Cenidet

Sistema Generador de Predicciones de Acceso para la Replicación de Sitios de la Web en Dispositivos Inalámbricos

Ing. Gabriel Hernández Méndez
M.C.Juan Gabriel González Serna
Ing. Juan Carlos olivares

Agenda

- Introducción.
- Arquitectura.
- Identificación de usuarios.
- Identificación de sesiones de usuarios.
- Mecanismos para la identificación de usuarios y sesiones.
- Búsqueda de patrones interesantes.
- Ítems frecuentes.
- Reglas de asociación.
- Minería de reglas de asociación
- Conclusiones.
- Bibliografía.

Introducción

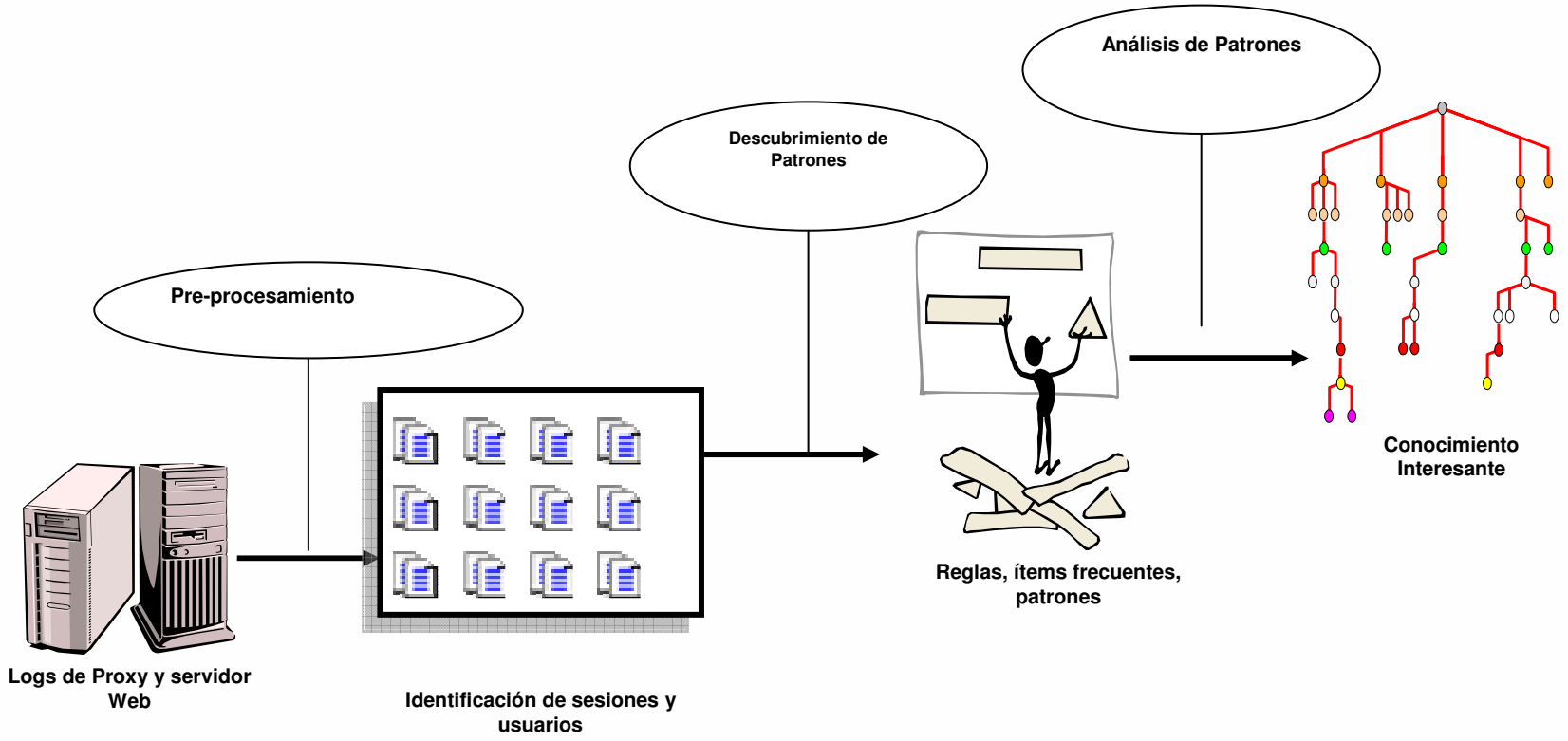
- El Web Mining se refiere a la aplicación de técnicas de Data Mining sobre la World Wide Web.
- De esta definición se deriva que WM es simplemente aprovechar las técnicas de DM para obtener conocimiento de la información disponible en Internet.

