Diseño e implementación de un generador de sitios web adaptativos automáticos:

Descubrimiento de patrones de navegación

Proyecto de Taller V

Estudiantes: Luis Do Rego
Leticia Pérez

Tutor: Ing. Eduardo Fernández

INCO - Facultad de Ingeniería
Universidad de la República

Setiembre 2001

Índice

Índice de Figuras

1 Introducción

En los últimos años Internet ha surgido como una herramienta poderosa para el intercambio de información de los más diversos tipos. Texto, hipertexto, imágenes y sonido son solo parte de los múltiples medios de comunicación a través de los cuales la tecnología de Internet permite intercambiar información. Muchas tareas que históricamente se realizaban en forma presencial han comenzado a realizarse a través de Internet. El entretenimiento, el comercio, las transacciones bancarias y la educación han comenzado a utilizar Internet como medio comunicación. Ejemplo de esto son las páginas personales, shoppings on-line, información sobre cursos en diversas universidades alrededor del mundo, cursos on-line y mucho más.

La taza de crecimiento en lo que respecta al tráfico a través de Internet y tamaño de los sitios web es muy alta. Cada vez más personas se acercan a Internet tanto para buscar información como para ofrecerla. Esto genera que tareas tales como el diseño de un sitio web o simplemente la navegación a través de ellos no sean tareas simples de realizar.

Diseñar un sitio web rico en información y ágil a la hora de ofrecerla es una tarea difícil. A menudo deberán contener innumerables objetos que ofrecer a una gran diversidad de visitantes y la consecuente elección de una estructura capaz de ofrecerlos en forma intuitiva, dificulta la tarea.

Los problemas de diseño de un sitio web se deben a varios factores. Como se menciona anteriormente la cantidad de información que deben ofrecer es uno de ellos, pero más allá de esto existen otros factores. Primero, diferentes visitantes a un sitio presentan diferentes intereses. Segundo, un mismo visitante puede requerir distintos tipos de información en distintos momentos. En tercer lugar, muchos sitios a medida que pasa el tiempo van cambiando su diseño original, acumulando vínculos y agregando páginas en lugares poco accesibles. Por último, un sitio puede haberse diseñado originalmente con un fin determinado, pero en la práctica termina siendo utilizado en forma completamente diferente violando o sobrepasando las expectativas del diseñador.

Debido a esto, se ha vuelto más importante obtener información acerca de cómo están siendo utilizados los sitios web por parte de los usuarios. Obtener esta información incluye obtener datos estadísticos tales como frecuencias de acceso a las diferentes páginas dentro de un sitio así como formas de análisis más sofisticadas, tales como encontrar los caminos más frecuentemente utilizados dentro de un sitio o descubrir patrones o conductas comunes de navegación que permitan definir perfiles de visitantes. Esta información se utiliza tanto para determinar estrategias de marketing como para reestructurar un sitio Web de forma tal de atender mejor las necesidades de los visitantes.

Diferentes algoritmos de Data Mining se han comenzado a aplicar para obtener los resultados antes mencionados, generando una nueva área dentro del Data Mining llamada Web Usage Mining.

En el presente documento se presenta una forma de utilizar determinados algoritmos de Web Usage Mining para crear sitios web adaptativos. Sitios que adaptan automáticamente su estructura y presentación para un visitante en particular, de forma de atender sus intereses a medida que este interactúa con el mismo, utilizando como base información acerca de los hábitos de navegación de los diferentes visitantes.

En la sección 1.1 se define el problema a resolver. Luego en las secciones 1.2 y 1.3 se plantean los objetivos del proyecto y se define el contexto de trabajo para alcanzar los mismos, respectivamente. Por último, en la sección 1.4, se detallan las contribuciones del proyecto.

El resto del documento se organiza de la siguiente manera. En la sección 2 se presentan los conceptos fundamentales utilizados en el desarrollo de la herramienta. En la sección 3 se presenta una aproximación a la solución propuesta. A continuación, en las secciones 4, 5 y 6 se describe la solución propuesta, se presenta la implementación de la misma y se analizan los resultados obtenidos para dicha implementación. Luego en la sección 7 se detallan las conclusiones. Por último en la sección 8 se presentan diferentes líneas de trabajo a seguir como forma de continuar las investigaciones acerca la generación de sitios web adaptativos automáticos.

1.1 Definición del problema

Muchos y diversos visitantes se acercan a los diferentes sitios web en busca de información, cada uno con su propia meta e intereses. Asimismo, un visitante puede tener diferentes objetivos en diferentes visitas al mismo sitio web. Además, con el paso del tiempo, los intereses y las motivaciones de los visitantes probablemente varíen. Por esta razón la navegación de los visitantes dentro del sitio es dinámica, depende del momento y las características de los mismos. Sin embargo los sitios web son en general ambientes estáticos, tanto en lo que refiere a la estructura de los mismos como de las páginas que los componen. Esto implica que la navegación a través de los mismos sea también estática y quede sugerida a partir del grafo de navegación que define su estructura. Asimismo dicho grafo de navegación suele ser diseñado de forma de atender lo que a criterio del diseñador podrían ser los intereses de la mayoría de los visitantes pero sin un conocimiento concreto de los mismos y sin considerar intereses individuales. Esto ocasiona que la búsqueda de información requiera en ciertos casos, realizar una navegación larga y tediosa hasta localizar la información deseada. Podría pensarse que organizar la información en forma jerárquica solucionaría este problema, pero la experiencia demuestra lo contrario. ¿Quién no se ha encontrado buscando información en un sitio durante horas al tiempo que repetía “Esta información tiene que estar aquí en alguna parte?”

La solución propuesta es entonces, construir sitios con estructuras y páginas dinámicas que contemplen los intereses de la población y las particularidades de cada visitante a medida que estos interactúan con el mismo. Sitios que anticipen las necesidades del usuario y se adapten a él dinámicamente de forma tal de satisfacer sus necesidades, requiriéndole el mínimo de esfuerzo posible.

1.2 Objetivos del Proyecto

El objetivo general del proyecto es la generación de sitios web adaptativos automáticos, sitios con estructuras y páginas dinámicas que contemplen los intereses de la población y las particularidades de cada visitante a medida que estos interactúan con el mismo.

Este objetivo general se divide en dos objetivos concretos, substancialmente distintos:

- Realizar recomendaciones dinámicas de páginas a visitar, a los usuarios de un sitio web. Dichas recomendaciones pretenden acercar al usuario a sus objetivos, deducidos a partir de su conducta de navegación.

- Implementar una herramienta que permita diseñar sitios con estructuras y páginas dinámicas que se adapten a los intereses de los visitantes. Esta herramienta debe permitir diseñar páginas que generen dinámicamente links a otras páginas en función de los intereses particulares del usuario.

Este informe trata exclusivamente del diseño e implementación de una herramienta que solucione el primero de los problemas planteados anteriormente y su interfaz de comunicación con otra herramienta que soluciona el segundo de ellos.

El objetivo es entonces, realizar el descubrimiento de información acerca de como interactúan los visitantes con un sito web (conductas de navegación), tendiente a construir sitios web capaces de adaptar dinámicamente su estructura y presentación a los hábitos de los visitantes (sitios web adaptativos automáticos). Asimismo, se debe desarrollar una interfaz de comunicación con la segunda herramienta de forma tal que integradas permitan generar sitios web adaptativos automáticos.

Con respecto a la herramienta se pretende lograr un diseño modular y abierto que permita su mantenimiento (tanto correctivo como perfectivo) y extensión por medio de cambios en los módulos existentes y/o el agregado de nuevos módulos que aumenten las funcionalidades.

1.3 Contexto de trabajo

A la hora de elegir el lenguaje para la implementación de la herramienta deben tomarse en cuenta varios aspectos. En primer lugar debe ser un lenguaje que permita la portabilidad de la herramienta, ya que existen en plaza diversos tipos de servidores web para entornos Unix o Windows. Luego el lenguaje debía implementar la metodología de orientación a objetos pues esta permite implementar más fácilmente un diseño modular y alcanzar los objetivos de extensibilidad y escalabilidad de la herramienta. Además debe proveer primitivas que permitan realizar programación concurrente en una forma simple y robusta, proveyendo facilidades para la creación de hebras de ejecución (tareas), compartir recursos, sincronización y comunicación entre tareas.

Por todas estas razones se elige Java como lenguaje de implementación, ya que reúne todas las características antes descriptas. Si bien otros lenguajes soportan la mayoría de las mismas, no permiten la portabilidad de la herramienta, característica fundamental del sistema que se quiere desarrollar.

Java es un lenguaje de programación orientado a objetos. Aunque es similar a C++, es más pequeño, portable y fácil de utilizar, puesto que es más robusto y gestiona la memoria por sí mismo. Fue diseñado para ser seguro y factible de ejecutarse sobre cualquier plataforma. Los programas Java se compilan en bytecodes, que se asemejan al código de máquina y no son específicos de una plataforma, pudiendo ejecutarse en cualquier computadora que disponga de una máquina virtual Java. Esto lo hace un lenguaje útil para la programación de aplicaciones web [Microsoft2001].

1.4 Contribución del proyecto

Cada vez existe mayor interés en desarrollar aplicaciones que brinden servicios a través de Internet. En particular, la generación de sitios que se adapten al usuario y que tomen en cuenta sus intereses, es un área aún en investigación. Asimismo, la aplicación de técnicas para el análisis estadístico de los hábitos de navegación de los usuarios en Internet puede ser de gran utilidad en el ámbito comercial. Pero para ello se debe pasar primero por un estudio académico de sus reales posibilidades.

No se puede dejar de lado la contribución del proyecto en la formación de los estudiantes que integraron el equipo de desarrollo. Han obtenido amplios conocimientos acerca del funcionamiento y posibilidades de desarrollo, de aplicaciones para Internet. Para la construcción de la herramienta se vincularon áreas diferentes como ser bases de datos, Data Mining, Web Usage Mining, comunicación entre procesos y tecnología Internet. Los estudiantes adquirieron fuertes conocimientos del lenguaje Java y sus beneficios en el manejo de hilos de ejecución, recursos compartidos y comunicación entre procesos. Asimismo, el proyecto permitió a los estudiantes que participaron en el mismo, experimentar el proceso de desarrollo de una herramienta de software a lo largo de un periodo prolongado de tiempo.

2 Conceptos fundamentales

En esta sección se brindan definiciones generales de los principales conceptos utilizados en el desarrollo de la herramienta. El objetivo es que estos conocimientos sirvan de base para una mejor comprensión de los objetivos planteados y las posteriores decisiones de implementación de la herramienta

2.1 Sitios Web

Un sitio web es un conjunto de documentos HTML y archivos asociados (imágenes, video, texto, etc.) que son ofrecidos por un servidor de objetos http en el World Wide Web. Los documentos HTML ofrecidos en un sitio web, generalmente cubren uno o más temas relacionados y se interconectan a través de hipervínculos [Microsoft2001].

En general los sitios web son ambientes estáticos en lo que respecta a su estructura ya que la misma se implementa a través de los hipervínculos que interconectan las páginas. La mayoría de los sitios tienen una página de inicio como punto de entrada al mismo, que funciona generalmente como índice. Desde allí los visitantes comienzan su navegación a través del sitio guiados por los hipervínculos que encuentran en cada página visitada. De esta forma dichos hipervínculos determinan la estructura o grafo de navegación del sitio.

2.2 Sitios Web Adaptativos Automáticos

La característica fundamental de un Sitio Web Adaptativo Automático es que utiliza información acerca de los patrones de navegación de los visitantes que lo han accedido, para adaptar su organización y presentación dinámicamente a medida que un visitante navega por el mismo.

