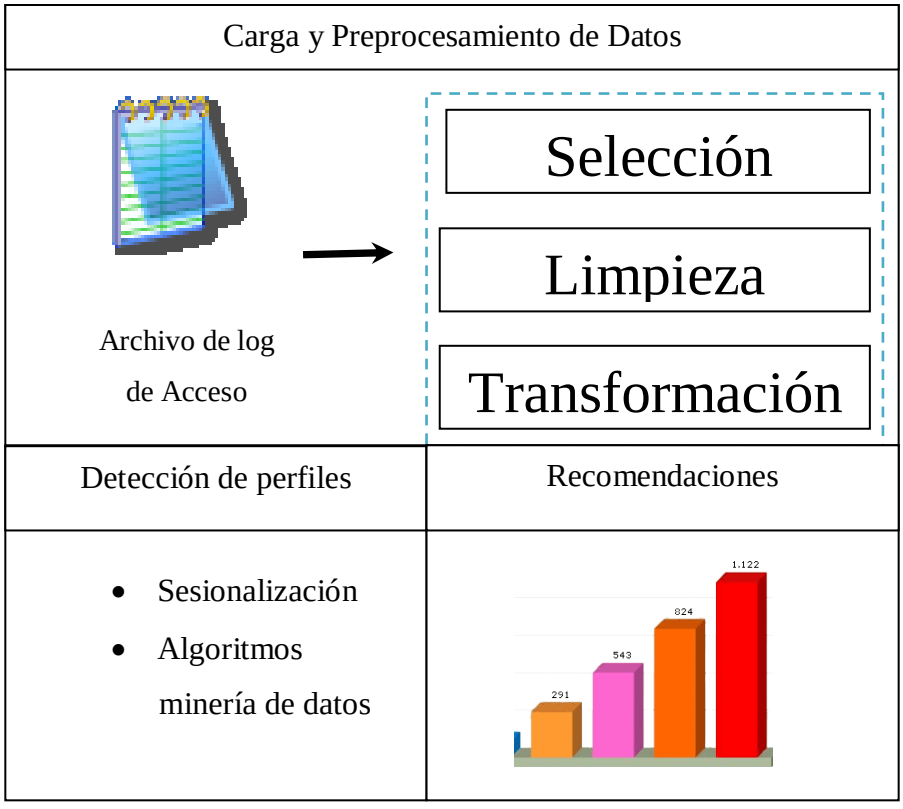


Fuente: El autor

6. ARQUITECTURA Y DISEÑO DEL PROTOTIPO

Los sistemas de identificación de perfiles de usuario al igual que la mayoría de las aplicaciones que se basan en minerías de datos presentan una arquitectura modular correspondiente con cada una de las etapas implementadas. Dicha modularidad es una característica implícita en los procesos de descubrimiento del conocimiento, podemos ilustrar esta arquitectura en el siguiente gráfico.

Figura 5. Arquitectura del Prototipo



Fuente: El autor

6.1 CARGA Y PREPROCESAMIENTO DE LOS DATOS

Usualmente los datos usados en los procesos de minería provienen de archivos planos o de grandes bases de datos, en las cuales se registra de forma permanente las transacciones que se realizan en una organización.

Para el caso de la minería Web, la información se obtiene de documentos, servicios, archivos de log generados por servidores web, datos de registro, cookies y demás. Los datos almacenados en un archivo de log de acceso hacen referencia a las solicitudes de recursos realizadas al servidor como consecuencia de la interacción de los usuarios con un sitio y son usados como insumo para el prototipo realizado.

Estos datos de log exigen un pre procesamiento antes de la aplicación de técnicas de minería enfocadas en la extracción de información, por lo cual es fundamental la realización de los siguientes pasos:

Selección: es necesario identificar los campos que conforman el archivo de log de acceso y establecer cuales contribuyen en la tarea de detección de perfiles de usuario.

Limpieza: este paso está enfocado hacia la eliminación de registros de acceso y columnas que no aportan información relevante en los perfiles de usuario. Registros que hacen referencia a accesos de robots, registros con código de estado no exitoso, duplicidad y accesos a páginas auxiliares son eliminados.

Transformación: en este paso los datos son transformados en objetos apropiados de tal forma que la aplicación de técnicas en etapas posteriores sea fácil y la calidad de los resultados no sea perjudicada. Por tanto, los datos limpios obtenidos en el paso anterior son utilizados para construir los perfiles de usuario e identificar las páginas que son visitadas.

6.2 SESIONALIZACIÓN

Luego de realizar la limpieza de los datos del archivo de log, es necesario identificar las rutas de navegación de cada usuario, este proceso de reconstrucción de los caminos de consulta tomados por los usuarios se denomina sesionalización [VEL08]. El objetivo de esta tarea es encontrar las sesiones reales de cada usuario, esta reconstrucción se realiza desde los archivos de log para asociar cada registro del archivo a una única sesión durante un periodo de tiempo.

6.3 CONSTRUCCIÓN DE PERFILES

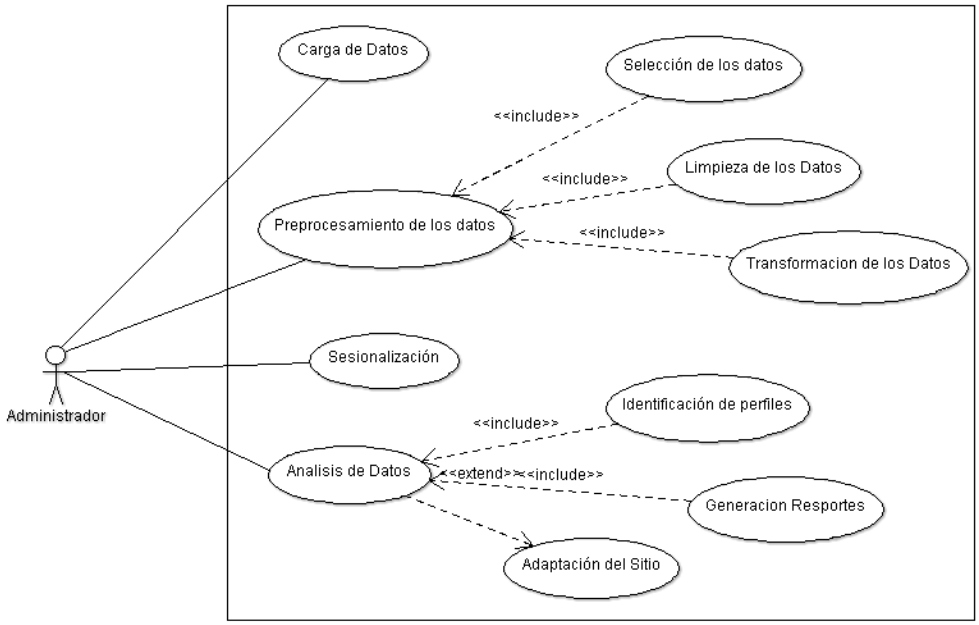
En la minería de datos existen diversos modelos aplicables según los datos disponibles y el conocimiento que desee obtenerse a partir de estos. El objetivo de esta etapa es detectar perfiles de usuario que representen comportamientos de navegación en un sitio Web.
Una vez los datos han sido limpiados y transformados, algunas técnicas son aplicadas para descubrir patrones frecuentes, lo que en términos de minería de uso Web se denominan patrones de navegación o perfiles de usuario. Un patrón o perfil se caracteriza por un conjunto de páginas que reflejan un comportamiento específico y está dado por la similitud existente entre las sesiones de usuario extraídas.

6.4 RECOMENDACIONES

En esta etapa se realizan recomendaciones al administrador del sitio web con el ánimo de mejorar su presentación y usabilidad. Informes como por ejemplo paginas más consultadas, tiempos es que son consultadas mayor número de páginas entre otras.

6.5 DIAGRAMA DE CASOS DE USO

Figura 6. Diagrama de Casos de Uso



Fuente: El autor

En el diagrama de casos para el prototipo desarrollado se identifica que solo interactúa un usuario con la aplicación, el administrador del sistema.

Sobresalen 3 casos de uso particulares: carga de datos, preprocesamiento de datos y análisis de datos. El caso de uso preprocesamiento de datos, incluye los casos de uso selección de datos, limpieza de datos y transformación de los mismo, es decir que la tarea de preprocesamiento debe realizar las tareas de selección, limpieza y transformación de los datos.