Introducción

- Existen ejemplos claros en lo resulta útil el análisis de los datos de uso Web.
  - Mejorar el diseño de la estructura de un sitio Web.
  - Planeación de campañas de marketing orientadas al comercio electrónico.
  - Mejoramiento de sistemas, ya sea en la calidad de su desempeño.
  - En el caso particular de este estudio, se utilizó el análisis para identificar patrones de acceso a recursos Web con el objetivo de seleccionar archivos para el acaparamiento en dispositivos inalámbricos.

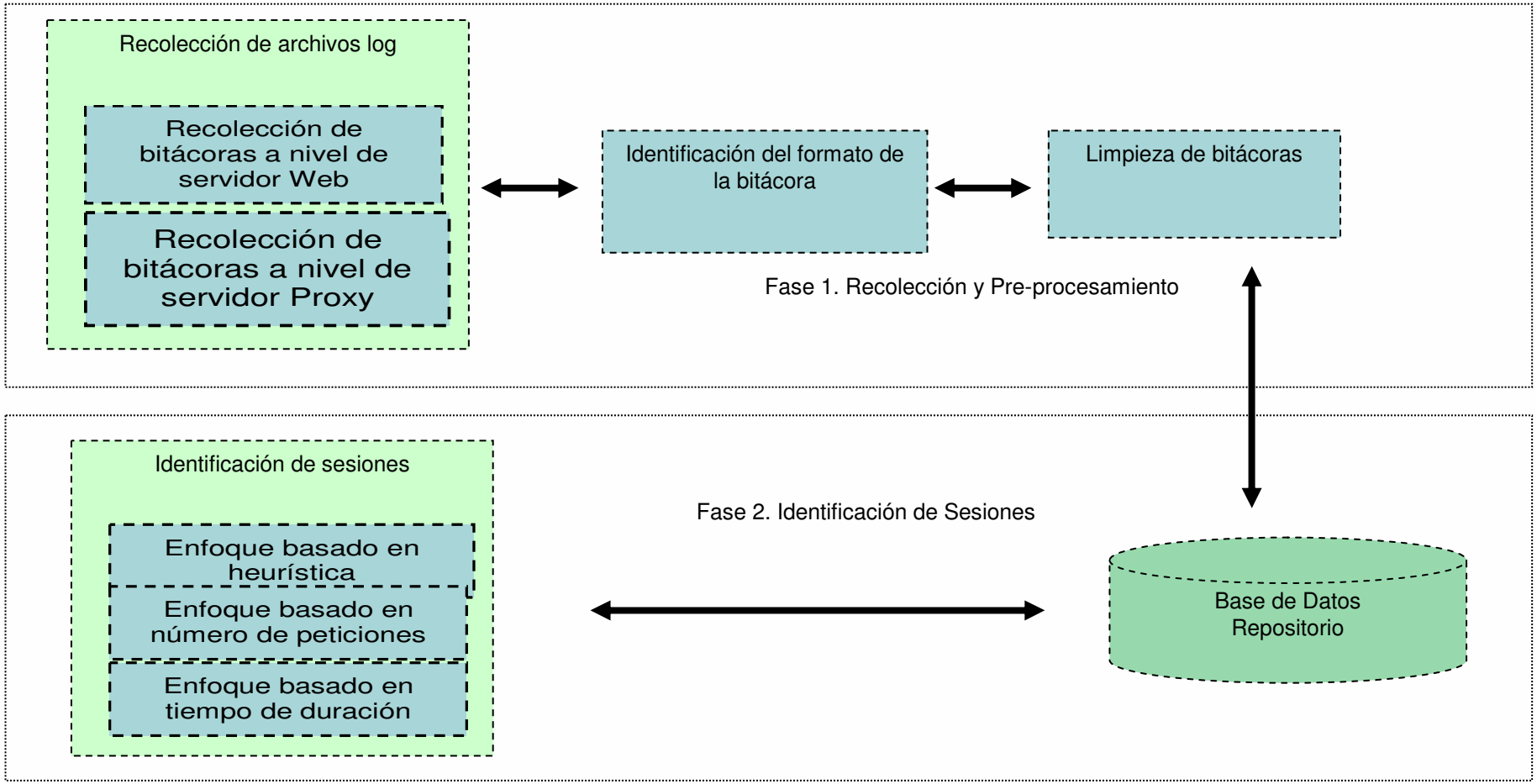
Arquitectura

- Este trabajo implementa el ciclo clásico utilizado para el descubrimiento del conocimiento