La mayoría de los sitios existentes en el World Wide Web presentan una estructura estática mientras que las necesidades de los visitantes cambian con el tiempo. Un sitio adaptativo aprende de los hábitos de navegación de sus visitantes y decide qué información presentar, cómo y cuándo. Asimismo, un diseñador elige una estructura determinada al diseñar un sitio web, pero esta estructura, no necesariamente es la mejor ni la más adecuada para cubrir las necesidades de todos los visitantes en cualquier caso. Un sitio adaptativo pretende reconocer los intereses del visitante y sugerirle, en función de estos, los caminos de navegación que mejor se adapten a ellos. Un sitio web adaptativo permite por ejemplo un rápido acceso a las páginas más populares, genera vínculos dinámicamente para conectar páginas que puedan estar relacionadas dados los intereses del visitante. En otras palabras reduce el largo del camino de navegación hacia la información buscada. Cabe aclarar que la estructura original del sitio se mantiene. No se destruyen vínculos sino que se agregan nuevos que sugieren caminos más directos para el acceso a la información.

Un sitio web puede ser adaptativo de dos formas [PE97]. La situación ideal es la personalización, donde el sitio web se adapta a cada visitante en particular, en tiempo real, atendiendo de esta forma las necesidades del mismo. Mediante la personalización, se busca predecir el objetivo final del visitante a medida que el mismo interactúa con el sitio, basándose en información obtenida en sus visitas anteriores acerca de sus gustos preferencias y patrones de navegación.

El segundo enfoque es la optimización, donde el sitio web modifica su estructura dinámicamente para atender las necesidades de cada usuario, basándose en los patrones de

navegación de todos sus anteriores visitantes. La idea de la optimización es aprender de la conducta de navegación de todos sus usuarios para modificar su estructura dinámicamente haciendo la navegación más sencilla a través de él, aún para quienes lo acceden por primera vez.

2.3 Captura de Información

Para construir sitios web adaptativos automáticos es necesario contar con información acerca del visitante, que permita inferir sus gustos y preferencias y generar, en base a esto, los vínculos entre páginas que correspondan. Esta información se puede obtener de dos formas básicas. Primero, un usuario puede proveerla explícitamente por ejemplo, a través de formularios que completa al momento de la navegación, con sus datos personales y demás datos requeridos. Sin embargo, debe tenerse en cuenta que esto no garantiza la veracidad de los datos ya que el usuario podría proporcionar datos erróneos. Otra forma de obtener información, es capturarla a medida que el usuario interactúa con el sito. Esto puede hacerse mediante el uso de cookies y/o archivos de registro de accesos al servidor web.

Cookies

El Protocolo HTTP (Hypertext Transfer Protocol) es el protocolo en el que se basa la conexión y transferencia de información entre navegadores y servidores web. Uno de sus principales inconvenientes es que es un protocolo sin estados, es decir no memoriza ni conserva información sobre anteriores conexiones del usuario. Se han desarrollado varias técnicas para evitar este inconveniente y posiblemente la más conocida y utilizada de ellas, sea el uso de cookies.

Una cookie no es más que una cadena de información, contenida en un archivo de texto con un formato determinado (ver apéndice 9.1), que el servidor web envía a un cliente acompañando la página web que este último le ha solicitado. El navegador del cliente se encarga de guardar esta información, en el disco duro del cliente, en un directorio particular. En futuros accesos al servidor web, el navegador del cliente le devolverá una copia de la cookie junto con la nueva solicitud. De esta manera el servidor web recibe la cookie y recupera la información que había enviado al cliente.

Las cookies representan una potente herramienta empleada por los servidores web para almacenar y recuperar información acerca de sus visitantes [6]. Proporcionan una forma de identificar usuarios. Permiten al servidor web recordar algunos datos concernientes al usuario, como ser sus preferencias para la visualización de las páginas, nombre y contraseña, productos que más le interesan, un número de identificación, entre otras cosas. En [8] se puede obtener mas información acerca de los diferentes usos de las cookies.

Archivos de registro de accesos al servidor web (Log de Accesos)

Por definición, en Internet el anonimato no existe, está anulado por defecto. Todas las personas que navegan por Internet dejan huellas a su paso, sin saberlo. Cada vez que un visitante solicita una nueva página, el servidor web se entera automáticamente de ciertos datos acerca de él. Esta información es enviada por el visitante al servidor web en forma automática y este último la almacena en archivos, que registran las solicitudes recibidas, llamados Archivos de registro de accesos al servidor web o también Logs del servidor web.

Las acciones que realiza el servidor web, en relación con el registro de las solicitudes recibidas, son las siguientes. Para cada archivo enviado al cliente (esto es, cada página HTML y cada elemento no textual que contiene, como botones, separadores, iconos, etc.), el servidor deja un registro en el archivo de registro de accesos. El formato e información específica que

contienen estos registros variará de un servidor web a otro, pero en un aspecto permanecen constantes: contienen toda la información necesaria para analizar la actividad del servidor web. Esta información incluye la dirección IP del usuario que accede al servidor, tiempo de acceso, url de la página accedida, un número de tres dígitos que codifica los errores de trasmisión y el número de bytes trasmitidos, entre otras cosas. Incluso, los servidores web, pueden configurarse para registrar también, en el archivo de registro de accesos, la información contenida en las cookies que envían a los navegadores que los acceden. Si pretendemos obtener datos sobre los usuarios que acceden a un servidor web o sobre qué buscan cuando acceden, no hay mejor sitio para mirar que el propio servidor.

Esta Información puede ser registrada en distintos formatos pero existen dos estándares para el registro de la información: como por ejemplo, Common Log File Format (CLF) y Extended (Combined) Log Format (ECLF) (ver apéndice 9.2).

2.4 Patrones de navegación

Como se definió anteriormente, en la sección 2.2, la característica fundamental de un Sitio Web Adaptativo Automático es que utiliza información acerca de los patrones de navegación de los visitantes que lo han accedido, para adaptar su organización y presentación dinámicamente a medida que un visitante navega por el mismo.

Los patrones de navegación de los usuarios representan el comportamiento de los usuarios al navegar por un sitio web. Más claramente, hacen referencia al transito de los usuarios a través de un sitio web.

El análisis de patrones de navegación se aplica en diferentes áreas dentro del World Wide Web. Por ejemplo, permite a las organizaciones dedicadas al comercio electrónico, crear estrategias de comercio cruzado entre productos y servicios o determinar campañas de promoción efectivas y eficientes, entre otras cosas. Asimismo, proporcionan información para reestructurar un sitio de forma de hacerlo más productivo. Permiten también, descubrir grupos de usuarios con conductas similares, que por ejemplo pueden ser utilizados a la hora de crear campañas publicitarias dirigidas a usuarios con determinado perfil.

Diferentes algoritmos de Data Mining, están siendo utilizados en la actualidad, para la obtención de patrones de navegación. Esto ha dado paso a una nueva área dentro del Data Mining, llamada Web Usage Mining. Web Usage Mining es la aplicación de técnicas de minería de datos (Data Mining) para el descubrimiento de patrones de navegación que permitan entender y atender mejor las necesidades de los usuarios en diferentes aplicaciones web. [SCDT2000].

2.5 Web Usage Mining

Web Usage Mining es el proceso de aplicar técnicas de minería de datos (Data Mining) para descubrir patrones de navegación, a partir de diferentes fuentes de datos del web, con el objetivo de utilizarlos para atender mejor las necesidades de los usuarios en diferentes aplicaciones web. [SCDT2000].

En cualquier actividad de minería en necesario seguir una determinada secuencia de tareas. Primero se debe obtener y preparar el conjunto de datos sobre el cual se realizará minería. Luego se aplican las técnicas de minería para descubrir la información deseada. Por último se analizan los resultados. De la misma forma, el proceso de Web Usage Mining se compone de tres fases llamadas preprocesamiento de datos, descubrimiento de patrones de navegación y

análisis de patrones, donde cada una de estas fases realiza cada una de las tareas antes mencionadas.

En la sección 2.5.1 se identifica el conjunto de datos sobre el cual se aplican técnicas de minería para el descubrimiento de patrones. A continuación, en las secciones 2.5.2 a 2.5.4 se detallan las distintas fases del proceso de Web Usage Mining. Finalmente, en la sección 2.5.5 se presentan diferentes herramientas que utilizan Web Usage Mining para el descubrimiento de patrones de navegación

2.5.1 Datos del Web

Una etapa clave para el proceso de descubrimiento de información o minería de datos, es la obtención del conjunto de datos sobre el cual aplicar técnicas de minería para el descubrimiento de patrones de navegación.

En Web Usage Mining los datos sobre los cuales aplicar técnicas de minería para el descubrimiento de patrones pueden obtenerse tanto del lado del cliente como del lado del servidor. La diferencia entre estos datos no esta dada únicamente por su localización física sino también por su disponibilidad y los métodos para recolectarlos. Tanto del lado del cliente como del lado del servidor existen diferentes fuentes de datos desde las cuales extraer información.

Fuentes de Datos

En el presente documento se hace referencia únicamente a la información disponible en los servidores web y la forma en que ésta puede ser utilizada en el descubrimiento de patrones de navegación. Una descripción más detallada de las diferentes fuentes de datos del lado del cliente y su utilización se detalla en [SCDT2000]

Como se mencionó anteriormente, en la sección 2.3, es posible obtener información acerca de los visitantes a un sitio y del uso que los mismos hacen de él, a medida que estos interactúan con el mismo. La captura de información a medida que el usuario navega por el sitio es realizada por los servidores web, a través del uso de cookies y archivos de registro de accesos.

Los archivos de registro de accesos de los servidores web son una importante fuente de información sobre la cual realizar minería, pues implícitamente almacenan la conducta de navegación de los visitantes al sitio. Los datos almacenados en estos archivos reflejan los accesos (posiblemente concurrentes) al sitio web, realizados por múltiples usuarios.

También, los servidores web pueden proporcionar otros recursos para obtener información como pueden ser las cookies. Un servidor web puede generar una cookie para cada navegador en particular que lo accede, como una forma de rastrear usuarios a través del sitio.

Abstracciones de datos

Luego de obtenidos los datos desde las fuentes anteriormente mencionadas, es necesario realizar abstracciones sobre ellos de forma tal de construir el conjunto de datos sobre el cual aplicar técnicas de minería para el descubrimiento de patrones. El conjunto de datos a construir a partir de los datos recolectados en los servidores web se basa en la identificación de usuarios y sesiones de usuarios.

Un usuario se define como un individuo que accede a un sitio web a través de un navegador[8].

Una sesión de usuario es la secuencia ordenada de páginas solicitadas por un usuario durante su visita a un sitio web [8].

2.5.2 Preprocesamiento de datos

Las anteriores definiciones de usuario y sesión de usuario pueden parecer triviales, pero en la práctica la identificación única y repetida de usuarios, así como la obtención de las sesiones de los diferentes usuarios dentro del sitio, no es una tarea simple de realizar. Esto se debe, entre otras cosas, a que por ejemplo un usuario puede acceder a una página desde diferentes máquinas o usar más de un navegador desde la misma máquina, lo cual dificulta su identificación. De la misma forma, un usuario puede acceder muchas veces a un sitio durante el período de captura de información a través del archivo de registro de accesos, lo cual genera que se deba implementar una forma de dividir el conjunto de sus accesos en las diferentes sesiones, dentro del período considerado. Por esta razón se hace necesaria una fase de preprocesamiento de los datos recolectados en los servidores web, de forma tal de prepararlos para la aplicación de las técnicas de minería.

La etapa de preprocesamiento del proceso de Web Usage Mining, consiste en la obtención de las abstracciones definidas anteriormente, usuarios y sesiones de usuarios, a partir de la información recolectada en los servidores web, tendiente a construir el conjunto de datos sobre el cual aplicar técnicas de minería para el descubrimiento de patrones de navegación.

2.5.3 Descubrimiento de patrones de navegación

Algunas de las técnicas utilizadas más comúnmente para el descubrimiento de patrones de navegación son análisis estadísticos, descubrimiento de reglas de asociación, descubrimiento de patrones secuenciales y clustering. [SCDT2000]

Análisis estadístico

El análisis estadístico es el método más común para obtener información acerca de los visitantes a un sitio web. Mediante el análisis del archivo de registro de accesos al servidor web, se puede obtener información estadística como por ejemplo frecuencia de acceso a las diferentes páginas, tiempos medios de duración de las visitas de los usuarios, largo promedio de las sesiones de los usuarios, entre otros.