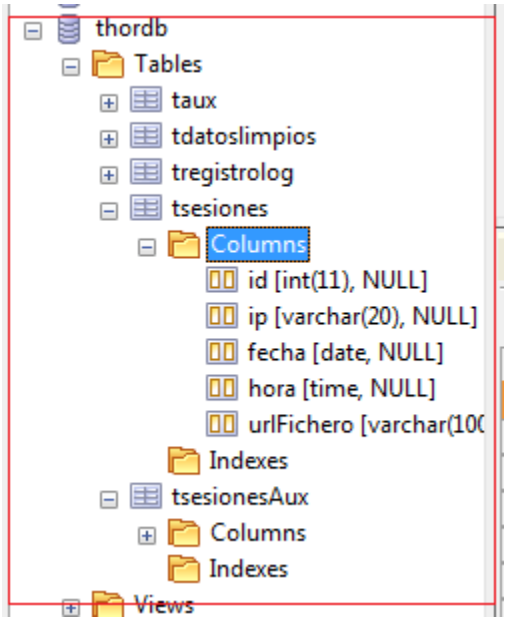
El caso de uso Análisis de datos incluye identificación de perfiles y generación de reportes, esto indica que se realiza la identificación de perfiles y generación de reportes antes de hacer el análisis de datos. El caso de uso adaptación del sitio extiende del caso análisis de datos, es decir cuando se realiza el análisis de los datos es posible que se realice o no la adaptación del sitio.

7. CONSTRUCCIÓN

7.1 BASE DE DATOS

Para la selección, limpieza y tratamiento de los registros se diseñó una base de datos consistente en cinco tablas, en las cuales se almacenan, seleccionan, limpian y transforman los registros del archivo de log.

Figura 7. Tablas Base de Datos

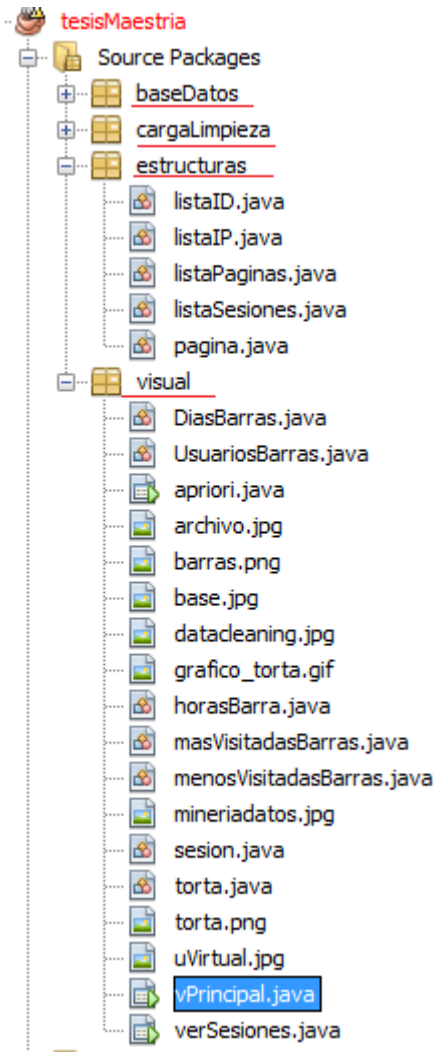


Fuente: El autor

Para el soporte de la base de datos se trabajó con el sistema de gestión de bases de datos MySQL, debido a sus características de multiusuario, multiplataforma y código abierto.

7.2 CODIFICACIÓN EN JAVA

Figura 8. Paquetes y Clases en Java



Fuente: El autor

Para la construcción del prototipo las clases se organizan en 3 Packages los cuales soportan las siguientes clases:

Tabla 1. Paquetes y Clases en Java

Package: baseDatos	
Clase	Métodos Principales
Basedatos.java	Funciones conexión, lectura y consulta en base de datos.
Package: CargaLimpieza	
archivoLog.java	Carga de los datos de log en la base de datos. Limpieza y transformación de los datos.
Package: Estructuras	
listaID.java	Almacenamiento y procesamiento id sesiones
listaIP.java	Clase para almacenar las diferentes direcciones ip de registradas en los archivos de log.
listaPaginas.java	Almacenamiento y procesamiento de las páginas de sesiones de usuario.
listaSesiones.java	Almacenamiento y procesamiento de las sesiones de usuario.
Pagina.java	Clase base para el almacenamiento de los datos de una página.
Package: Visual	
Sesion.java	Extracción direcciones Ip Creación sesiones Restar tiempo Guardar sesiones

	Extraer Sesiones según Tamaño Indexar Paginas Algoritmo apriori
verSesiones.java	Formulario que soporta una tabla para el listado de las sesiones creadas.
vPrincipal.java	Esta clase tiene los atributos y método que soportan la interfaz gráfica que permite la interacción del prototipo con el usuario.
Torta.java	Tiene todos los métodos para generación de los informes en formato torta.
menosVisitadasBarras.java **masVisitadasBarras.java** **horasBarra.java** **UsuariosBarras.java** **DiasBarras.java**	Generación de dataSet Generación y muestra de grafico según informe
A priori	Formulario que soporta el llamado al algoritmo apriori, en él se muestras las reglas de asociación generadas.

Librerías: para el desarrollo del prototipo se utilizaron algunas librerías de uso libre:

- **mysql-connector-java-5.0.4-bin:** API usada para la conexión y ejecución de operaciones sobre la base de datos desde el lenguaje de programación Java. Esta API es independiente del sistema operativo donde se ejecute o de la base de datos a la cual se accesa.

- jcommon-1.0.16: Jcommon es una librería libre de clases de java de propósito general que se utiliza en varios proyectos de la jfree.org, como JFreeChart y JFreeReport.

- jfreechart-1.0.13: Librería grafica libre (LGPL) para Java. Esta librería soporta los gráficos de barras, gráficos circulares, gráficos de líneas, gráficos de series de tiempo, diagramas de dispersión, histogramas, diagramas de Gantt simple entre otros.

- Liquidlnf: Librería usada para el manejo del look and feel de la aplicación.

7.3 HERRAMIENTAS DE DESARROLLO

- **Mysql**: Sistema de gestión de bases de datos relacional, con licencia GPL, tiene soporte multihilo y multiusuario lo que le permite soportar una gran carga de datos de manera muy eficiente.

- **Java:** Lenguaje de programación simple, orientado a objetos, distribuido, robusto, seguro, independiente de la arquitectura, portable, multiusuario entre otras características.

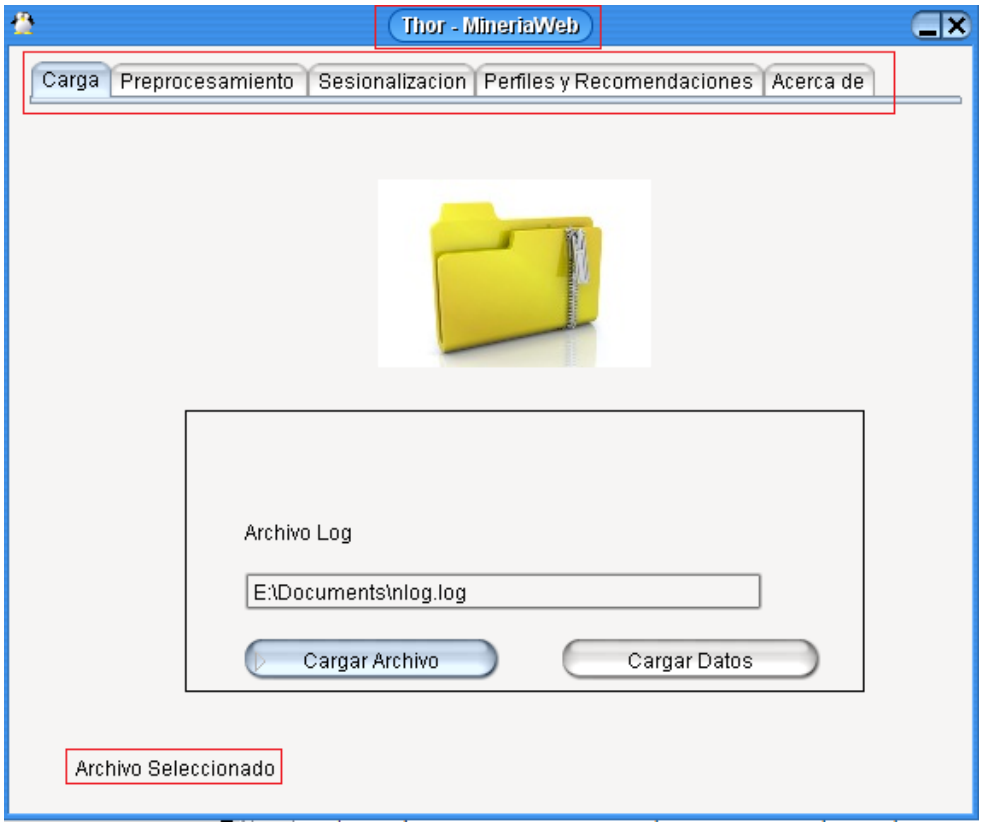
- **Netbeans:** Entorno de desarrollo libre, usado principalmente para programación con el lenguaje java.

- **ArgoUml:** Aplicación de diagramado de Uml escrita en Java y publicada bajo la licencia BSD.

7.4 DESCRIPCIÓN DE PANTALLAS CONSTRUIDAS

Ventana Principal

Figura 9. Ventana Principal de la Aplicación



Fuente: El autor

La ventana principal de la aplicación está compuesta por área de título, menú y barra de estado.