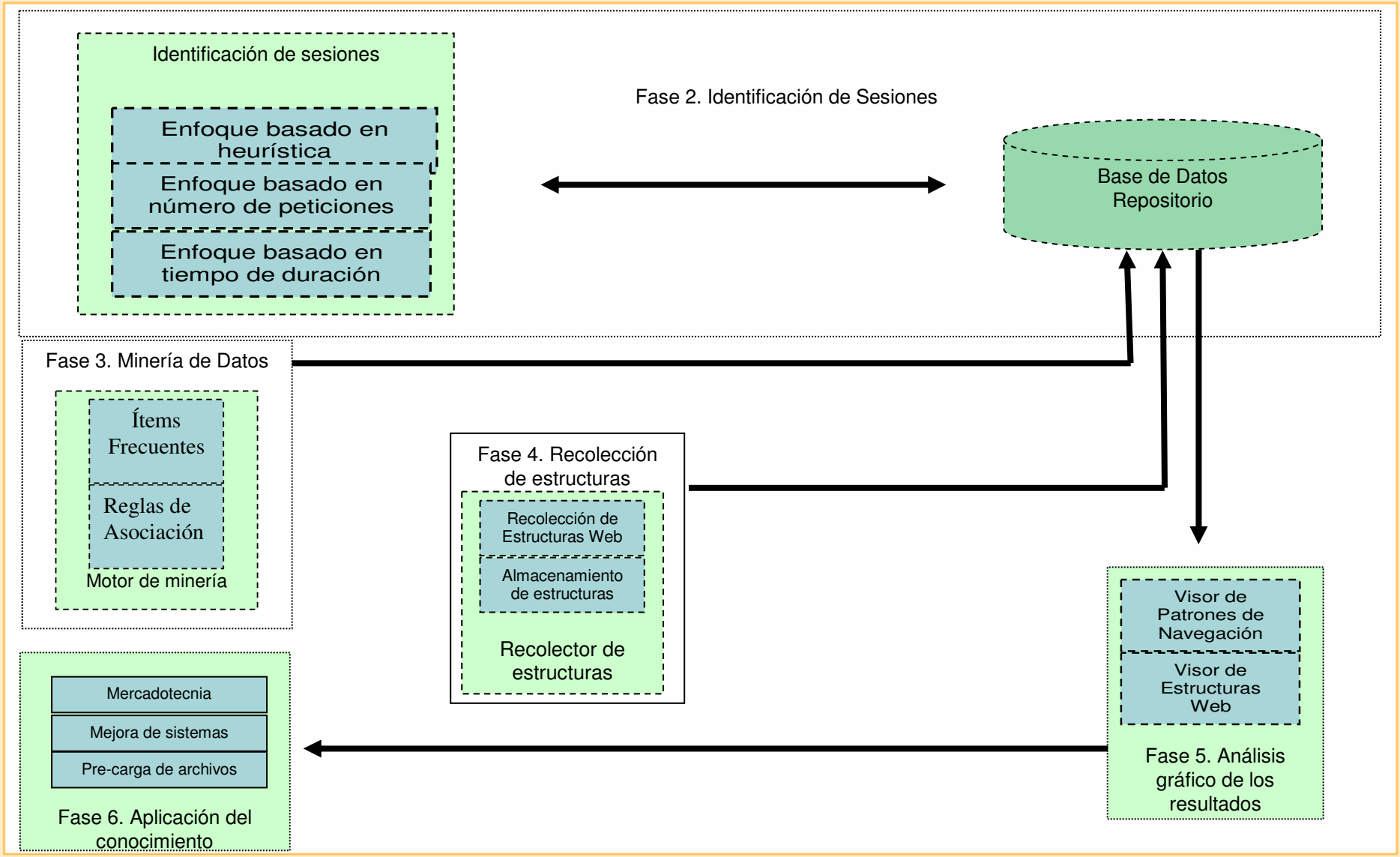
Arquitectura

- La arquitectura utilizada se detalla en la siguiente figura.



Cenidet

Arquitectura



Identificación de usuarios

- En el mejor de los casos el nombre del usuario va implícito en los archivos log, desafortunadamente, muy pocos recursos Web solicitan la identificación del usuario.

- En la ausencia de tal información, el nombre del host, el recurso Web solicitado por el usuario y el agente navegador utilizado