Existen diversas herramientas en el mercado que realizan este tipo de análisis estadístico basadas en el archivo de registro de accesos al servidor web [11]. Ejemplo de estas herramientas son *Analog* [1], *Wusage7* [2], *Openwebscope* [3], *Webalizer* [4], Getstat [9], WWWStat [10] entre otras. Todas estas herramientas toman el archivo de registro de accesos al servidor web, generalmente en alguno de los formatos estándar, y lo analizan para obtener información estadística, que en general puede ser seleccionada por el usuario. Luego, con esta información generan reportes, por lo general en formato html. La información que se puede obtener cubre desde descubrir las páginas más populares del sitio hasta identificar desde que países acceden los usuarios a un sitio web.

Este tipo de información es especialmente útil para mejorar la performance de las aplicaciones web, permite a los diseñadores reestructurar los sitios web para volverlos más productivos o también puede ser utilizada como soporte para la toma de decisiones de marketing.

Reglas de asociación

La técnica de descubrimiento de reglas de asociación, en el contexto del Web Usage Mining, se utiliza para descubrir todas las asociaciones y correlaciones entre accesos a las diferentes páginas disponibles en un sitio web, donde la presencia de un conjunto de páginas en una sesión de usuario implica, (con cierto grado de confianza) la presencia de otras páginas.

[SCDT2000]. Estas páginas no necesariamente deben estar interconectadas a través de hipervínculos.

Usando estas técnicas se pueden encontrar correlaciones o asociaciones tales como que un visitante que accede a una página con información acerca de instrumentos musicales, más precisamente saxo, también accede a páginas con información acerca de libros de música jazz y páginas de discos de Charlie Parker.

Esta información permite a las organizaciones relacionadas con el comercio electrónico desarrollar estrategias eficaces de comercialización. La existencia y también la ausencia de tales reglas pueden dar una indicación de cómo reestructurar el sitio web de una organización, a fin de tornarlo más productivo.

Patrones secuenciales

La técnica de descubrimiento de patrones secuenciales permite encontrar patrones, dentro de las sesiones de usuario, en los cuales la presencia de un conjunto de páginas es seguida por otra página. En este caso, a diferencia de las reglas de asociación, importa el orden en que son visitadas las páginas. Esta técnica permite descubrir relaciones tales como que, usuarios que acceden previamente a páginas acerca de libros de tecnología Internet luego acceden a la página que contiene definiciones de términos Internet.

El descubrimiento de patrones secuenciales permite predecir la conducta de los visitantes dentro de los sitios web. Esto puede utilizarse para diseñar campañas publicitarias y otras estrategias de marketing en forma eficiente.

Clustering

La técnica de clustering permite agrupar conjuntos de items con características similares bajo algún criterio definido por el usuario. En el contexto del Web Usage Mining existen dos tipos interesantes de clusters o agrupaciones a descubrir, *clusters de sesiones de usuario* y *clusters de páginas.*

Los clusters de sesiones de usuario, representan un grupo de sesiones de usuario similares según la ocurrencia de accesos a páginas durante las mismas. De esta forma se establecen grupos de usuarios que presentan conductas o patrones de navegación similares. Esta información es especialmente útil para facilitar el desarrollo y ejecución de estrategias de comercialización, en y fuera de línea, tal como el envío de correo automatizado a los usuarios que presenten un determinado comportamiento dentro del sitio.

Por otra parte, los clusters de páginas, agrupan páginas que tienden a ocurrir juntas, dentro de las sesiones de los usuarios. Si consideramos que las páginas que un usuario visita durante una sesión dentro del sitio, tienden a estar conceptualmente relacionadas, entonces los clusters de páginas descubren grupos de páginas cuyo contenido esta relacionado de alguna forma para los usuarios. No se asume que todas las páginas visitadas por un usuario dentro de una sesión deban estar relacionadas, por esta razón debe trabajarse con información acumulada durante largos períodos de tiempo acerca de las visitas realizadas a un sitio por parte de los diferentes usuarios.

Ambas técnicas permiten modificar dinámicamente la estructura de un sitio web mediante la creación de hipervínculos entre páginas cuyo contenido esta conceptualmente relacionado para determinados usuarios según el análisis de su conducta a través del sitio, la cual se ve reflejada en los clusters obtenidos.

2.5.4 Análisis de patrones

El análisis de patrones es el último paso en el proceso de Web Usage Mining (ver figura 2.1). El conjunto de patrones de navegación obtenido en la etapa anterior es analizado para eliminar de él aquellos patrones que no resulten de interés. Según el tipo de aplicación web para la cual se quieren aplicar los patrones obtenidos, serán las técnicas de análisis que se deben utilizar. Las técnicas más comúnmente utilizadas consisten en la aplicación de lenguajes de consulta, como por ejemplo SQL, sobre los patrones obtenidos. Otro método es almacenar los patrones en un cubo de datos y realizar sobre este, operaciones con herramientas OLAP. También pueden ser empleadas técnicas de visualización, como por el ejemplo la representación gráfica de los patrones descubiertos.

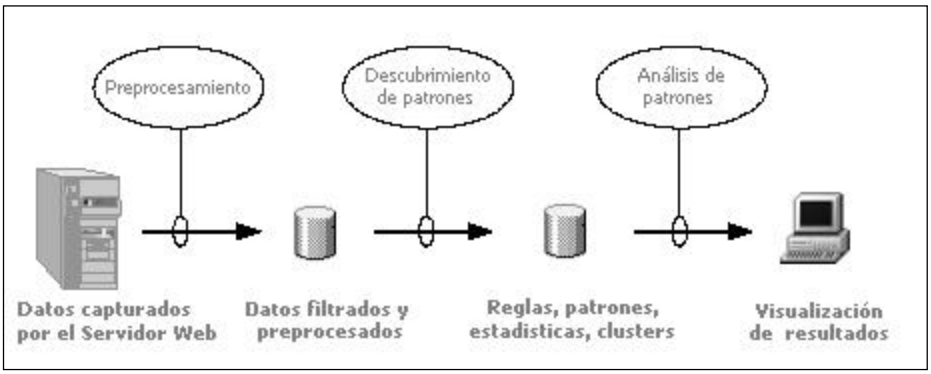


Figura 2.1: Proceso de Web Usage Mining

.

2.5.5 Aplicaciones

Diferentes estudios se han realizado sobre la aplicación de las distintas técnicas de Web Usage Mining para el descubrimiento automático de patrones de navegación.

Spiliopoulou et al. [SF99], Cooley et al. [CMS99] aplican técnicas de Web Usage Mining para extraer patrones de navegación, a partir de los archivos de registro de accesos a servidores web, para ser aplicados en estrategias de marketing.

Spiliopoulou et al. [SPF99] propone una metodología basada en el descubrimiento y comparación de patrones de navegación de clientes y no clientes a un sitio web comercial, como forma de mejorar la estructura del mismo.

Yan et al. [YJGGD96] y Nasraoui et al [NFJK99] proponen el uso de clusters de sesiones de usuario para predecir la conducta futura de los usuarios dentro de un sitio web.

Perkowitz y Etzioni [PE97, PE98] proponen la generación de sitios web adaptativos mediante la aplicación de técnicas de clustering, aplicadas a la información contenida en los archivos de registro de accesos a servidores web.

Mobasher et al. [MJHS97] describen la aplicación de las técnicas de descubrimiento de reglas de asociación y patrones secuenciales, para obtener patrones de navegación a partir de los datos recolectados por un servidor web.

Mobasher, Cooley y Srivastaba [MCS99, MCS2000] han realizado estudios en el área de la generación de sitios web adaptativos. En particular utilizan y comparan las técnicas de clustering y descubrimiento de reglas de asociación, aplicadas a las sesiones de usuarios descubiertas en los archivos de registro de accesos a los servidores web, en la generación de sitios web adaptativos.

3 Aproximación a la solución

Anteriormente, en la sección 2, se presentó una definición detallada de sitio web adaptativo automático. Asimismo, se mencionó el hecho de que para alcanzar la adaptabilidad del sitio se pueden seguir dos criterios diferentes, la personalización o la optimización. En esta sección se comparan ambos criterios y se define cuál será el criterio a utilizar y de que forma.

3.1 Personalización vs. Optimización

Tanto la personalización como la optimización adaptan dinámicamente la estructura de un sitio web a los posibles intereses de los usuarios basándose en su conducta de navegación actual. La diferencia radica en la información adicional utilizada para esto.

Ambos criterios adaptan la estructura del sitio web a cada visitante en tiempo real, atendiendo de esta forma sus necesidades. Para realizar esta tarea el criterio de personalización se basa exclusivamente en la conducta de navegación de cada visitante, en cambio el criterio de optimización toma en cuenta la conducta de navegación de todos los visitantes al sitio. La personalización adapta el sitio web basándose en los intereses particulares de cada visitante mientras que la optimización se basa en los intereses de todos los visitantes. La optimización aprende de la conducta de todos los visitantes para que el sitio sea más fácil de utilizar, incluso para aquellos que nunca lo han utilizado antes.

Para lograr la personalización es necesario capturar, procesar y almacenar grandes volúmenes de datos ya que por cada visitante se debe mantener información en forma individualizada, de manera tal de recuperarla en futuros accesos del mismo visitante. Sin embargo la optimización, si bien también maneja grandes volúmenes de datos, requiere una capacidad de almacenamiento mucho menor ya que guarda información no individualizada, manteniendo solamente grupos o perfiles de visitantes.

Bajo el criterio de personalización es necesario reconocer un visitante, cuando este accede al sitio, de forma tal de recuperar la información que se tiene del mismo, con el fin de personalizar el sitio según sus intereses. Esto implica desarrollar una técnica de identificación de usuarios que permita reconocerlos cada vez que acceden nuevamente. Esta tarea no es simple de realizar y las técnicas que lo consiguen no son aplicables en todos los casos. Por el contrario la optimización no requiere identificar a un visitante cuando este accede al sitio. Sin importar quien sea el visitante se busca una relación entre su sesión activa y alguno de los grupos o perfiles obtenidos, para adaptar el sitio automáticamente a lo que se asume, a partir de estos perfiles, pueden ser sus intereses.

3.2 Solución propuesta

Si bien la personalización es el ideal, la optimización se torna más aplicable. Por esta razón se elige la optimización como forma de implementar sitios adaptativos automáticos.

El primer paso para lograr la optimización del sitio es utilizar los métodos de captura de información, descriptos en la sección 2.3, para recoger datos acerca de los visitantes. Sobre estos datos se aplican técnicas de Web Usage Mining, a fin de obtener patrones de navegación que permitan clasificar a los visitantes en diferentes grupos o perfiles. Luego de establecidos los grupos, en tiempo real durante la sesión de un visitante dentro del sitio, se observa su conducta de navegación actual para determinar a qué grupo pertenece y así adaptar dinámicamente la estructura del sitio a sus intereses.

Según lo descrito anteriormente se distinguen tres etapas en la generación de sitios web adaptativos automáticos bajo el criterio de la optimización. Estas etapas se definen como *captura y procesamiento de información, descubrimiento de patrones de navegación* y *recomendación dinámica de páginas*

Captura y procesamiento de información

Para capturar información se utilizan los archivos de registro de accesos al servidor web durante un período prolongado de tiempo que permita disponer de una muestra uniforme de accesos. También se utilizan cookies, enviadas por el servidor web a los navegadores que lo acceden, cuyo valor es un número de identificación de usuario. Este número de identificación se registra junto con la información de la solicitud, en el archivo de registro de accesos al servidor web. Esto permite identificar a que usuario pertenece cada acceso registrado en el archivo. Luego se procesa la información contenida en dicho archivo para generar el conjunto de sesiones de los usuarios que accedieron al servidor web durante el período de captura de información.

Descubrimiento de patrones

Para obtener los patrones de navegación, se utiliza una de las técnicas de Web Usage Mining propuestas en [MCS2000], basada en la utilización de algoritmos de clustering de páginas, aplicados sobre el conjunto de sesiones de usuario, que permitan descubrir grupos de páginas cuyo contenido está conceptualmente relacionado para los usuarios del sitio web.