Título: Es igual para toda la aplicación en este caso “Thor – Minería Web”

Menú: Basado en pestañas con 5 opciones: carga, Preprocesamiento, sesionalización, Perfiles y Recomendaciones y Acerca de.

Barra de estado: común para las 3 primeras opciones del menú, en ella se indica el estado en que se encuentra de la aplicación.

Carga.

Figura 10. Ventana Carga de Datos

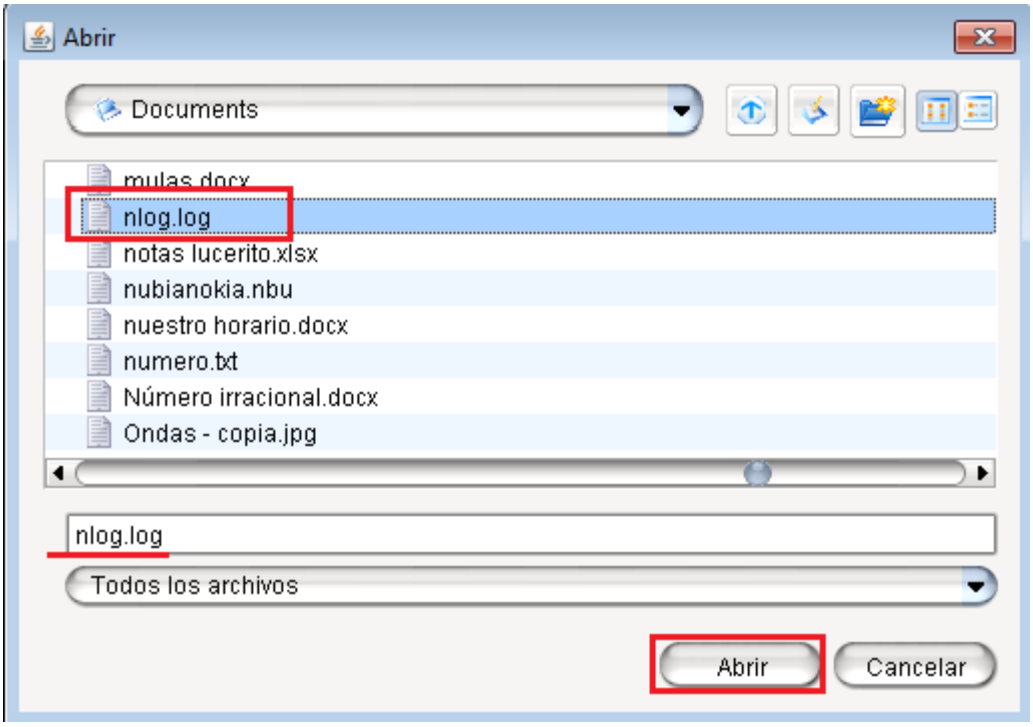
Fuente: El autor

Esta ventana tiene una caja de texto, y 2 botones.

Caja de texto: En esta caja de texto cargamos el nombre del archivo de log.

Cargar Archivo: botón que nos permite seleccionar el archivo de log, al dar clic en él se abre la ventana abrir.

Figura 11. Ventana Abrir Archivo

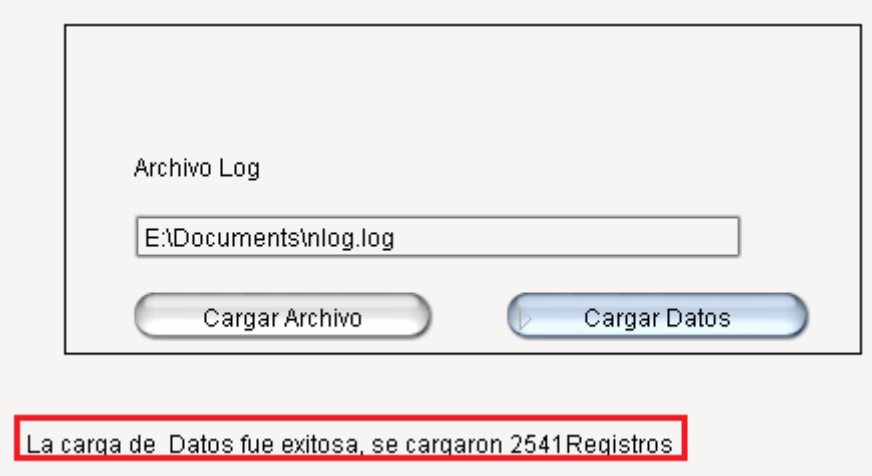


Fuente: El autor

Ventana abrir: En esta pantalla seleccionamos el archivo de log en la ubicación que se encuentre, al dar clic en abrir, se carga el nombre del archivo y se indica en la barra de estado que el archivo fue seleccionado.

Cargar Archivo: al dar clic en este botón se cargan al sistema todos los registros del archivo de log, al terminar esta tarea, se indica en la barra de estado el número total de registros cargados.

Figura 12. Ventana reporte carga de los datos



Fuente: El autor

PreProcesamiento.

Figura 13. Ventana Pre Procesamiento de los datos

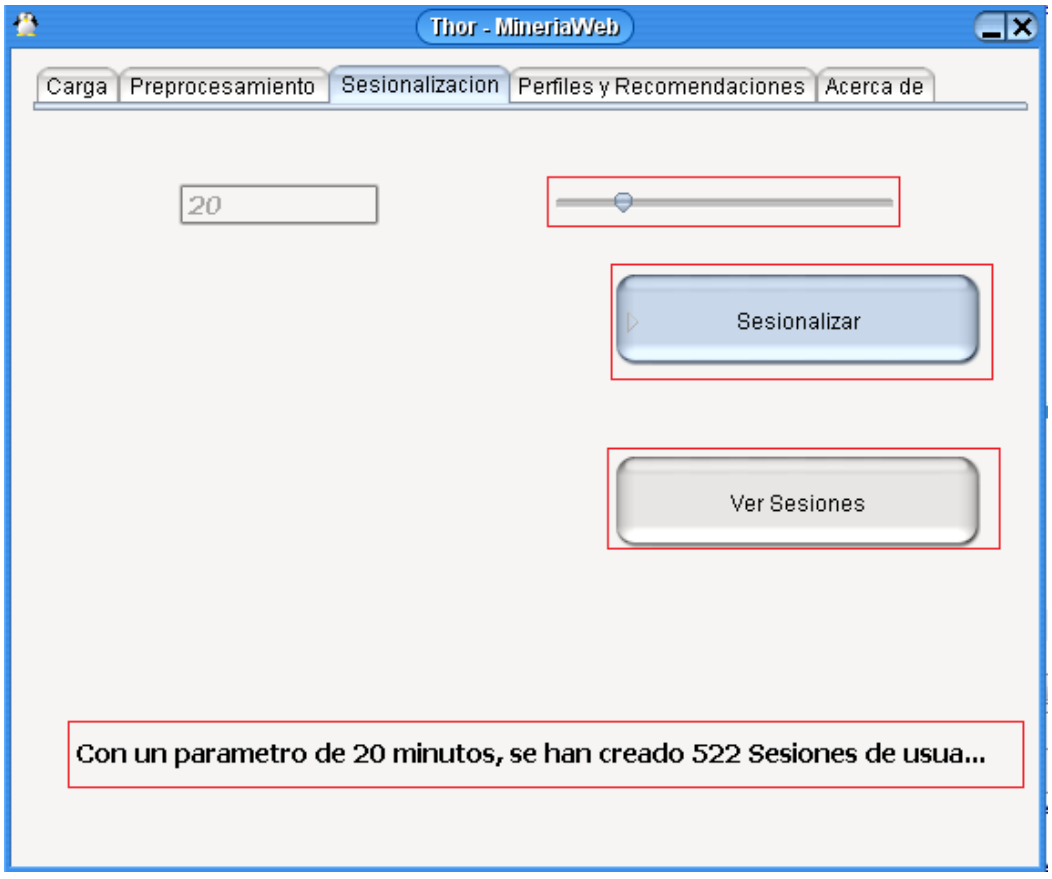
Fuente: El autor

En esta ventana encontramos un único botón y la barra de estado.

Limpieza & transformación: Este botón nos permita hacer una selección y limpieza y transformación de los datos, cuando termina la tarea se indica en la barra de estado cuando registros fueron procesados.

Sesionalización.