por el usuario, son las únicas opciones que se tienen para llevar acabo la identificación de los usuarios que visitan un sitio Web.

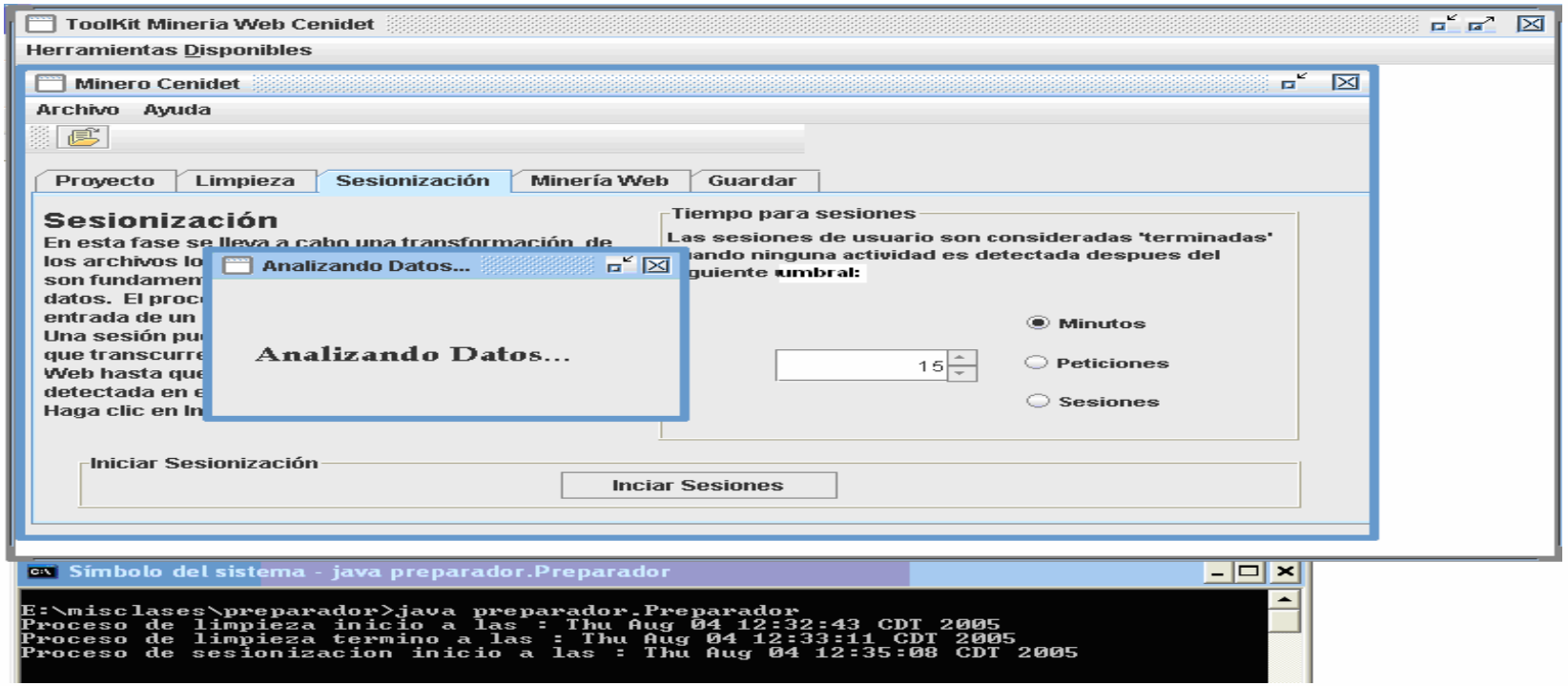
- La identificación de usuario sería trivial si se asume que cada uno de los visitantes tienen un única dirección IP asignada, pero desafortunadamente no es así, ya que la presencia de servidores Proxy por parte de los proveedores de Internet y en redes locales enmascaran a los usuarios.