Recomendación dinámica de páginas

La recomendación dinámica de páginas es la técnica utilizada para adaptar dinámicamente la estructura de un sitio web a los intereses de los visitantes. Se realiza en tiempo real, a medida que un usuario navega por el sitio, sobre la base de los grupos de páginas descubiertos previamente, información estadística y la sesión activa cada visitante.

4 Descripción de la solución

4.1 Arquitectura general

La arquitectura propuesta para la generación de sitios web adaptativos automáticos, según el criterio de la optimización, se divide en dos etapas [MCS2000]. La primer etapa se realiza off-line y se encarga de preparar los datos obtenidos, para aplicar sobre ellos técnicas de minería, y de la posterior aplicación de las mismas. La segunda etapa se realiza on-line y se encarga de adaptar dinámicamente la estructura del sitio a través de recomendaciones dinámicas de páginas. Estas dos etapas se especifican como dos componentes bien diferenciadas.

La primer componente, off-line, lleva a cabo las dos primera fases del proceso de Web Usage Mining, preprocesamiento de datos y descubrimiento de patrones de navegación. Se encarga de preprocesar los datos contenidos en el archivo de registro de accesos al servidor web, obteniendo así el conjunto de sesiones de usuario. Luego, sobre el conjunto obtenido, se aplica la técnica de clustering de páginas, obteniendo como resultado un conjunto de clusters de páginas que representan distintos patrones de navegación dentro del sitio.

La segunda componente, on-line, utiliza los clusters obtenidos por la componente off-line y datos estadísticos que obtiene analizando el conjunto de sesiones de usuario, para proporcionar recomendaciones dinámicas de páginas, a los visitantes del sitio según sus sesiones activas. Esta componente mantiene la sesión activa de un visitante mientras que el navegador del mismo hace peticiones HTTP. Las recomendaciones son solicitadas a medida que el usuario navega por el sitio y se envían al visitante, a través de un protocolo de comunicación diseñado para esto.

A continuación se presenta la arquitectura propuesta para la herramienta:

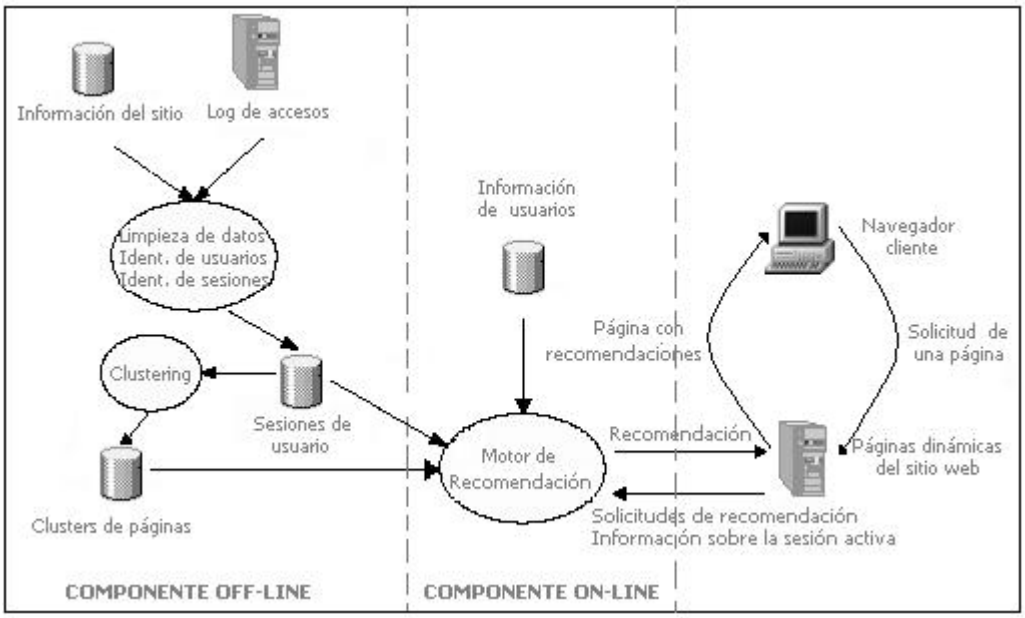


Figura 4.1: Arquitectura General

4.2 Componente off-line

4.2.1 Preprocesamiento de datos

El preprocesamiento de datos es el paso previo a la aplicación de cualquier técnica de minería. Aquí se analiza el log del servidor web[1], para obtener el conjunto de las sesiones de los usuarios. Esto implica analizar cada entrada en el archivo de registro de accesos, seleccionando solo aquellas que provean información útil. Luego se debe determinar que usuario originó cada entrada, de forma de poder obtener las diferentes sesiones de los usuarios.

Limpieza de los datos

La limpieza o filtrado de los accesos registrados en el log del servidor web tiene como objetivo eliminar registros no deseados, o sea registros que no aportan información relevante para la minería de datos. La información descubierta o los datos estadísticos obtenidos son útiles solamente si los datos contenidos en el log del servidor presentan un cuadro real de los accesos de los usuarios al sitio Web.

El protocolo HTTP (protocolo de trasmisión de datos por Internet) requiere realizar una nueva conexión con el servidor web por cada archivo que se solicite al mismo. Dado que una página web contiene diferentes objetos, (imágenes, sonido, etc.), además del archivo html en sí, una solicitud de un visitante de acceder a una página determinada da lugar a menudo a varias entradas en el log del servidor, una por cada objeto enviado al visitante. En la mayoría de los casos, solamente la entrada correspondiente a la petición del archivo HTML es relevante y debe ser utilizada para obtener las sesiones de usuario. Esto es así porque, en general, un visitante no solicita explícitamente todos los gráficos que están en una página web, sino que estos le son enviados automáticamente por el servidor web según lo indicado por las etiquetas HTML, que construyen la página. Puesto que el objetivo principal al aplicar técnicas de Web Usage Mining es obtener los patrones de navegación de los usuarios dentro del sitio, no tiene sentido incluir las solicitudes que no fueron realizadas explícitamente por el usuario. La eliminación de los registros no deseados puede ser lograda controlando el sufijo de la dirección URL del objeto solicitado. Por ejemplo, todas las entradas en el log del servidor con sufijo GIF, JPEG, JPG pueden ser quitadas, ya que corresponden a imágenes, contenidas en una página solicitada por el visitante.

También es importante considerar que no todas las solicitudes realizadas por un usuario a un servidor web terminan en accesos exitosos. Toda solicitud http a un servidor web, sea exitosa o no, genera una entrada en el log de accesos. El servidor web registra en el log, el resultado de la solicitud (ver apéndice 9.2). Dado que en la tarea de minería interesan las páginas efectivamente accedidas del sitio, las entradas en el registro que indiquen un acceso insatisfecho son considerados registros no deseables y por lo tanto no son utilizados.

Por último es necesario filtrar los accesos contenidos en el archivo de log, en función de la estructura física del sitio para evitar considerar páginas caducas o actualmente inexistentes.

Identificación de usuarios

Otra de las tareas de preprocesamiento implica la identificación de usuarios, para poder luego identificar las sesiones de los mismos. Como se mencionó anteriormente, en el log de acceso del servidor web queda registrada la dirección IP de cada usuario que accede al mismo. Podría pensarse que esta información es suficiente para lograr la identificación de los usuarios.

[1] *Log del servidor web* es el otro nombre con el que se conoce al *archivo de registro de accesos al servidor web.*

Esto no ocurre así, principalmente a causa de la asignación dinámica de direcciones IP y la existencia de los servidores proxy.

Los servidores proxy mapean todos los host de su red en una sola dirección IP, la suya. Esto implica que en el log de acceso del servidor web, todas las peticiones que llegan a través de un proxy tienen la misma dirección IP, aunque provengan de más de un usuario.

La forma en que se resuelve este problema es a través del uso de cookies. La solución implica enviar una cookie al navegador de cada visitante la primera vez que este accede al sitio. La cookie contendrá un número de identificación, que el navegador almacena en el disco duro del visitante. Luego en futuras solicitudes al servidor web, el navegador envía la cookie junto con la solicitud. Basta configurar el servidor web para que registre el valor de la cookie, en el archivo de log junto con la restante información de la solicitud, para poder identificar que accesos corresponden a que usuario.

Sin embargo, estas técnicas no siempre pueden ser utilizadas debido a niveles de seguridad impuestos por el navegador del usuario. No se puede forzar a un visitante a aceptar las cookies. Por esta razón se han desarrollado varias heurísticas simples basadas en los campos del referrer y agente en log del servidor web (ver apéndice 9.2), que también pueden utilizarse para identificar sesiones de usuario [CMS99].

Identificación de sesiones

La tarea de identificación de sesiones de usuario requiere de una previa identificación de usuarios. Como se mencionó anteriormente esto puede ser resuelto mediante el uso de cookies. Luego de identificados los usuarios se debe afrontar el problema de identificación de sesiones.

Dado que el log de accesos de un servidor web almacena información de largos períodos de tiempo, es probable que un usuario visite el sitio más de una vez durante este período. El objetivo de identificar las sesiones de usuario es dividir el conjunto de todos los accesos a páginas, en sesiones individuales. El método más simple de obtener las sesiones es mediante un timeout donde, si el tiempo transcurrido entre dos solicitudes sucesivas excede cierto limite se asume que el usuario ha comenzado una nueva sesión. Datos empíricos sugieren que 25.5 minutos es un timeout apropiado. [CMS99].

Existen ciertos problemas en la identificación de sesiones de usuario, que es importante tener en cuenta y para los cuales no existen soluciones definitivas. Idealmente, las sesiones de los usuarios que accedieron a un sitio web permitirían extraer información de quienes tuvieron acceso al sitio web, qué paginas solicitaron, en qué orden y cuanto tiempo duró la visita a cada página. Pero existen impedimentos para poder identificar en forma exacta una sesión de usuario. Dos de los impedimentos más grandes son los servidores proxy y el almacenamiento local (caché) de los navegadores.
Para mejorar el funcionamiento y reducir al mínimo el tráfico de la red, la mayoría de los navegadores almacenan las páginas solicitadas en memoria. De esta forma, cuando un usuario quiere volver a visitar una página, su navegador la busca en su archivo caché y se la ofrece sin realizar una nueva solicitud al servidor web. De igual forma, para acelerar la navegación, los servidores proxy almacenan en archivos caché las páginas solicitadas por un usuario, para ofrecerlas al próximo usuario que las solicite, sin realizar un nuevo acceso al servidor web. Esto produce que no todas las páginas que visita un usuario queden registradas en el log del servidor afectando la veracidad de las sesiones de los usuarios.

4.2.2 Descubrimiento de patrones de navegación

Como se explicó anteriormente, en la sección 2.5, Web Usage Mining es la técnica utilizada para realizar el descubrimiento de patrones de navegación. En particular, para la generación de sitios web adaptativos automáticos se necesita clasificar a los visitantes en diferentes grupos o perfiles según sus patrones de navegación, los cuales serán utilizados en la etapa de recomendación descripta mas adelante en el documento.

La técnica de clustering aplicada al conjunto de sesiones de usuario, obtenido a partir del archivo de registro de accesos al servidor web, permite descubrir clusters o grupos de sesiones de usuario similares. La similitud entre sesiones se define en base a la ocurrencia de accesos a páginas en dichas sesiones. En otras palabras, la similitud se mide en función de la cantidad de páginas en común que presenten las sesiones. De esta forma, los clusters de sesiones representan conductas o patrones de navegación similares dentro del sitio. Intuitivamente, usuarios con una conducta similar dentro de un sitio web poseen intereses similares al acceder al mismo. Entonces los clusters de sesiones sugieren los distintos perfiles de usuario que navegan a través del sitio.

Por esta razón, la técnica de clustering es la elegida para realizar el descubrimiento de patrones de navegación que permitan lograr la adaptabilidad de un sitio web.