Figura 14. Ventana Sesionalización



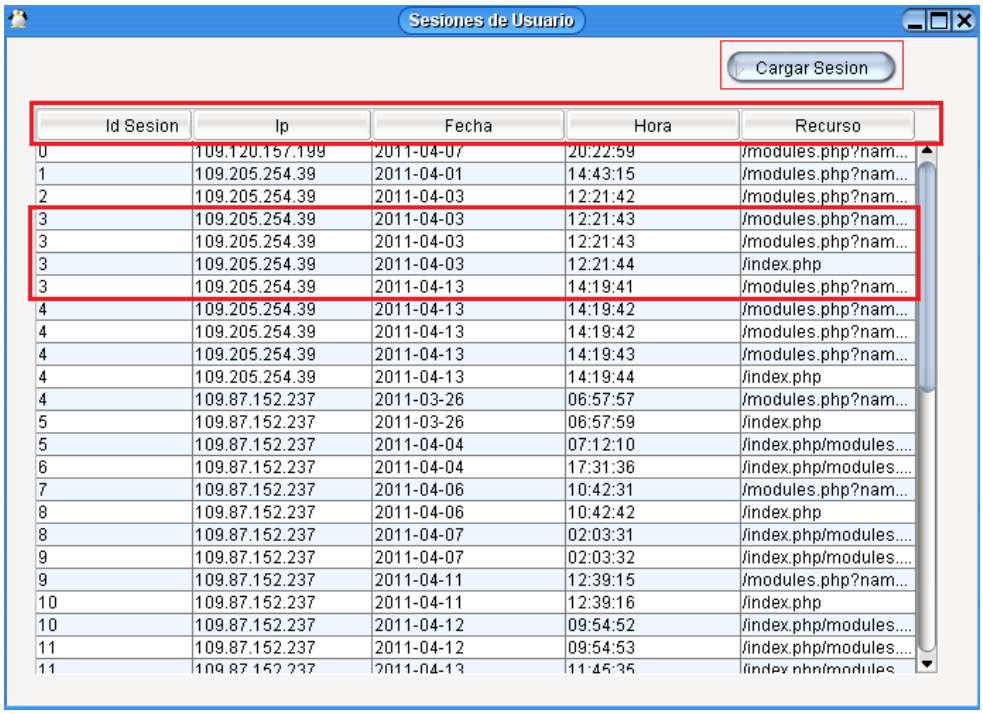
Fuente: El autor

Para realizar la sesionalización esta ventana nos provee de 3 controles, 1 barra de slide que nos permite seleccionar el valor del parámetro en min para crear las sesiones, por defecto este valor es de 20 minutos, pero lo podemos modificar entre 10 y 60 minutos.

El botón sesionalizar con crear y guarda las sesiones de usuario y se nos indica en la barra de estado el número de sesiones creadas.

El botón ver sesiones por carga otra ventana en la cual podemos observar las sesiones anteriormente creadas.

Figura 15. Ventana Sesiones de Usuario



Id Sesion	Ip	Fecha	Hora	Recurso
0	109.120.157.199	2011-04-07	20:22:59	/modules.php?nam...
1	109.205.254.39	2011-04-01	14:43:15	/modules.php?nam...
2	109.205.254.39	2011-04-03	12:21:42	/modules.php?nam...
3	109.205.254.39	2011-04-03	12:21:43	/modules.php?nam...
3	109.205.254.39	2011-04-03	12:21:43	/modules.php?nam...
3	109.205.254.39	2011-04-03	12:21:44	/index.php
3	109.205.254.39	2011-04-13	14:19:41	/modules.php?nam...
4	109.205.254.39	2011-04-13	14:19:42	/modules.php?nam...
4	109.205.254.39	2011-04-13	14:19:42	/modules.php?nam...
4	109.205.254.39	2011-04-13	14:19:43	/modules.php?nam...
4	109.205.254.39	2011-04-13	14:19:44	/index.php
4	109.87.152.237	2011-03-26	06:57:57	/modules.php?nam...
5	109.87.152.237	2011-03-26	06:57:59	/index.php
5	109.87.152.237	2011-04-04	07:12:10	/index.php/modules....
6	109.87.152.237	2011-04-04	17:31:36	/index.php/modules....
7	109.87.152.237	2011-04-06	10:42:31	/modules.php?nam...
8	109.87.152.237	2011-04-06	10:42:42	/index.php
8	109.87.152.237	2011-04-07	02:03:31	/index.php/modules....
9	109.87.152.237	2011-04-07	02:03:32	/index.php/modules....
9	109.87.152.237	2011-04-11	12:39:15	/modules.php?nam...
10	109.87.152.237	2011-04-11	12:39:16	/index.php
10	109.87.152.237	2011-04-12	09:54:52	/index.php/modules....
11	109.87.152.237	2011-04-12	09:54:53	/index.php/modules....
11	109.87.152.237	2011-04-13	11:45:35	/index.php/modules

Fuente: El autor

En esta ventana observamos un botón que nos permite cargar las sesiones de usuario, una tabla que nos muestra un encabezado con los títulos del ítem mostrado y posterior a esto las sesiones de usuario.

Perfiles y recomendación.

Figura 16. Ventana Perfiles y Recomendaciones

Fuente: El autor

Esta ventana nos provee de 4 secciones: Informe, cantidad, presentación y apriori.

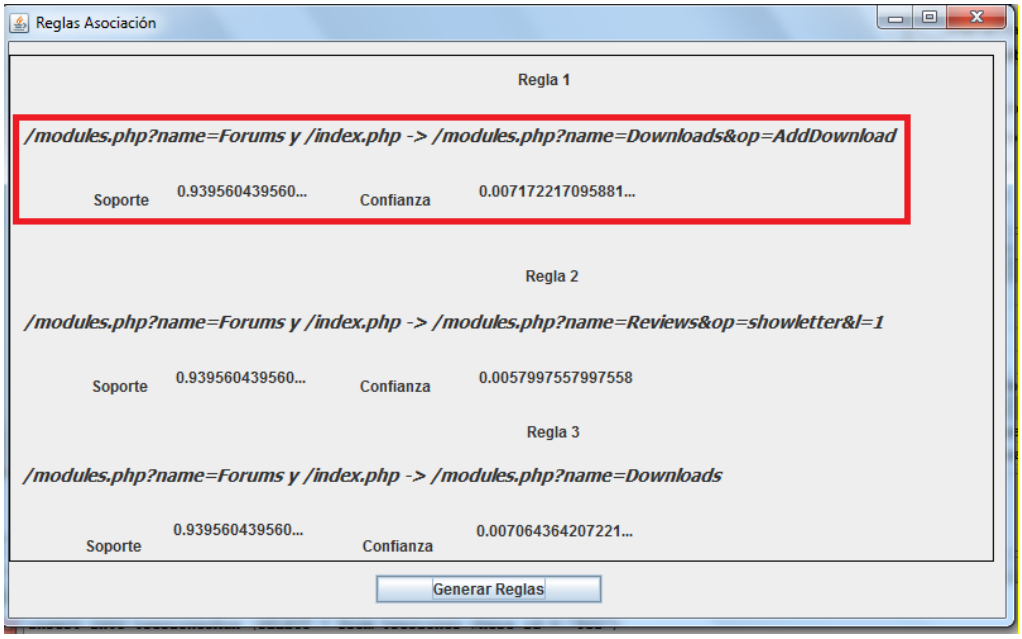
Informe: En esta sección podemos escoger el informe que queramos generar: Páginas más visitadas, Usuarios más Activos, Paginas menos visitadas, Visitas por hora del día y Visitas por el día de la semana.

Cantidad: Para los informes de Páginas más visitadas, usuarios más activos y paginas menos visitadas este campo se activa para que el usuario pueda escoger cuantos registros desea ver en el informe, por defecto toma el valor de 10, para los informes visitas por hora del día y visitas por días de la semana este campo se inhabilita.

Presentación: En ese apartado se puede seleccionar entre 2 tipos de gráficas: barras y torta. Estos tipos de presentación sirven para cualquiera de los informes.

Apriori: Este botón nos hace el llamado a otro formulario que nos muestra las reglas de asociación generadas por el algoritmo apriori.