Identificación de sesiones de usuario

- Una sesión en las bitácoras de solicitudes de los servidores Web incluye todos los recursos Web que un visitante solicitó durante su estancia en el sitio Web.
- Desafortunadamente, las bitácoras de los servidores Web no mantienen un control sobre los recursos solicitados durante una visita.
- Es por ello que la identificación de usuarios y sesiones de usuarios en bitácoras de servidores Web se tiene que realizar mediante mecanismos específicos.
- En este trabajo se incluyen 3 mecanismos para la identificación de usuarios y sesiones.

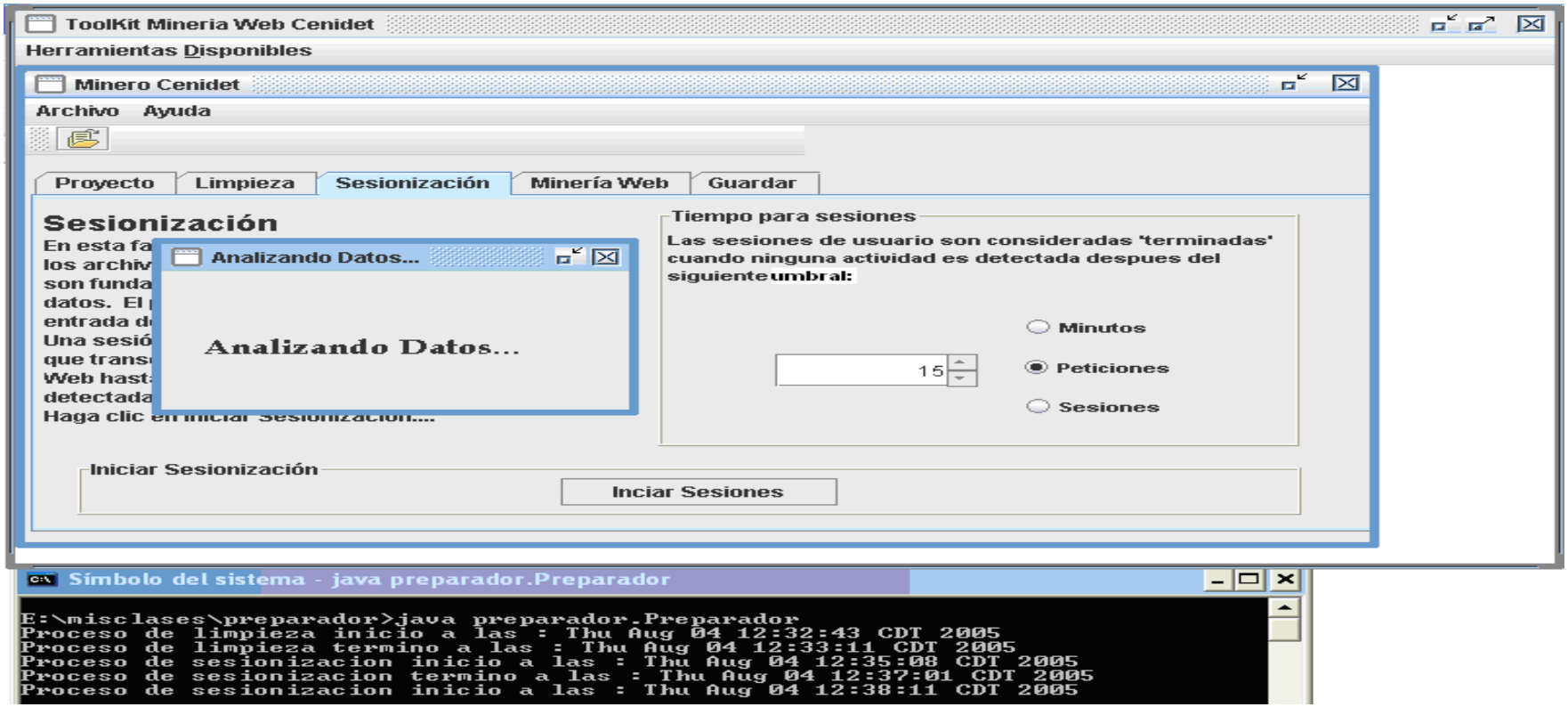
Mecanismos para la identificación de usuarios y sesiones