Aplicación de la técnica de clustering

Para aplicar algoritmos de clustering sobre un conjunto de datos, es necesario representar dichos datos en un formato que facilite la aplicación del algoritmo. En este caso en particular el algoritmo se aplica sobre el conjunto de sesiones de usuarios, entonces se debe elegir una forma de representación para las mismas
Suponiendo que existen ***N*** páginas dentro del sitio web, se representa una sesión ***s*** de un usuario mediante un vector de dimensión ***N*** con el siguiente formato:

$\boldsymbol{s = (p_1,p_2,p_3,...,p_N)}$ donde p_i=1 si la i-ésima página pertenece a la sesión y 0 en caso contrario.

Ciertas investigaciones [SZAS97, YJGD96] sugieren utilizar, en lugar de los valores binarios (0,1) para los p_i, valores basados en el tiempo que un usuario permanece en una página determinada o la frecuencia de aparición de una página dentro de la sesión de usuario considerada. Sin embargo, ninguna de estas propuestas es más intuitiva o justificable que otra en el contexto de las sesiones de usuario. Estudios realizados [KMM+97] sugieren que para un usuario determinado, la cantidad de tiempo que el mismo permanece en una página generalmente no es un buen indicador del interés del mismo en dicha página. Esto se debe a que el tiempo trascurrido entre solicitudes consecutivas de un usuario no necesariamente es el tiempo que el usuario permanece interactuando con la página. Del mismo modo, la cantidad de veces que se accede a una página en una sesión, no es generalmente una buena medida de la importancia de la página para el usuario. Basta con pensar en páginas que solo contienen hipervínculos a otras páginas y por lo tanto pueden llegar a ser accedidas varias veces en el transcurso de una sesión, sin que su contenido sea de interés sino para ser utilizadas como índices.
Lo que realmente importa es determinar si un usuario visita o no a una determinada página. Potencialmente cualquier página del sitio puede ser de interés para algún usuario, y lo importante es determinar si una página fue visitada o no en el transcurso de una sesión. Por esta razón se eligen valores binarios para p_i[MCS2000].

Para obtener los clusters de sesiones de usuario se necesita alguna medida (métrica) que permita medir la similitud entre sesiones de forma de agrupar aquellas que sean similares.

Anteriormente se mencionó que dos sesiones de usuario se definen similares en función de la ocurrencia de accesos a páginas dentro de las mismas. Por ello se utiliza como métrica el coseno normalizado del ángulo que forman los vectores que las representan [MCS2000]. Este coseno se calcula como el producto escalar entre los vectores, divido entre el producto de sus normas, o sea:

$$similitud(t,s) \equiv \frac{\sum_{i\equiv 1}^{N}(t_i \times s_i)}{|t| \times |s|}$$ Siendo *t* y *s* dos vectores que representan, cada uno, una sesión de usuario

El producto escalar entre los vectores *t* y *s*, $\sum_{i\equiv 1}^{N}(t_i \times s_i)$, da como resultado la cantidad de páginas que tienen en común las sesiones representadas por ellos. La multiplicación de la coordenada t_i por la coordenada s_i da como resultado 1 si la página *i* fue visitada en ambas sesiones y 0 en otro caso. Luego al realizar la sumatoria se obtiene la cantidad total de páginas en común. Finalmente se normaliza este resultado para que la cantidad de páginas visitadas en cada sesión afecte él calculo al compararse sesiones con distinta cantidad de paginas visitadas. De esta forma, a través del cálculo del coseno normalizado del ángulo que forman los vectores, se obtiene el grado de similitud entre las sesiones que representan.

En base a la métrica definida anteriormente, que permite cuantificar la similitud entre sesiones, se aplica un algoritmo de clustering que permita agrupar aquellas que sean similares. Existen diversos algoritmos de clustering que pueden utilizarse en esta tarea..

Una vez particionado el conjunto de sesiones de usuario, mediante la aplicación de un algoritmo de clustering se obtiene un conjunto de clusters de sesiones C de la forma:

$C = \{C_1, C_2, ..., C_K\}$

Cada cluster de sesiones agrupa sesiones de usuario similares, bajo el criterio de similitud definido anteriormente. De esta forma, se puede decir que los clusters obtenidos agrupan rutas de navegación similares a través del sitio. Por lo tanto cada cluster obtenido representa un perfil de usuario dentro del sitio considerado.

El objetivo final a alcanzar es la recomendación de páginas a un usuario que se encuentra navegando por el sitio, en función de las que ya ha visitado y de las que han visitado usuarios con una conducta similar a la suya. Estas últimas se obtienen a partir de los clusters hallados. Sin embargo, los clusters de sesiones de usuario hallados, no son una representación adecuada para realizar la recomendación de páginas. Esto se debe a que cada cluster puede tener cientos de sesiones de usuario que a su vez hacen a varias páginas. Conviene convertir los clusters de sesiones a clusters de páginas. Así, cada cluster de páginas agrupa páginas del sitio cuyo contenido esta relacionado para un determinado perfil de usuario.

Para obtener el cluster de páginas asociado a un cluster de sesiones, se calcula un vector m_c de la siguiente forma:

$m_c = (m_1, m_2, ..., m_N)$ donde m_i es el número promedio de apariciones de la i-ésima página considerando todas las sesiones que integran el cluster ***c*** y ***N*** es la cantidad de páginas del sitio web.

Cada m_i se calcula como el total de las visitas que obtuvo la página i, considerando todas las sesiones del cluster c, dividido entre la cantidad de sesiones que integran dicho cluster. Así, el vector m_c obtenido representa un conjunto de páginas, las cuales tienen asociado un peso m_i. Dicho peso m_i corresponde al promedio relativo de visitas que recibió la página i, por parte de usuarios que presentaron un determinado patrón de navegación.

Es recomendable aplicar un filtro a los grupos o clusters de páginas obtenidos para eliminar de ellos aquellas páginas con un bajo promedio de visitas. Es posible definir una cota μ tal que todas las páginas con un promedio de visitas menor que dicha cota, sean quitadas del cluster.

En este punto se han obtenido los clusters de páginas a partir de los cuales se realizan las recomendaciones dinámicas a los usuarios. El proceso de recomendación es tarea de la componente on-line y se detalla en la sección siguiente.

4.3 Componente On-line

Una vez finalizadas las tareas de Web Usage Mining y habiéndose obtenido los clusters de páginas que representan los diferentes patrones de navegación dentro del sitio, la componente on-line genera recomendaciones en forma dinámica a los usuarios a medida que estos navegan por el sitio. Para esto utiliza su sesión activa, los clusters de páginas obtenidos y también información estadística.

Dado que uno de los factores a considerar a la hora de realizar recomendaciones es la conducta actual del usuario dentro del sitio, es necesario mantener información acerca de las sesiones activas de los usuarios que se encuentran navegando a través del sitio web. Esta tarea es también llevada a cabo por la componente on-line.

Por último esta componente brinda la interfaz de comunicación con las herramientas cliente, de forma de recibir peticiones de recomendación y enviar las respuestas correspondientes con el fin de adaptar el sitio a los intereses del usuario que encuentra visitándolo.

4.3.1 Motor de recomendación

La tarea del motor de recomendación es determinar un conjunto de recomendaciones para un usuario a partir de su sesión activa, los clusters de páginas obtenidos y también de resultados estadísticos. El motor de recomendación acepta solicitudes de recomendación a medida que el usuario navega por el sitio y envía las correspondientes respuestas a través de una protocolo de comunicación diseñado para ello.

La sesión activa muestra la conducta de navegación del usuario indicando las páginas que este ha visitado hasta el momento. Los clusters de páginas se utilizan para identificar que páginas visitan los usuarios con patrones de navegación similares. Por último los resultados estadísticos permiten obtener información como por ejemplo las páginas mas visitadas del sitio o las páginas mas visitadas por un usuario en particular.

- **Recomendación basada en patrones de navegación**

La recomendación de páginas basada en patrones de navegación consiste en obtener un conjunto de páginas que un usuario podría estar interesado en visitar, dado que así lo hicieron los usuarios que presentaron un patrón de navegación similar al suyo. Las páginas a recomendar se obtienen a partir de los clusters de páginas.

Se deben tener en cuenta tres factores al momento de realizar las recomendaciones [MCS2000], estos son:

1. Definir un criterio para determinar que cluster de páginas debe ser utilizado en la recomendación, en función de la sesión activa del usuario.
2. No recomendar páginas que ya han sido visitadas por el usuario en su sesión activa.
3. Determinar que porción de la sesión activa de un usuario será considerada para decidir que cluster de páginas utilizar en recomendación de páginas.

Criterio de comparación

Como se explicó anteriormente los clusters de páginas obtenidos agrupan páginas cuyo contenido esta relacionado para un determinado perfil de usuario. Estos clusters son representados como vectores de la forma:

$c = (u^c_1, u^c_2, ..., u^c_N)$ donde **u^c_i** es el cociente entre el total de apariciones de la i-ésima página del sitio, considerando todas las sesiones que integran el cluster de sesiones desde el cuál se obtuvo **c,** entre la cantidad total de sesiones de dicho cluster.

El valor **u^c_i** se conoce como el peso de la i-ésima página, dentro del cluster *c*. ***Peso(i,c)*** **$= u^c_i$**

Como se definió anteriormente las sesiones de usuario se representan a través de vectores de la siguiente forma:

$s = (s_1, s_2, ..., s_n)$ donde s_i vale 1 si el usuario ha visitado la página i o 0 en caso contrario.

Cuando un usuario comienza una sesión dentro del sitio, se debe observar su conducta de navegación, representada por su sesión activa dentro del sitio, para determinar qué cluster de páginas utilizar al momento de realizar recomendaciones. De esta forma, el primer paso a seguir es decidir a cuál de los patrones de navegación descubiertos, se asemeja mas a la conducta del usuario que se encuentra navegando por el sitio.
Para esto se calcula el grado de similitud entre cada cluster y la sesión activa del usuario [MCS2000]. Se aplica el mismo criterio utilizado para obtener el grado de similitud entre sesiones:

$$similitud(s,c) \equiv \frac{\sum_{i=1}^{N}(c_i \times s_i)}{|s| \times |c|}$$

Donde, ***s*** es la sesión activa de usuario, ***c*** es un cluster de páginas y ***N*** es la cantidad de páginas del sitio web.

Se calcula el producto escalar entre el vector ***c*** que representa el cluster de páginas y el vector ***s*** que representa la sesión del usuario. Luego se normaliza este resultado dividiéndolo entre la norma del vectores ***c*** y ***s***.
El producto escalar entre los vectores ***c*** y ***s*** determina el grado de similitud entre el cluster y la sesión en función de la cantidad de páginas que tienen en común, ponderadas por el peso de las mismas dentro del cluster ***c***.

Porción de sesión activa utilizada

Para determinar la porción de la sesión activa de un usuario s_L, que será considerada en el proceso de realizar recomendaciones, es utilizada una ventana corrediza de tamaño fijo ***L*** sobre la sesión activa. Esto quiere decir que solamente se toman en cuenta las ***L*** últimas páginas visitadas por un usuario en el momento de realizar la recomendación. Esto se debe principalmente a que un usuario podría presentar múltiples y variados objetivos al visitar un sitio, durante una misma sesión. Entonces no se debe permitir que las elecciones de navegación

realizadas por un usuario en un momento, con un objetivo específico, afecten al resto de la sesión donde su objetivo pudo haber cambiado.

Una tarea importante dentro del proceso de recomendación es determinar el tamaño óptimo de la ventana. Análisis realizados en esta área concluyen que la cantidad promedio de páginas visitadas en las sesiones de usuario dentro de un sitio, es un valor adecuado a utilizar [MCS2000].