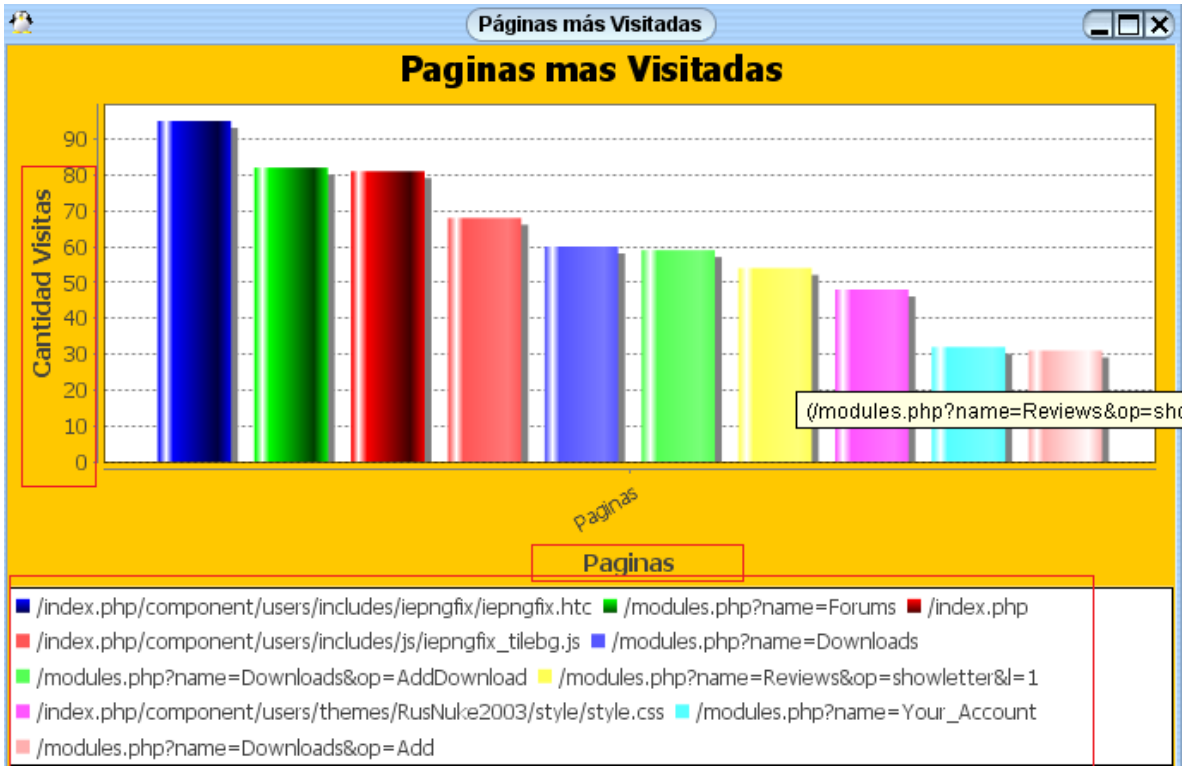
Figura 17. Ventana Reglas de Asociación



Fuente: El autor

Informe por barras.

Figura 18. Ejemplo Informe Presentación por Barras



Fuente: El autor

En este tipo de informes se presentan los ítems de una categoría en el eje horizontal y sobre el eje vertical los valores que toman cada ítem, en la parte inferior de grafico se muestra los ítems que representa cada barra. En el gráfico de ejemplo se puede observar las páginas más visitadas, cada barra vertical representa una página y la altura de cada una de ellas la cantidad de visitas obtenidas.

Informe en torta o circular.

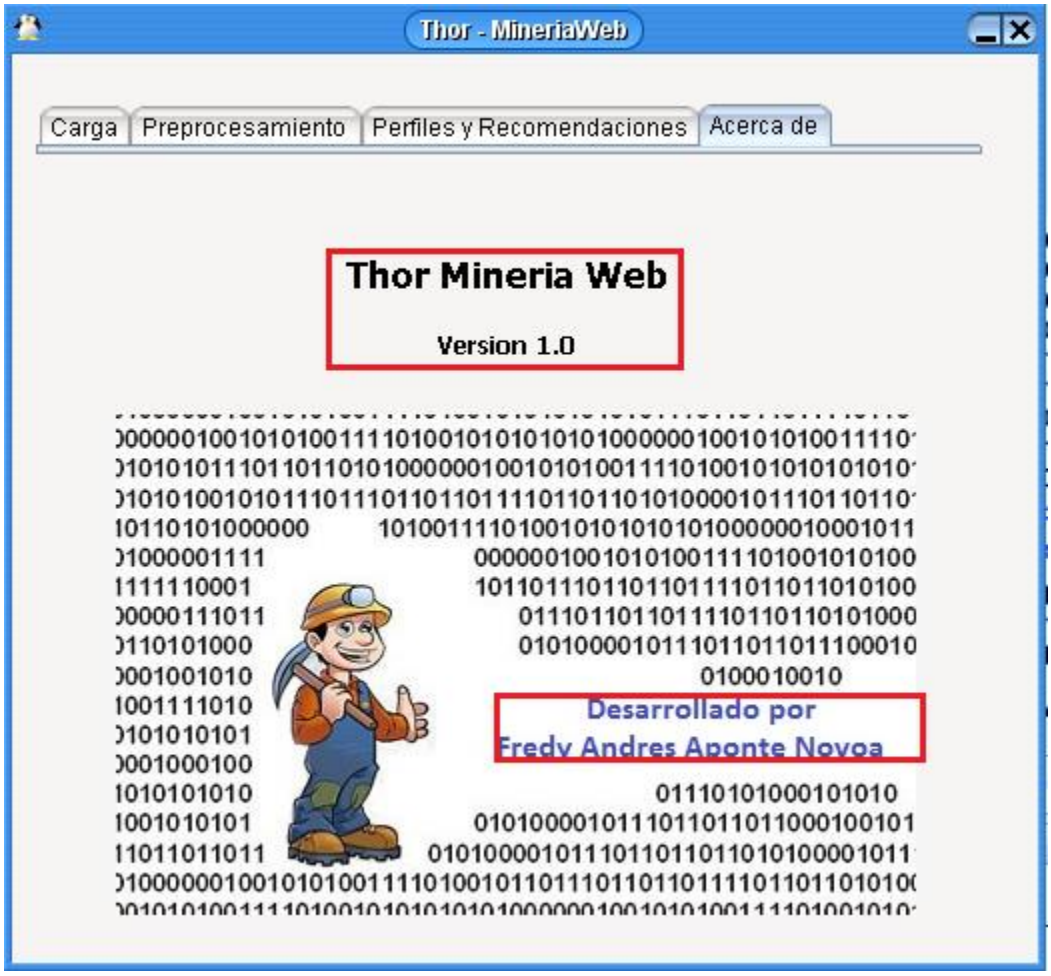
Figura 19. Ejemplo Informe Presentación Circular

Fuente: El autor

En este tipo de informes se presentan los ítems de una categoría en pedazos de una circunferencia (torta), cada ítem se representa de un color en particular. En la parte inferior del grafico se muestra los ítems que representa cada pedazo de la torta. En el gráfico de ejemplo se puede observar las peticiones por día de la semana, cada pedazo de la torta representa un día de la semana página y su tamaño la cantidad de consultas en ese día.

Acerca de.

Figura 20. Ventana Acerca de



Fuente: El autor

En esta pantalla se muestra la información de la aplicación y del desarrollador de la misma.

8. ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

Los resultados de la investigación obtenidos por el prototipo desarrollado se basan en los datos alojados en el archivo de log de acceso del sitio HTTP://ALLTSK.TOMSK.RU, un portal ruso de una comunidad desarrolladores de software libre, la cual promueve la filosofía del software libre además de ofrecer servicio de hosting, publicidad en internet y servicio de pagos electrónicos. Actualmente el sitio está desarrollado en el sistema de gestión de contenidos PHP-NUKE, el log utilizado fue extraído con registros referentes al proyecto PHP-NUKE.

Figura 21. Ventana Principal Sitio ALLTSK.TOMSK.RU

Fuente: El autor

El prototipo permite general informes estadísticos lo que permite una mejor interpretación de los datos alojados en el servidor web, facilitando al administrador del sitio tomar decisiones para adoptar estrategias que contribuyan a mejorar la interactividad del sitio con sus usuarios.

Los registros del archivo log representan los patrones de comportamiento del usuario en el sitio web, cada registro del archivo de log tiene entre otra información la siguiente: dirección ip, fecha y hora de acceso, requerimiento, Url de recurso, el protocolo utilizado, código de error, agente que realizó el requerimiento, y el número de bytes transmitidos. Esta información es almacenada en un archivo plano, donde cada renglón representa un registro y cada campo es separado por un espacio (“ ”).