- El primer mecanismo se enfoca en identificar usuarios y sus sesiones que tengan un tiempo de duración determinado, es decir, una sesión inicia cuando el usuario entra al sitio Web y termina cuando el tiempo de duración indicado se alcanza o se dejan de registrar peticiones.



Mecanismos para la identificación de usuarios y sesiones

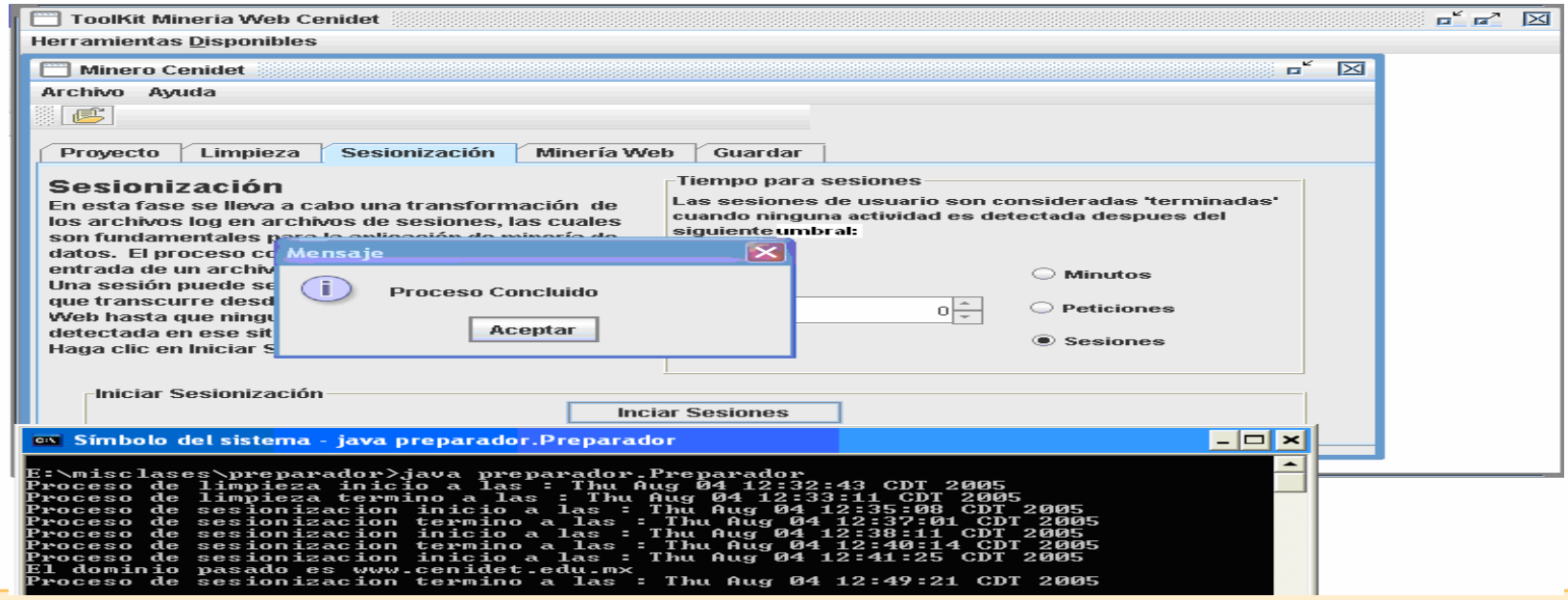
- Otro mecanismo para crear sesiones es definiendo un número máximo de recurso de pueden estar dentro de una sesión, el usuario decide el numero de recursos que puede estar contenidos durante una sesión.