Proceso de recomendación a partir de los clusters de páginas

Al momento de realizar recomendaciones a un usuario, se debe decidir que clusters utilizar y cuáles de sus páginas conviene recomendar. Para esto es asignando un valor de recomendación para cada página de cada cluster, calculado en función de la sesión activa del usuario y del grado de similitud de esta con el cluster al que pertenece la página [MCS2000]. Por lo tanto, el valor de recomendación de una página es calculado como:

$recomendacion(s,u,c) \equiv \sqrt{peso(u,c) \times similitud(s,c)}$ siendo *u* una página del sitio, *s* la sesión activa del usuario y *c* el cluster al que pertenece la página *u*

El grado de similitud entre el cluster *c* y la sesión activa *s* determinan el grado de semejanza entre el patrón de navegación representado por el cluster *c* y la sesión activa del usuario. Cuanto mayor sea la semejanza, mayor será el valor de similitud. Por su parte, el peso de la página *u* dentro del cluster *c* es un indicador del interés que presenta la página *u* para los usuarios con conducta de navegación similar a la representada por el cluster *c*, por lo tanto su valor es utilizado en el cálculo del valor de recomendación.
De esta forma, al recomendar una página, se toma en cuenta la conducta actual del usuario, para determinar a que perfil corresponde y en función de esto recomendar las páginas mas visitadas por usuarios con este perfil.

El cálculo puede aplicarse a todas las páginas de todos los clusters. Así, todas las páginas tendrán un valor de recomendación asignado. El conjunto de páginas a recomendar se construye ordenando el conjunto de páginas del sitio en orden descendente de valor de recomendación.

Se pueden utilizar diferentes criterios a la hora de decidir que páginas recomendar al usuario. Se puede decidir recomendar solamente las páginas con mayor valor de recomendación o definir una cota η que determine que solo se recomendarán paginas con un valor de recomendación mayor que η. Esto dependerá de cada caso particular.

- **Recomendación basada en resultados estadísticos**

La recomendación de páginas basada en resultados estadísticos consiste en recomendar una página que presenta determinadas características, desde el punto de vista estadístico, que la vuelven lo suficientemente interesante como para recomendarla a un visitante.
En particular se resuelven dos consultas estadísticas:

- La página mas accedida por el visitante que se encuentra navegando por el sitio.
- La pagina mas accedida por todos los visitantes del sitio.

La información para resolver estas consultas se obtiene a partir de los datos obtenidos luego de realizar la fase de preprocesamiento de datos.

4.3.2 Mantenimiento de sesiones de usuarios

Las sesiones activas de los usuarios que se encuentran navegando por el sitio son un factor importante cuando se quieren realizar recomendaciones dinámicas a los mismos. Por ello, es necesario mantenerlas con el fin de poder consultarlas y decidir en base a ellas, qué páginas recomendar. El mantenimiento de sesiones requiere realizar el seguimiento de los usuarios a través del sitio obteniendo cada una de las páginas que visitan. A medida que se obtiene esta información es necesario enviarla a la herramienta para la construcción y mantenimiento de las sesiones activas de cada usuario que se encuentra visitando el sitio.

El seguimiento de usuarios se realiza a través de cookies. Cuando un cliente realiza su primer petición de una página a un servidor web mediante un navegador, se le envía una cookie con un número de identificación. Este número de identificación es utilizado a medida que el usuario navega por el sitio, para poder determinar que páginas visita.

A medida que cada usuario realiza solicitudes de páginas estas solicitudes deben enviarse a la herramienta para que construya y actualice la sesión activa de cada usuario. El envío de esta información se realiza a través de la interfaz de comunicación que se describe en la siguiente sección.

4.4 Interfaz de comunicación

Como se explicó anteriormente la herramienta desarrollada debe interactuar con otras herramientas que permitan diseñar sitios web adaptativos. La interacción de las mismas permite la generación de sitios web adaptativos automáticos. De esta forma se pueden considerar a las herramientas que diseñan sitios web adaptativos como herramientas cliente de la herramienta desarrollada. Con el fin de obtener una herramienta independiente del diseño e implementación elegidos para las herramientas cliente, se define una interfaz de comunicación basada en la recepción de solicitudes de recomendación y el envío de las respuestas. Para esto se implementa un protocolo de comunicación basado en el envío y recepción de paquetes que transportan las solicitudes y las respuestas. Asimismo se define un lenguaje para realizar las peticiones de recomendación. De esta forma las herramientas cliente envían solicitudes de recomendación utilizando el protocolo definido y reciben las recomendaciones que deberán hacer llegar a cada visitante del sitio.

También esta interfaz de comunicación es utilizada para enviar a la herramienta la información de que páginas solicita cada usuario, permitiendo mantener las sesiones activas de los mismos.

Lenguaje de consulta (DQL)

El lenguaje de consulta se utiliza para realizar solicitudes de recomendación (apéndice 9.4). Estas pueden ser de dos tipos. Se pueden solicitar recomendaciones de tipo estadístico, como por ejemplo las páginas mas visitadas del sitio, o solicitudes a partir de los patrones de navegación.

Protocolo de comunicación

El protocolo de comunicación establece cómo se deben realizar las solicitudes de recomendaciones y como se envían las respuestas. Asimismo, establece como se envía la información acerca de que páginas solicita cada usuario a medida que navega por el sitio. Este protocolo maneja diferentes tipos de paquetes: *paquetes de consulta*, *paquetes respuesta y paquetes de seguimiento.* Estos paquetes se envían y se reciben utilizando UDP.

- *Paquetes de consulta*

Estos paquetes son utilizados por la herramienta cliente para realizar un pedido de recomendación. Se identifican a partir del campo Tipo de paquete, donde deberán contener una letra "C". El pedido se especifica mediante el lenguaje de consulta definido (DQL). La carga útil del paquete es la consulta DQL.

Tipo de paquete C (consulta)	Identificador	Consulta en lenguajeDQL	Fin de paquete

- Paquetes respuesta

Este tipo de paquetes es utilizado por la componente on-line para enviar a la herramienta cliente la respuesta a sus solicitudes de recomendación. Se identifican a partir del campo Tipo de paquete, donde deberán contener una letra "R". La carga útil de los mismos contiene la página que se recomienda según la petición recibida.

Tipo de paquete R (respuesta)	Identificador	Recomendación	Fin de paquete

- *Paquetes de seguimiento*

Estos paquetes son utilizados para brindar información a la componente on-line acerca de las páginas que visitan los usuarios y así poder realizar el mantenimiento de las sesiones activas de los mismos. Son enviados por el cliente y no requieren respuesta ni confirmación. Se identifican a partir del campo Tipo de paquete, donde deberán contener una letra "T"La carga útil de los mismos contiene un par (usuario, página) que indica que un determinado usuario visita una determinada página.

Tipo de Paquete T (seguimiento)	(Usuario, Página)	Fin de paquete

- Especificación del protocolo (SDL)

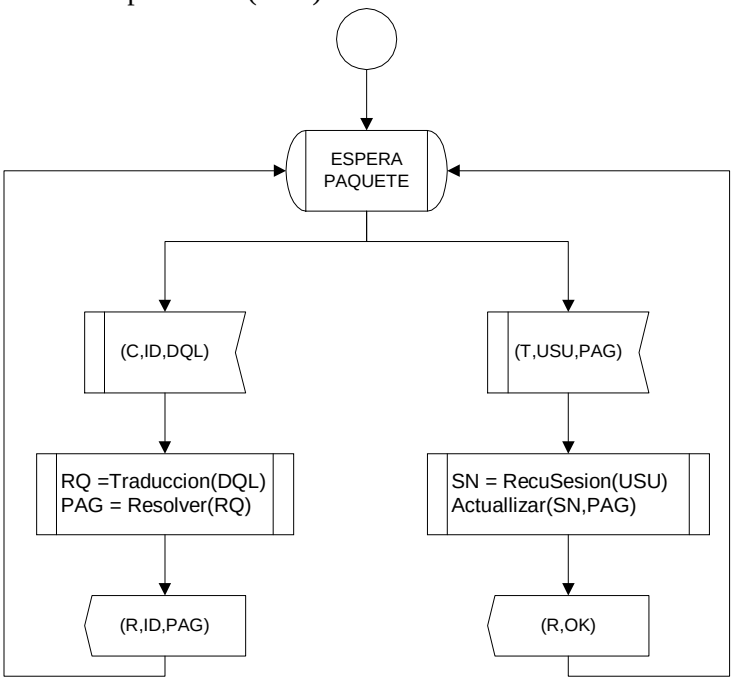


Figura 4.2: Especificación del protocolo de comunicación.

5 Implementación de la solución

En esta sección se detalla la implementación de la herramienta para la generación de sitios web adaptativos automáticos. Se presenta el diseño modular donde cada módulo cumple una función especifica. Dentro de estos módulos se encuentran las componentes on-line y off-line descriptas en la sección anterior

5.1 Diseño modular

El enfoque utilizado en la implementación de la herramienta es un diseño modular y orientado a objetos. Esto permite alcanzar el objetivo de extensibilidad de la herramienta por medio de cambios en los módulos existentes y/o el agregado de nuevos módulos que aumenten las funcionalidades.

Para alcanzar dicho objetivo la herramienta desarrollada se divide en módulos, relacionándose estos según muestra la figura 5.1

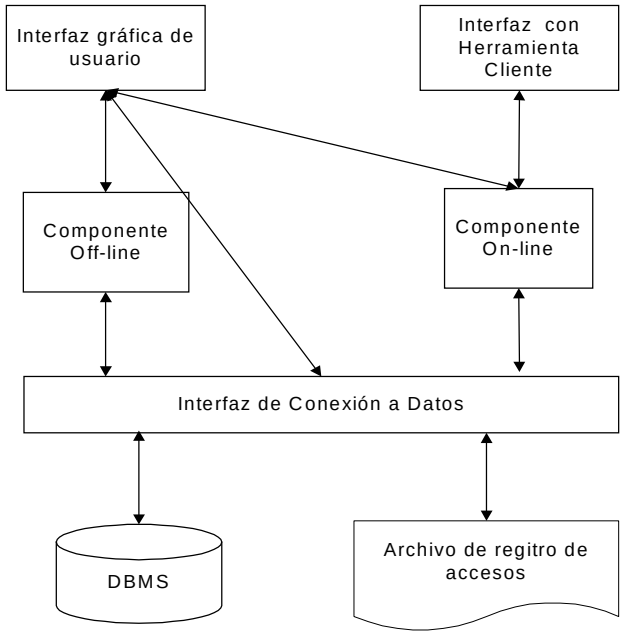


Figura 5.1: Diseño modular

DBMS

Hace referencia al almacenamiento físico de los datos en un manejador de base de datos.

Archivo de registro de accesos

Hace referencia al archivo de registros de acceso al servidor web que se utiliza como información de partida para el proceso de generación de sitios web adaptativos automáticos

Interfaz de conexión a datos

Para lograr independencia entre las distintas componentes y el almacenamiento físico de los datos, se implementa un módulo que actúa como interfaz entre los mismos. Esta interfaz se utiliza para el acceso a la base de datos y para el acceso al archivo de registro de accesos al servidor web.

Al ofrecer una interfaz para la comunicación con el manejador de base de datos se reduce el impacto producido en la herramienta al cambiar el manejador. De esta forma frente a un cambio del manejador, los cambios a la herramienta quedan encapsulados dentro del módulo interfaz de conexión a datos.

De igual forma, ofrecer una interfaz de comunicación con el archivo de registro de accesos al servidor web, permite independizar las componentes que realizan el procesamiento de los datos, del formato del archivo. Como se explicó anteriormente existen diferentes formatos para el registro de accesos al servidor web, al tener un módulo que se encargue de la lectura de los registros y su correspondencia en estructuras internas independientes del formato de los mismos se logra que el formato del archivo sea transparente para las componentes.

Se implementa a través de clases que realizan la conexión al almacenamiento físico de datos, tanto para recuperarlos como para almacenarlos y clases para representar estos datos en un formato adecuado

Componente Off-Line

Este módulo implementa la componente off-line descripta en la sección anterior. Realiza las etapas de preprocesamiento de datos y descubrimiento de patrones de navegación, del proceso de Web Usage Mining.

Se implementa a través de clases que se encargan de realizar las distintas tareas de la etapa de preprocesamiento: limpieza de datos, identificación de usuarios e identificación de sesiones y clases que se encargan de implementar el algoritmo de clustering para el descubrimiento de patrones de navegación

Componente On-line

Este módulo implementa la componente on-line descripta en la sección anterior. Realiza el mantenimiento de las sesiones de usuario y de la recomendación dinámica de páginas a medida que los usuarios navegan por el sitio web.