Figura 22. Muestra Archivo log de Acceso

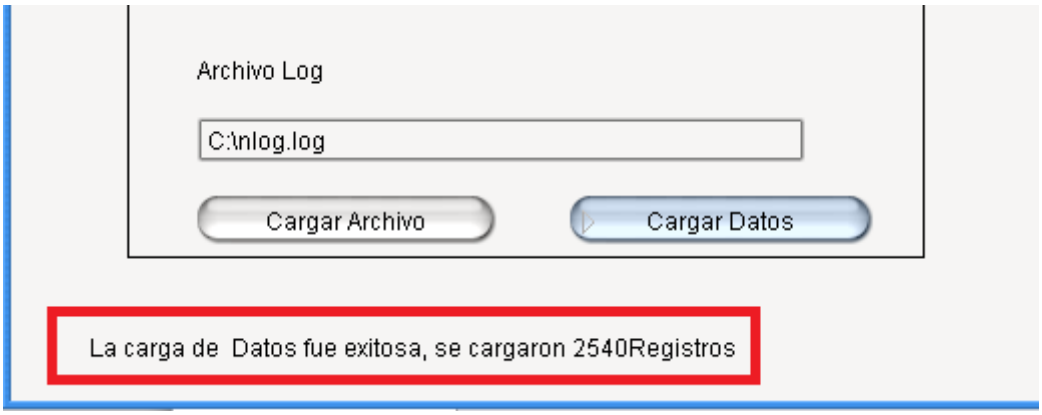
```
95.108.248.29 - - [24/Mar/2011:15:03:25 +0600] "GET /modules.php?
name=Downloads&op=NewDownloadsDate&selectdate=1300045121 HTTP/1.1" 200 21552
95.108.248.29 - - [24/Mar/2011:15:29:25 +0600] "GET /modules.php?
name=Downloads&op=NewDownloadsDate&selectdate=1298749122 HTTP/1.1" 200 21556
66.249.66.198 - - [24/Mar/2011:15:48:09 +0600] "GET /kcaptcha/index.php?
PHPSESSID=6859bfsvld18r21molm5bjdec5&rnd=7870&knum=4037 HTTP/1.1" 200 3649
95.108.248.29 - - [24/Mar/2011:15:55:18 +0600] "GET /modules.php?
name=Downloads&op=NewDownloadsDate&selectdate=1298317122 HTTP/1.1" 200 21555
95.108.248.29 - - [24/Mar/2011:16:21:11 +0600] "GET /modules.php?
name=Downloads&op=NewDownloadsDate&selectdate=1300822721 HTTP/1.1" 200 21553
95.108.248.29 - - [24/Mar/2011:16:47:01 +0600] "GET /modules.php?
name=Downloads&op=NewDownloadsDate&selectdate=1300477121 HTTP/1.1" 200 21553
66.249.66.198 - - [24/Mar/2011:17:06:23 +0600] "GET /index.php/modules.php?
name=News&file=article&sid=1 HTTP/1.1" 200 22049
79.142.79.11 - - [24/Mar/2011:17:06:42 +0600] "GET /logs/access.log HTTP/1.1" 200
1095
95.108.248.29 - - [24/Mar/2011:17:12:50 +0600] "GET /modules.php?
name=Downloads&op=NewDownloadsDate&selectdate=1299958721 HTTP/1.1" 200 21553
95.108.248.29 - - [24/Mar/2011:17:38:39 +0600] "GET /modules.php?
name=Downloads&op=NewDownloadsDate&selectdate=1298489922 HTTP/1.1" 200 21556
```

Fuente: El autor

8.1 CARGA DE DATOS

Se realizó la carga de los 2540 registros almacenados en el archivo de log

Figura 23. Ventana Carga de Datos



Fuente: El autor

8.2 LIMPIEZA DE DATOS

Al realizar la limpieza de los datos se obtuvieron 2410 registros limpios y transformados. Esto indica que un 94,88% de los datos son minables y 5.11% de ellos son datos basura para el sistema.

Figura 24. Ventana Resultados Preprocesamiento de datos

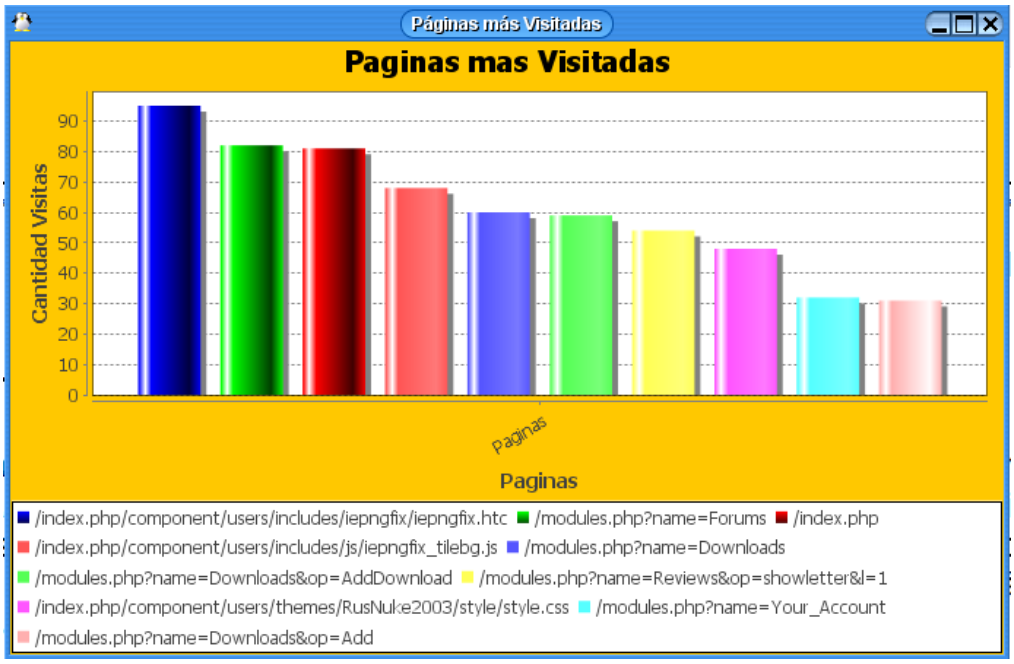
Fuente: El autor

8.3 INFORMES

Cada uno de los informes presentados por la aplicación son actualizados cada vez que se cargue el archivo log y se realice el proceso pre procesamiento de los datos, el resultado de cada informe debe ser usado realizar un mejor diseño y administración del sitio web.

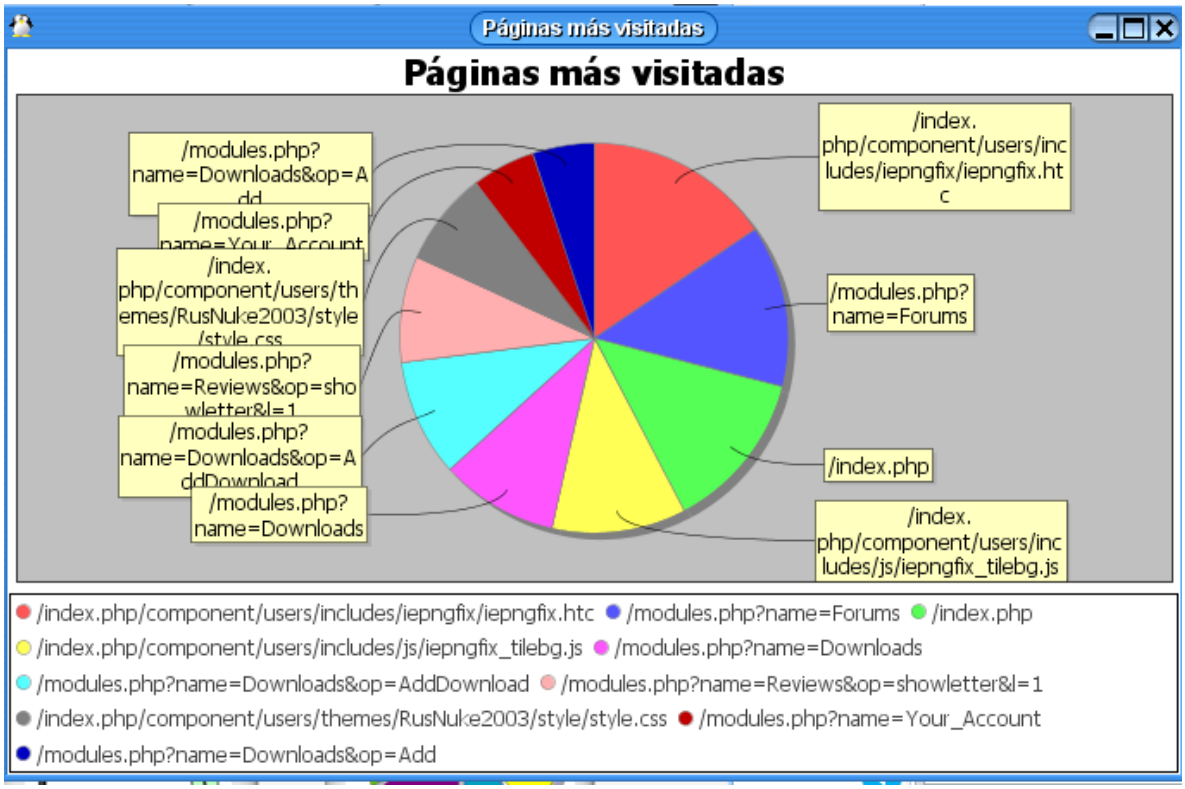
Páginas más visitadas. Se generó el informe de páginas más visitadas con un parámetro de 10 páginas.