Mecanismos para la identificación de usuarios y sesiones

- El tercer mecanismo es un algoritmo heurístico basado en la problemática de que un visitante no siempre está un tiempo determinado en un sitio además de que el número de recursos Web solicitados nunca está definido.

- Dicho algoritmo es capaz de identificar usuario y los recursos Web solicitados durante su visita a un sitio Web incluyendo la identificación de aquellos múltiples usuarios que se encuentran detrás de un servidor Proxy.



Búsqueda de patrones interesantes

- Una vez que se ha localizado las sesiones de usuarios, es posible aplicar técnicas para el descubrimiento de patrones sobre los datos almacenados.
- Algunos algoritmos desempeñan el análisis estadístico y otros la minería de datos
- En este trabajo se utilizaron principalmente búsqueda de ítems frecuentes y minería de reglas de asociación.

Ítems frecuentes

- Los conjuntos de ítems frecuentes pueden ser usados para relacionar las páginas más frecuentes en consultarse conjuntamente durante una sesión de usuario. Algunos ejemplos de ítems frecuentes so como los siguientes:

  - La página index.html y aspirantes.html del dominio www.cenidet.edu.mx son accedías juntas en un 20% de las sesiones de usuario registradas.

  - El archivo minero.zip y el documento minero.doc son accedidos juntos en un 12% de las sesiones de usuario.

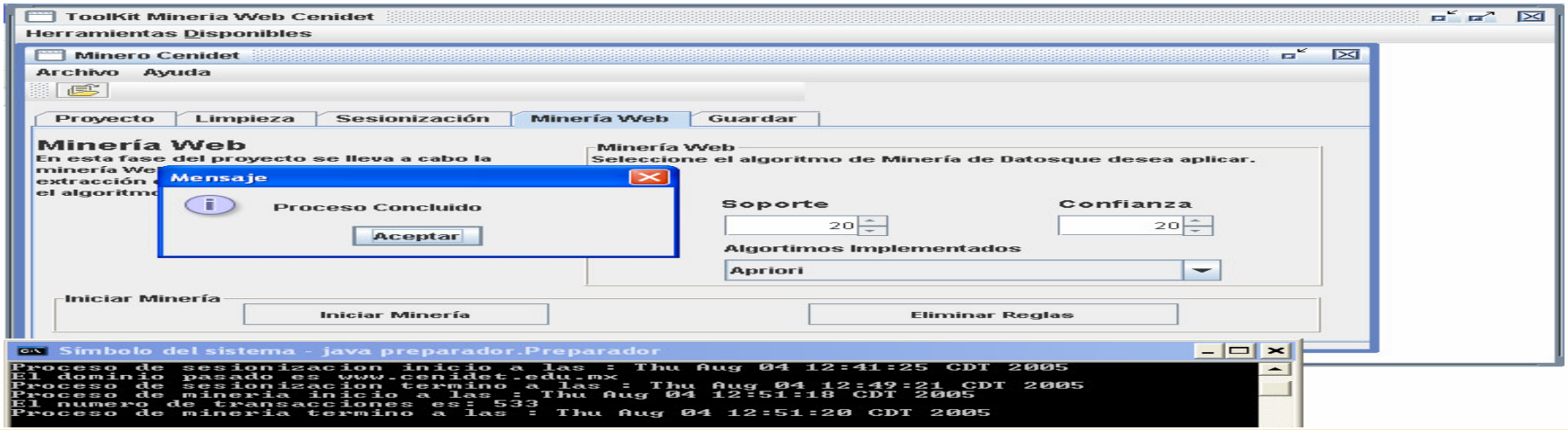
Reglas de asociación

- Cualquier conjunto de ítems frecuentes puede ser tan distante como la profundidad del árbol de navegación del sitio Web al que corresponda.
- Un conjunto de ítems frecuentes esta dado por dos elementos (A y B) los cuales puede llevar a dos reglas de asociación representadas por A→B y B→A aunque los valores que definen el nivel de interés de cada una de las reglas sea distinto, por ejemplo:
- Cuando la página index.html es accedida en una sesión, la página aspirantes.html tiene un 90% de probabilidad de ser accedida en la misma sesión.
- Cuando la página aspirante.html es accedida en una sesión, la pagina index.html tiene un 20% de probabilidad de ser accedida en la misma sesión.