Interfaz Gráfica de usuario

Este módulo permite la interacción entre el usuario y la herramienta. Esta interacción se realiza a través de componentes gráficos como ser ventanas y formularios que permiten al usuario ingresar al sistema y utilizar las distintas funcionalidades de la herramienta.

Interfaz con Herramientas Cliente

La herramienta desarrollada debe interactuar con herramientas que permitan diseñar sitios web adaptativos. Esta interfaz permite la comunicación con dichas herramientas cliente mediante la recepción de solicitudes de recomendación y el envío de la respuesta. También esta interfaz de comunicación es utilizada para enviar a la herramienta, la información de qué páginas solicita cada usuario, permitiendo el mantenimiento de las sesiones activas de los mismos.

Mediante un conjunto de clases se implementa el protocolo de comunicación con la herramienta cliente y un traductor para el lenguaje de consulta DQL diseñado para solicitar las recomendaciones.

6 Resultados experimentales

Para evaluar los resultados de la solución propuesta se utilizó el archivo de registro de accesos al servidor web de la Facultad de Ingeniería. Los accesos registrados en dicho archivo cubren el período comprendido entre el 8 de febrero del 2001 al 2 de marzo del 2001.

Para realizar la tarea de preprocesamiento del archivo de log se realizaron las siguientes consideraciones. Se tomó 25 minutos como el máximo tiempo transcurrido entre dos accesos consecutivos dentro de una misma sesión [CMS99]. Asimismo, se tomaron en cuenta solamente, sesiones en las cuales se hubieran visitado por lo menos tres páginas. Esta depuración se realiza, pues sesiones en las que se hayan visitado una o dos páginas no aportan información relativa a los patrones de navegación del usuario. De esta forma se obtuvo un conjunto de 5826 sesiones de usuario. El número de páginas diferentes accedidas durante el trascurso de estas sesiones fue 2261.

6.1 Evaluación de patrones descubiertos

Para particionar el conjunto de sesiones en los distintos clusters de páginas se utilizó el algoritmo Vector Quantization (ver apéndice 9.3), utilizando como métrica el grado de similitud entre una sesión y un cluster[2], el cuál implica el grado de similitud entre una sesión y todas las sesiones pertenecientes al cluster[3].

A diferencia de otros algoritmos de clustering, Vector Quantization (VQ) no requiere conocer a priori el número de clusters que se deben obtener, sino que genera un nuevo cluster cada vez que lo considera necesario. Para su ejecución, solo se necesita el valor de una constante ρ que indica el mínimo grado de similitud entre un cluster y una sesión, para que dicha sesión pueda pertenecer al cluster.
El algoritmo comienza con un conjunto vacío de clusters, la primera sesión examinada dará lugar a la creación del primer cluster. Luego para cada sesión de usuario se determina a que cluster debe pertenecer, en función de su grado de similitud con cada uno de ellos. La nueva sesión pertenecerá al cluster para el cuál haya obtenido mayor grado de similitud siempre y cuando sea mayor que la cota mínima ρ. Si el grado de similitud entre esta nueva sesión y los clusters existentes no supera la cota mínima ρ, se crea un nuevo cluster que solo contiene dicha sesión.

Dado que el algoritmo VQ requiere una constante ρ de entrada, para evaluar los resultados del mismo se hizo variar la constante ρ. El valor de la constante varía entre 0 y 1 ya que el grado de similitud entre la sesión y el cluster se calcula en función del coseno del ángulo que forman los vectores que los representan. Por esta razón se analizaron los resultados para ρ = {0, 0.2, 0.4, 0.6, 0.8, 1}

---

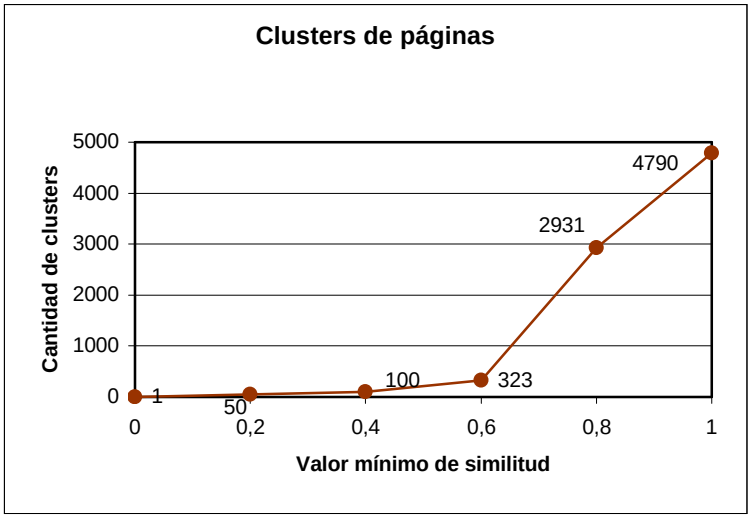
[2] $similitud(s,c) \equiv \frac{\sum_{i=1}^{n}(c_i \times s_i)}{|s| \times |c|}$ Donde, s es la sesión activa de usuario, c es un cluster de páginas y N es la cantidad de páginas del sitio web.

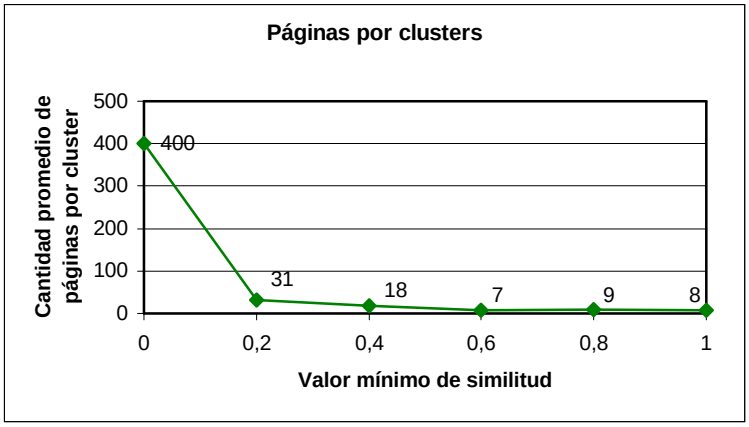
[3] $similitud(t,s) \equiv \frac{\sum_{i=1}^{N}(t_i \times s_i)}{|t| \times |s|}$ Siendo *t* y *s* dos sesiones de usuario y N es la cantidad de páginas del sitio web.

Los resultados en todos los casos mostraron que las páginas web de cada área de trabajo de la facultad (cursos, cuestiones administrativas, investigación, extensión, gestión), se agruparon en diferentes clusters. Se encontró un cluster especifico de cada curso de la facultad, un cluster para las páginas de biblioteca, un clusters para bedelía, otro para los servicios administrativos, etc. También se encontraron clusters que vinculaban dos áreas de trabajo, por ejemplo dos cursos o dos áreas de investigación, diferentes y no necesariamente del mismo instituto. La variación en los resultados se aprecia en que a medida que ρ aumenta, los clusters tienden a subdividirse, volviéndose cada vez más específicos dentro de cada área de trabajo y los pesos de las páginas que los componen son cada vez mas altos, tendientes a 1. Esto ocasiona el aumento del número de clusters descubiertos y la disminución del número de páginas por cluster.

Cuantificando los resultados se observa:

Cota	***Clusters***
0	1
0,2	50
0,4	100
0,6	323
0,8	2931
1	4790





Cuanto mayor es ρ, mayor es el número de clusters que se obtienen. Esto se debe a que cuanto más grande se hace ρ, más restrictiva es la condición de similitud entre una sesión y un cluster, para que esta pueda pertenecer a él. Se observa por ejemplo que cuando ρ=0 se crea un solo cluster, ya que ρ=0 indica que no es necesario que exista similitud alguna entre las sesiones

que integran un mismo cluster. Por el contrario, en el caso ρ=1 se forman 4490 clusters pues ρ=1 implica que las sesiones que integran un mismo cluster deben ser idénticas.

También es posible observar que cuanto mayor es ρ, menor es la cantidad de páginas en cada cluster. Esto se debe a que cuanto mayor es ρ menor es la cantidad de sesiones que integran cada cluster ya que la restricción en la similitud se hace mas fuerte. Al disminuir las sesiones por cluster, es probable que el número de páginas por cluster también disminuya

Dada la cantidad de clusters obtenidos se hace imposible mostrar los resultados relativos a las páginas que los componen en cada caso, parte de ellos se presentan en el apéndice 9.5.1. Mediante el análisis de los mismos, se observa que a medida que aumenta ρ, no solo aumenta la cantidad de clusters sino que estos se hacen más específicos dentro de cada área de trabajo y los pesos[4] de las páginas que los componen están cada vez mas cerca de uno. Esto se debe a que cuanto mayor es la similitud entre las sesiones que componen un cluster mayor es la cantidad de páginas que tienen en común. Al mismo tiempo disminuye él número de sesiones que integran cada cluster pues se hace mas fuerte la restricción. Así, el peso de cada página, o lo que es lo mismo, el número promedio de apariciones de cada página considerando todas las sesiones que integran un cluster, aumenta tendiente a 1.

Los resultados obtenidos muestran que es posible descubrir patrones o conductas de navegación comunes entre los usuarios de un sitio web, a través de un algoritmo de clustering aplicado a información obtenida desde un archivo de registros de acceso a un servidor web. Mas precisamente se obtuvieron conjuntos de páginas que son visitados por la mayoría de los usuarios que presentan una conducta de navegación determinada.

Para el caso particular del algoritmo Vector Quantization, el grado mínimo de similitud a utilizar para obtener los clusters, depende del grado de especificidad dentro de cada área de trabajo, que se pretenda alcanzar en ellos y este dependerá de cada sitio en particular donde se quiera aplicar la herramienta.

6.2 Recomendaciones dinámicas

Para evaluar las recomendaciones dinámicas, se eligió un tamaño de ventana de 4, ya que se observo que este es el promedio de páginas visitadas en cada sesión de usuario dentro del sitio web de la Facultad de Ingeniería. Asimismo, se eligió un conjunto de sesiones reales, extraídas del archivo de registro de accesos de la Facultad de Ingeniería, para realizar las pruebas.

A continuación se presenta una tabla que muestra una de las sesiones de usuario utilizadas para la evaluación de resultados y las recomendaciones obtenidas, con su correspondiente valor de recomendación. Otros ejemplos de sesiones utilizadas pueden encontrarse en él apéndice 9.5.2. El valor de recomendación mínimo a partir del cuál se ofrecen recomendaciones es 0.3.

Sesión de usuario paso a paso	Página recomendada	Valor
/~electiva/	/~electiva/Semestre1/index.html	0,42
	/~electiva/Semestre1/GestCalidad.html	0,42
	/~electiva/Semestre1/CompGrafica.htm	0,42
/~electiva/Semestre1/admin.htm	/~electiva/Semestre1/SistInfoGeografica2.html	0,4

[4] $peso(i,c) \equiv u_i^c$ donde $\mathbf{u^c_i}$ es el número promedio de apariciones de la i-ésima página considerando todas las sesiones que integran el cluster *c*.