Figura 25. Informe Páginas más visitadas barras



Fuente: El autor

Figura 26. Informe Páginas más visitadas Circular



Fuente: El autor

Figura 27. 10 Páginas más Visitadas

/index.php/component/users/includes/iepngfix/iepngfix.htc	95
/modules.php?name=Forums	82
/index.php	81
/index.php/component/users/includes/js/iepngfix_tilebg.js	68
/modules.php?name=Downloads	60
/modules.php?name=Downloads&op=AddDownload	59
/modules.php?name=Reviews&op=showletter&l=1	54
/index.php/component/users/themes/RusNuke2003/style/style.css	48
/modules.php?name=Your_Account	32
/modules.php?name=Downloads&op=Add	31

Fuente: El autor

Análisis del reporte

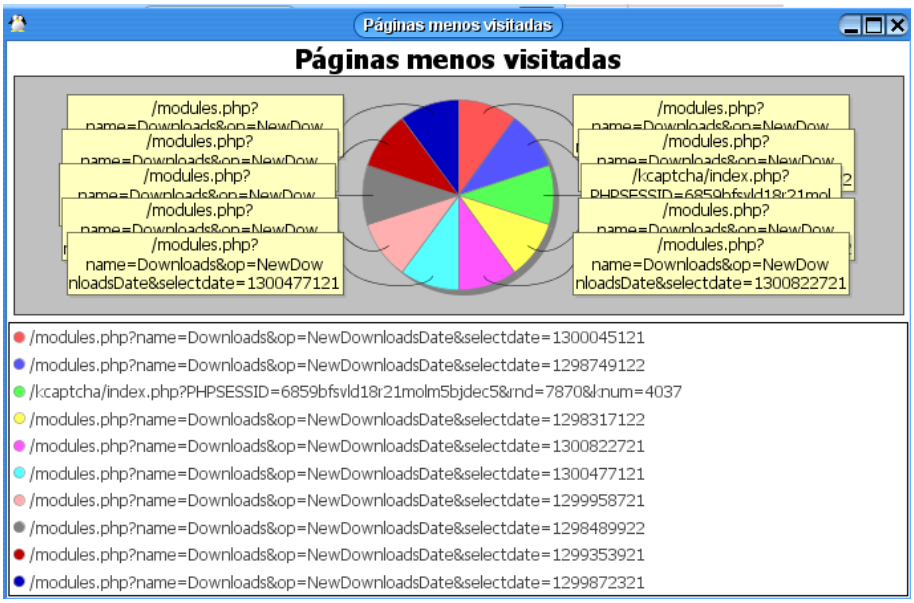
El informe nos muestra que la pagina más visitadas corresponden a la relacionada con la cuenta del usuario con 95 visitas, seguida de la página principal de foros con 82 visitas, seguida de la página principal del sitio con 81 visitas, también entre estas 10 páginas más visitadas se destacan 3 paginas relacionadas con descargas que entre ellas suman 122 visitas. Lo que nos indica que las preferencias de los usuarios sobre el sitio web son las descargas así como los foros que esta ofrece.

Con base este informe también podemos concluir que los usuarios visitan directamente sus páginas de mayor interés (descargas y foros) sin pasar por la página principal, lo cual surgiere hacer la página principal más atractiva para que además de las sus preferencias los usuarios visiten la demás páginas del sitio.

Páginas menos visitadas. Se generó el informe de páginas menos visitadas con un parámetro de 10 páginas.

Figura 28. Informe Páginas menos Visitadas barras

Figura 29. Informe Páginas menos Visitadas Circular



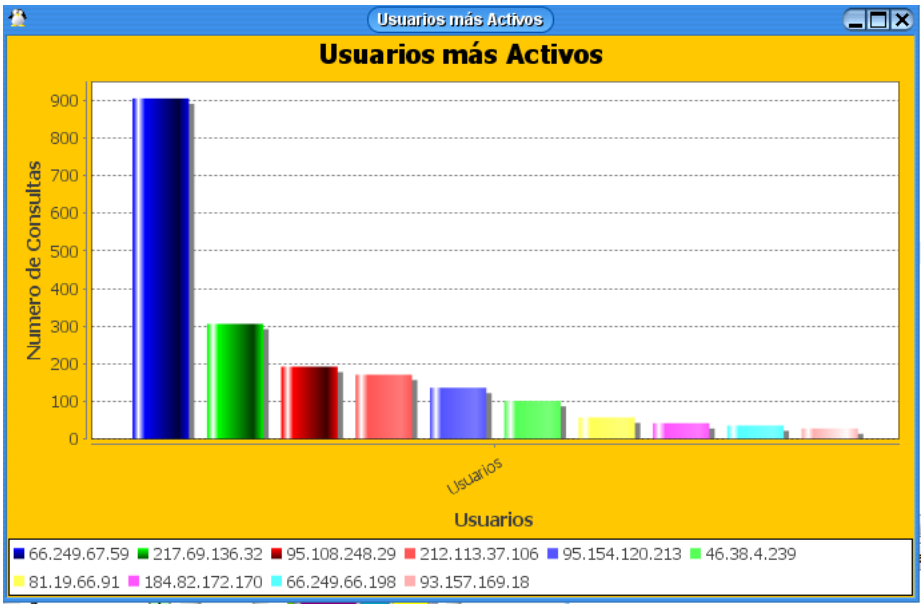
Fuente: El autor

Este informe es uno de los de mayor importancia para el administrador del sitio web ALLTSK.TOMSK.RU, acá se indican las páginas que registran menos visitas en el sitio.

9 de las 10 páginas menos visitadas corresponden a los nuevos ítems a ser descargados en el sitio, en base a esto se recomienda al administrador del sitio web realizar una mejor ubicación de estas páginas en los encales principales así como mejorar la presentación de estas para que sean más atractivas para el usuario del sitio web.

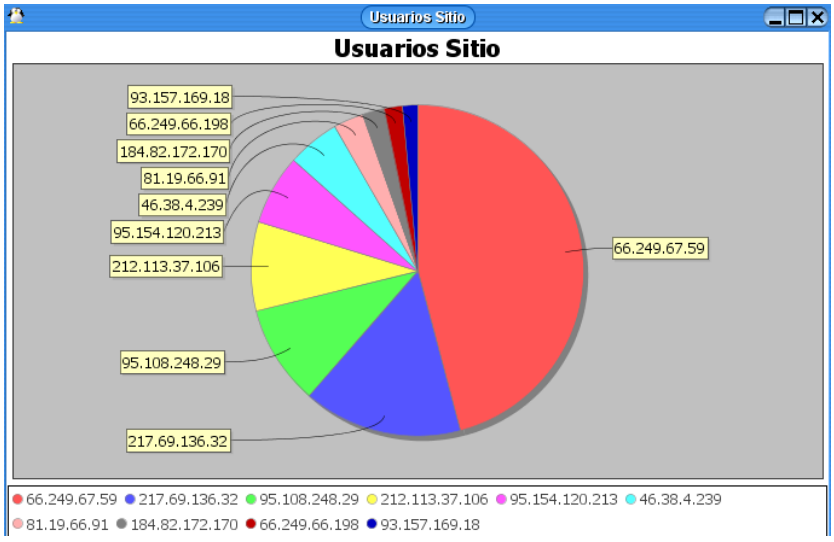
Usuarios más Activos. Se generó el informe de usuarios más activos con un parámetro de 10 páginas.

Figura 30. Informe Usuarios más activos Barras



Fuente: El autor

Figura 31. Informe Usuarios más Activos Circular



Fuente: El autor

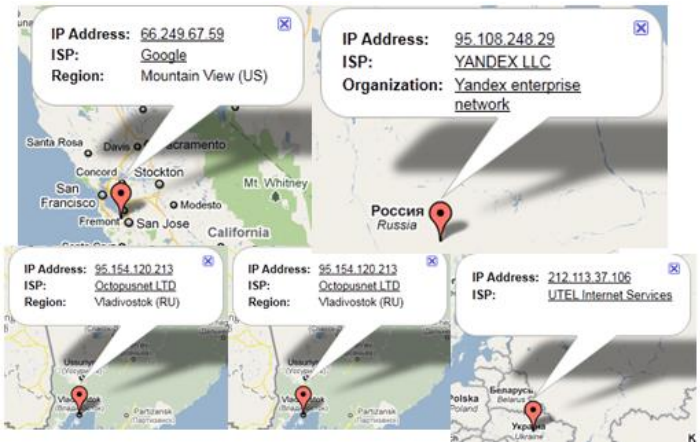
El usuario que con mayor frecuencia visita el sitio presenta la dirección IP 66.249.67.59 que representa el 37,55 de las visitas del sitio, se identifica que esta dirección corresponde a la california en estados unidos, en la siguiente tabla se relacionan las 5 direcciones y IP más frecuentes, el porcentaje de frecuencia y la ubicación de cada una de ellas.

Tabla 2. Frecuencia y Ubicación Usuarios más Activos

IP	Porcentaje Frecuencia	Ubicación
66.249.67.59	37,55	Mountain (US)
217.69.136.21	12,69	Moscú, Rusia
95.108.248.29	7,96	Moscú, Rusia
212.113.37.106	7,09	Ucrania
95.154.120.213	5,64	Vladivostok, Rusia

En base a este informe podemos concluir que de las 1710 consultas visitas de los 5 usuarios más activos 805 que correspondes al 47,07 % corresponden a usuarios de Rusia y Ucrania países que habla ruso, este comportamiento se refleja debido al idioma del sitio (ruso), por lo cual se sugiere al administrador del sitio adecuar la internacionalización del sitio para que pueda ser presentado en varios Idiomas y así captar más visitantes.