Minería de reglas de asociación

- En el contexto de minería de uso Web el conjunto de ítems frecuentes y reglas de asociación se refieren a un conjunto de páginas Web que son accedidas juntas y cuya frecuencia de acceso supera un umbral mínimo especificado representado por los valores de soporte y confianza.

- Dependiendo del valor del umbral, el nivel de interés de las reglas crece o decrementa y así mismo sirve para delimitar el número de reglas generadas y permitir su manipulación y análisis.



Conclusiones

- Hemos presentado un trabajo que es capaz de encontrar reglas interesantes a partir de archivos log generados por el servidor Web y el servidor Proxy.
- Nuestro sistema puede ser usado por expertos en el área de minería de uso Web y por no expertos y cualquier administrador de sitios Web pueda analizar sus archivos log sin tener conocimientos sobre minería de datos
- Los módulos presentados y analizados en este estudio, se han implementado y probado en el laboratorio de sistemas distribuidos del Centro Nacional de Investigación y Desarrollo Tecnológico.
- Este trabajo forma parte de la plataforma middleware que dará soporte a desconexiones de usuarios en una red inalámbrica, los patrones generados por la herramienta servirán como datos entrada a un sistema que lleva a cabo el acaparamiento de archivos en dispositivos inalámbricos.

Bibliografia

- [1] Robert Cooley, Pang-Nim Tan, Jaideep Srivastava. “WebSIFT: The Web Site Information Filter System”. University of Minnesota. 1999.
- [2] F. Masseglia, P. Poncelete, M. Teisseire. “Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure”. University of Versailles.
- [3] Raymon Kosala, Hendrik Blockeel, Frnak Neven. “Web Mining Research: A Survey”. Departament of Computer Science, Katholieke Univeriteit Leuven, Belgium. 2000.
- [4] Myra Spiliopoulou, Lukas C. Faulstich. “WUM: A Web Utilization Miner“. Institut fûr Wirtschaftsinformatik, Humboultord Berlin.
- [5] R. Cooley, B. Mobasher. “Web Mining: Information and Pattern Discovery, Department of Computer Science and Engineering”. University of Minnesota, Minneapolis, USA, 1997.
- [6] Myra Spiliopoulou, Lucas C. Faulstich. “A Data Miner analyzing the Navigational Behaviour of Web Users”. Institut fûr Wirtschaftsinformatik, Humboultord Berlin.
- [7] Jaideep Srivastava, R. Cooley. “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data”. Department of Computer Science and Engineering, University of Minnesota, Minneapolis, USA.
- [8] Robert Cooley, Bamshad Mobasher, “Data preparation for Mining World Wide Web Browsing Patterns”, Department of Computer Science, University of Minnesota, October 1998.
- [9] Rakesh Agrawal, Ramakrishnan Srikant, “Fast Algorithms for Mining Association Rules”, IBM Almaden Research Center, San Jose CA, USA.
- [10] Rakesh Agrawal, Tomasz Imielinski, “Mining Association Rules between Sets Items in Large Databases”, IBM Almaden Research Center, San Jose CA, USA.
- [11] Behzad Mortazavi-Asl, “Discovering and mining user web-page traversal patterns”, Simon Fraser University, 1999.
- [12] David René Valenzuela Molina, “Mecanismos para predicción de acaparamiento de datos en sistemas cliente/servidor móviles”, CENIDET, 2002.
- [13] Robert Walker Cooley, “Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data”, Universidad de Minnesota, Mayo 2000.

¡¡¡Gracias por su atención!!!

¿Preguntas, Comentarios?