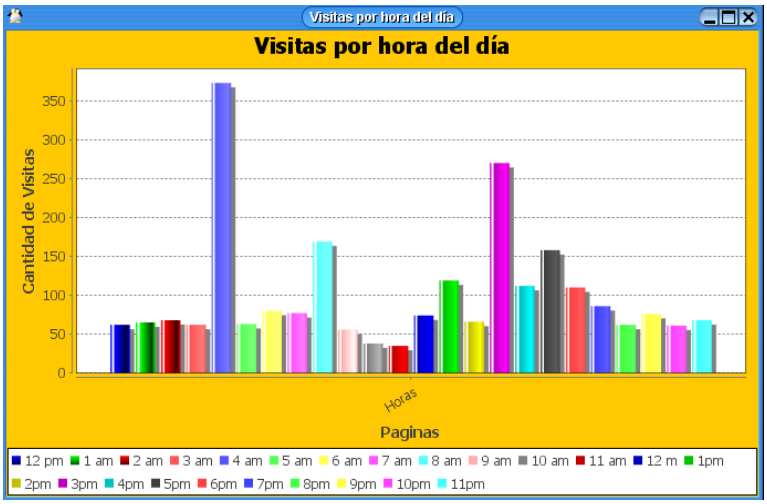
Figura 32. Ubicación geográfica usuarios más activos



Fuente: El autor

Visitas por hora del día. Este informe es muy importante para poder identificar las horas que presenta mayor tráfico el sitio, así como las horas que en el uso es bajo, esto con el objetivo de buscar estrategias de optimización de los recursos del sitio.

Figura 33. Informe Visitas por hora del día barras



Fuente: El autor

Figura 34. Informe Visitas por hora del día Circular

Fuente: El autor

En la siguiente tabla se muestran las 3 horas más frecuentes y el porcentaje de visitas para cada hora en el sitio web

Tabla 3. Horas más Frecuentes de visitas

Hora	Porcentaje de visitas
4 am	15,47
2 pm	11,20
8 am	7,01

En la siguiente tabla se muestran las 3 horas menos frecuentes y el porcentaje de visitas para cada hora en el sitio web

Tabla 4. Horas menos frecuentes de visitas

Hora	Porcentaje de visitas
11 am	1,45
10 am	1,57
9 am	2,32

Con base en la información reflejada en las tablas anteriores se puede concluir que la franja horaria en la cual el servidor presenta más carga y en el cual podría llegar a colapsar es entre las 4:00 am y las 4:59 am y la franja en que presenta el menor número de visitas en entre las 9:00 am y las 9:59 am.

Visitas por día de la semana. La importancia de este informe radica al igual que el anterior en la identificación de días de mayor tráfico y de menor tráfico. Con esta información y la obtenida en el informe anterior se programan estrategias para la realización de mantenimiento del sitio web en horas y día de muy bajo tráfico.

Figura 35. Informe visitas por día de la semana barras

Fuente: El autor

Figura 36. Informe visitas por día de la semana circular

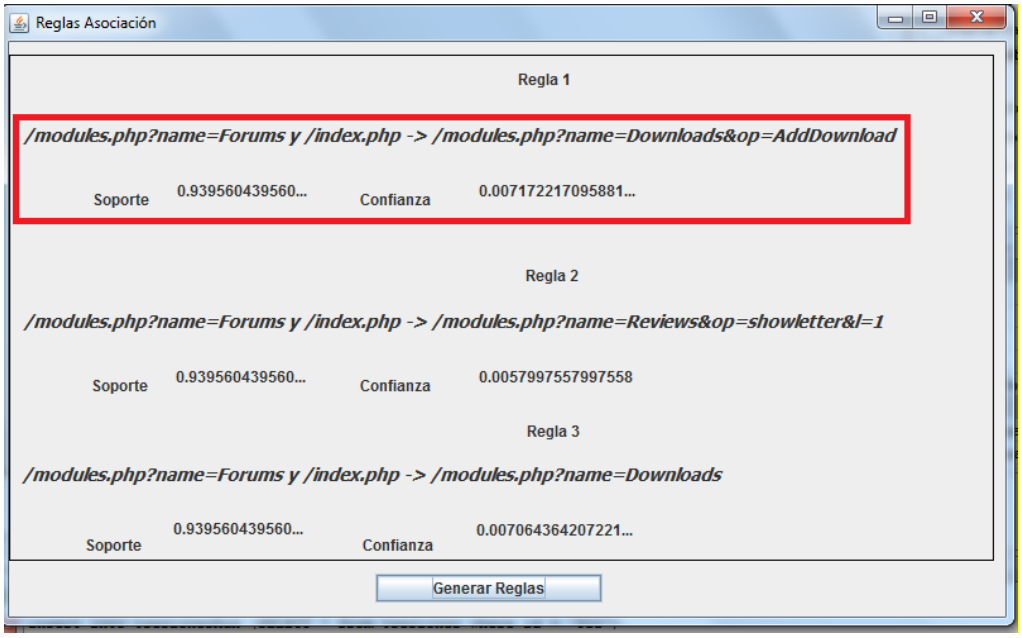
Fuente: El autor

Este informe nos indica que el día de mayor tráfico es el miércoles con un 27,05 % del total de las visitas, así mismo nos muestra que el domingo es el día de menor tráfico con tan solo un 5,85 % del total de las visitas.

Con la información obtenida en los informes de hora y día de la semana podemos recomendar al administrador del sitio realizar el mantenimiento del mismo los domingos entre las 11:00 am y las 11:59 am.

Reglas de asociación.

Figura 37. Informe Reglas de asociación



Fuente: El autor

Este informe nos muestra 3 reglas de asociación generadas por el algoritmo apriori, la primera de ellas nos indica que los usuarios que visitaron las paginas "/modules.php?name=Forums" y "/index.php" posiblemente visiten la pagina

“/modules.php?name=Downloads&op=AddDownload” esto con un soporte de 0,93 y una confianza de 0,0071.

Lo anterior nos indica que un 93% de los registros cubren correctamente la regla y un 0,71 % de los registros hacer verdadero el antecedente de la regla, es decir que las paginas “/modules.php?name=Forums” y “/index.php” se presentan un 93% en el conjunto de sesiones y en un 0,71% los usuarios que visitaron “/modules.php?name=Forums” y “/index.php” también visitaron “/modules.php?name=Downloads&op=AddDownload”.

9. CONCLUSIONES

- El análisis de la información registrada en los archivos de log de acceso por medio de la herramienta desarrollada si facilita la toma de decisiones y la conformación de alertas e informes oportunos al administrador de un sitio web.

- La técnica de clasificación en minería web, permiten la creación de un perfil para visitantes para ser utilizado en la clasificación de nuevos usuarios del sitio.

- El uso de herramientas de software libre contribuye de manera significativa en el desarrollo de proyectos de investigación en la comunidad académica.

- La categorización del comportamiento de los usuarios realizada por la herramienta, permite a los administradores del sitio web la toma de decisiones para mejorar la usabilidad del sitio.

- Las reglas de asociación en la minería de datos nos ayudan a la identificación de patrones de comportamiento de usuario con una técnica estadística aceptada para hacer una proyección a futuro hacia la inteligencia artificial.

10. BIBLIOGRAFÍA

[HAN01] HAN Jiawei, KAMBER Micheline. (2001). Data Mining Concepts and Techniques. Morgan and Kaufman.

[HAN05] HAN Xiwu, ZHAO Tiejun. (2005). Auto-K Dynamic Clustering Algorithm.

[KAN03] KANTARDZIC Mehmed. (2003). Data Mining Concepts, Models, methods and algorithms.

[MAR] MARTÍN Bautista, MARÍA J, Vila María Amparo. Building adaptive user profiles by a genetic fuzzy classifier with feature selection.

[MIT03] MITRA Sushmita, ACHARYA Tinku. (2003). Data Mining Multimedia, Soft Computing And Bioinformatics. New Jersey: John Wiley & Sons, Inc.

[NAS99] NASRAOUI Olfa, FRIGUI Hichem, JOSHI Anupam, KRISHNAPURAM Raghu. (1999). Mining Web Access Logs Using Relational Competitive Fuzzy Clustering.

[LAB03] LABROCHE Nicolas, MONMARCH´E Nicolas, VENTURINI Gilles. (2003). Web sessions Clustering with artificial ants colonies.

[ROM05] ROMÁN Ulises, ALARCÓN Luis. (2005). Minería De Uso De Web Para Predicción De Usuarios En La Universidad.

[SAR00] SARWAR Badrul, KARYPIS George, KONSTAN Joseph, RIED John. (2000). Analysis of Recommendation Algorithms for E-Commerce.

[SHA05] SHANKAR Bhushan , SHIRI Nematollaah, Sudhir P. (2005). Mudur Incremental Relational Fuzzy Subtractive Clustering for Dynamic Web Usage Profiling.

[STU02] STUMME Gerd, HOTHO Andreas, BERENDT Bettina. (2002). Usage Mining for and on the Semantic Web.

[VEL08] VELASQUEZ J. D, PALADE V. (2008). Adaptive Web site: A Knowledge Extraction from Web Data Approach. IOS Press.