	/~electiva/Semestre1/prolog.html	0,4
	/~electiva/Semestre1/ProgGenerica.htm	0,4
	/~electiva/Semestre1/ProgEntera.htm	0,4
	/~electiva/Semestre1/Interop.htm	0,4
	/~electiva/Semestre1/index.html	0,4
	/~electiva/Semestre1/GestionSistemasInfo.htm	0,4
	/~electiva/Semestre1/CompGrafica.htm	0,4
	/~electiva/Semestre1/index.html	0,39
	/~electiva/Semestre1/GestCalidad.html	0,39
	/~electiva/Semestre1/ProgFunc.htm	0,32
/~electiva/Semestre1/CompGrafica.htm	/~electiva/Semestre1/index.html	0,57
	/~electiva/Semestre1/GestCalidad.html	0,57
	/~electiva/Semestre1/SistInfoGeografica2.html	0,49
	/~electiva/Semestre1/prolog.html	0,49
	/~electiva/Semestre1/ProgGenerica.htm	0,49
	/~electiva/Semestre1/ProgEntera.htm	0,49
	/~electiva/Semestre1/Interop.htm	0,49
	/~electiva/Semestre1/ProgFunc.htm	0,39
	/~electiva/Semestre1/ProgEntera.htm	0,38
	/~electiva/Semestre1/GestionSistemasInfo.htm	0,38
/~electiva/Semestre1/GestionSistemasInfo.htm	/~electiva/Semestre1/index.html	0,56
	/~electiva/Semestre1/GestCalidad.html	0,56
	/~electiva/Semestre1/SistInfoGeografica2.html	0,55
	/~electiva/Semestre1/prolog.html	0,55
	/~electiva/Semestre1/ProgGenerica.htm	0,55
	/~electiva/Semestre1/ProgEntera.htm	0,55
	/~electiva/Semestre1/Interop.htm	0,55
	/~electiva/Semestre1/ProgFunc.htm	0,44
/~electiva/Semestre1/index.html	/~electiva/Semestre1/GestCalidad.html	0,64
	/~electiva/Semestre1/SistInfoGeografica2.html	0,62
	/~electiva/Semestre1/prolog.html	0,62
	/~electiva/Semestre1/ProgGenerica.htm	0,62
	/~electiva/Semestre1/ProgEntera.htm	0,62
	/~electiva/Semestre1/Interop.htm	0,62
	/~electiva/Semestre1/ProgFunc.htm	0,49
/~electiva/Semestre1/Interop.htm	/~electiva/Semestre1/GestCalidad.html	0,69
	/~electiva/Semestre1/SistInfoGeografica2.html	0,68
	/~electiva/Semestre1/prolog.html	0,68
	/~electiva/Semestre1/ProgGenerica.htm	0,68
	/~electiva/Semestre1/ProgEntera.htm	0,68
	/~electiva/Semestre1/ProgFunc.htm	0,54
/~electiva/Semestre1/ProgEntera.htm	/~electiva/Semestre1/GestCalidad.html	0,79
	/~electiva/Semestre1/SistInfoGeografica2.html	0,74
	/~electiva/Semestre1/prolog.html	0,74
	/~electiva/Semestre1/ProgGenerica.htm	0,74
	/~electiva/Semestre1/ProgFunc.htm	0,74
	/~electiva/Semestre2/index.html	0,59

Los resultados obtenidos para la sesión utilizada, permiten observar que al acceder a la página de las materias electivas de la Facultad de Ingeniería (/~electiva/), se obtiene una recomendación para visitar la página que contiene el índice de todas las electivas del primer semestre, lo cuál es perfectamente lógico si consideramos que el archivo de log utilizado para obtener la información cubre el mes de febrero. Asimismo, se recomiendan dos páginas que pertenecen a dos electivas diferentes. A medida que se van solicitando nuevas páginas dentro de la misma área de trabajo se van recomendando las diferentes páginas que contienen información de las materias electivas. Los resultados anteriores fueron también observados en todas las sesiones que se utilizaron en la evaluación de resultados. Las recomendaciones con valores mas altos se ubicaron siempre dentro de la misma área de trabajo que las páginas accedidas.

Se observa también que parte de las páginas que se visitan en la sesión extraída del log, se encontraron entre las recomendaciones recibidas. Esto fue también observado en la mayoría de las sesiones utilizadas para la evaluación.

Estos resultados muestran que es posible inferir los intereses de un usuario a partir de su conducta de navegación y los patrones de navegación descubiertos, dentro del sitio, utilizando técnicas de Web Usage Mining. Es posible acercar al usuario a sus objetivos dentro del sitio a través de recomendaciones dinámicas a páginas.

Considerando los resultados obtenidos para la evaluación del proceso de descubrimiento de patrones de navegación y del posterior proceso de recomendación dinámica, se muestra que es posible identificar patrones o conductas de navegación, comunes dentro de un sitio web y utilizar esta información en tiempo real para realizar recomendaciones dinámicas de páginas a visitar. Se logra inferir los intereses de los usuarios a partir de sus conductas de navegación actuales y los patrones descubiertos dentro del sitio. Luego en base a estos intereses se pueden realizar recomendaciones dinámicas que intenten acercar a los usuarios a sus objetivos dentro del sitio, sin requerirle mayores esfuerzos.

7 Conclusiones

El objetivo principal de la herramienta desarrollada es realizar el descubrimiento de información acerca de como interactúan los visitantes con un sito web (conductas de navegación), tendiente a construir sitios web capaces de adaptar dinámicamente su estructura y presentación a los hábitos de los visitantes (sitios web adaptativos automáticos).

Este objetivo fue alcanzado a través de una combinación de técnicas. Primero se utilizan técnicas de minería de datos, mas precisamente técnicas de Web Usage Mining, para el descubrimiento de patrones de navegación. Luego de obtenidos dichos patrones se utiliza una técnica que se basa en ellos para realizar recomendaciones dinámicas de páginas a visitar a los usuarios de un sitio web, a medida que estos se encuentran navegando por el mismo. Dichas recomendaciones pretenden acercar al usuario a sus objetivos, inferidos a partir de su conducta de navegación. La combinación de ambas técnicas permite adaptar dinámicamente la estructura del sitio web a los hábitos de los visitantes, ofreciéndoles nuevos caminos de navegación dentro del sitio.

El análisis de resultados realizado en la sección anterior muestra que a través de la técnica de Web Usage Mining es posible identificar los distintos perfiles de usuarios que acceden a un sitio web, a partir de sus conductas o patrones de navegación. Luego estos perfiles son utilizados, en tiempo real, mientras un usuario se encuentra navegando por el sitio, como base para realizarle recomendaciones. Identificando a cada visitante, con alguno de los perfiles descubiertos, es posible inferir sus intereses y ofrecerle dinámicamente, recomendaciones acerca de diferentes páginas a visitar.

Las posibilidades que ofrecen los sitios web adaptativos, basados en la recomendación dinámica de páginas, son vastas. Pueden ser utilizados en sitios dedicados al comercio electrónico, por ejemplo, como una forma automática de comercio cruzado. Si un cierto número de clientes que presentaron un determinado perfil, compraron un determinado producto, a un usuario que presente el mismo perfil, se le ofrecerá dicho producto, en un momento determinado de su navegación. También la identificación de perfiles permite implementar campañas de publicidad personalizada, por ejemplo, lograr que a los usuarios con determinadas conductas dentro del sitio se les muestren determinados artículos publicitarios a media que navegan o se les envíen automáticamente correo electrónico con determinada información de interés comercial. También en sitios académicos, los sitios adaptativos pueden ser de gran utilidad, ya que permiten acelerar la búsqueda de información en ellos.

Dado que la herramienta fue implementada en lenguaje Java esto la convierte en una herramienta portable y extensible. Por ser Java un lenguaje multiplataforma fue posible independizar la herramienta desarrollada de la plataforma sobre la cual es instalada. Esto hace que pueda ser utilizada de igual forma en ambientes Unix, Windows u otros. Asimismo, Java es un lenguaje orientado a objetos que permitió implementar el diseño modular propuesto en la sección 5.1. Así, la herramienta desarrollada puede ser fácilmente extendida para contener nuevas funcionalidades, que atiendan las necesidades particulares de cada sitio web, si es que esto fuera necesario.

8 Trabajos futuros

Los objetivos propuestos fueron alcanzados satisfactoriamente. De todos formas, durante los procesos de búsqueda y análisis de información se observó que existen muchas y diferentes técnicas para alcanzar el objetivo de adaptabilidad dinámica de un sitio web a partir del estudio de los patrones o conductas de navegación de los usuarios del mismo. La técnica de Web Usage Mining ofrece diferentes posibilidades tanto en la etapa de preprocesamiento de datos como en la etapa de descubrimiento de patrones, que deberían ser estudiadas y evaluadas.

8.1 Preprocesamiento de datos

Como se explicó anteriormente, antes de aplicar las técnicas para el descubrimiento de patrones de navegación es necesario pasar por un proceso de identificación de usuarios y posteriormente de sesiones. Para ello se utilizó la metodología de cookies que hacen llegar un número de identificación al usuario. Este no es el único método existente, en particular en [CMS99] se describen otras técnicas que podrían ser analizadas y puestas en práctica para evaluar su eficacia en relación con el uso de cookies.

8.2 Descubrimiento de patrones

En la sección 2 del presente documento, se desarrollan las diferentes posibilidades que presentan las técnicas de Web Usage Mining para el descubrimiento de patrones de navegación. En la implementación de la herramienta se utilizó la técnica de clustering. Sería interesante continuar la investigación dentro de las otras técnicas de Web Usage Mining, *reglas de asociación* y *patrones secuenciales*, y estudiar las posibilidades que ofrecen para la generación de sitios adaptativos. Ya existen algunos trabajos en esta línea de análisis como ser [MJHS97] que podrían tomarse como base para futuros estudios. También existen propuestas originales acerca de nuevas técnicas de minería como por ejemplo [PHMZ2000]

De igual forma la técnica de clustering presenta diferentes posibilidades a la hora de descubrir patrones de navegación, ya sea en lo referente a la forma de aplicación del algoritmo como al algoritmo utilizado. Sería importante analizar como afecta la aplicación del algoritmo o la selección del algoritmo a aplicar, en la calidad de los resultados obtenidos. En la presente implementación se utiliza el algoritmo Vector Quantization, pero por ejemplo podrían utilizarse redes neuronales para la implementación de un algoritmo de clustering más eficiente [PM96]. Algunos autores han incurrido ya en esta línea de trabajo y sus estudios [PE98, YJGD96, MCS99, NFJK99] podrían ser tomados como base para futuras investigaciones.

9 Apéndices

9.1 Cookies

9.1.1 Funcionamiento

Las cookies son pequeñas porciones de datos (archivos de texto), no son código ejecutable. Se envían desde el servidor web hacia el cliente cuando este accede al mismo realizando una solicitud HTTP. En futuros accesos a este mismo servidor web, el navegador del cliente le devolverá una copia de esta cookie junto con la nueva solicitud. Estos archivos se almacenan en el disco duro del cliente, liberando así al servidor web de una importante sobrecarga. Es el propio cliente el que almacena la información y quien se la devolverá posteriormente al servidor web cuando este la solicite.

9.1.2 Formato de cookies

Cada cookie es una pequeña porción de información, con una fecha de caducidad opcional, contenida en archivo de texto con el siguiente formato [5]:

NOMBRE=VALOR; expires=FECHA; camino=PATH; dominio=DOMAIN_NAME; secure

- NOMBRE=VALOR :
  Nombre es el nombre del dato almacenado y valor representa su valor. Valor, es el dato que el servidor pretende le sea devuelto cuando el navegador realiza una nueva solicitud. El nombre nos permite identificar la cookie y recuperar el dato almacenado.

- Expires=FECHA
  La fecha de caducidad es un parámetro opcional que indica el tiempo que se conserva la cookie. Si no se especifica el valor expires la cookie caduca cuando el usuario sale de la sesión en curso con el navegador. Por consiguiente el navegador conservará y recuperará las cookies solo si su fecha de caducidad aún no ha expirado.

- Dominio=DOMAIN_NAME
  Se trata de un nombre de dominio parcial o completo para el cual será valida la cookie. El navegador devolverá la cookie a todo host que coincida con el nombre de dominio DOMAIN_NAME. Por defecto el atributo contiene la dirección del servidor que envió la cookie y él es el único que recibirá una copia cuando el navegador requiera otro archivo desde el servidor.

- Path=PATH
  El valor PATH se utiliza para especificar el conjunto de URLs en el dominio para las cuales es válida la cookie. Por defecto el valor path se establece al camino del objeto HTTP que emitió la cookie.

- Secure
  Este atributo indica que la cookie solo será trasmitida a través de un canal seguro.

9.2 Formatos para Archivos de Registros de Accesos

9.2.1 NCSA Common Log File Format

El archivo con formato CLF [12,13] contiene una línea por cada archivo enviado al cliente. Cada una de estas líneas se compone de cadenas de caracteres separados por espacios.

Si una cadena no tiene un valor se representará por un guión (-). El significado de cada una de estas cadenas es el siguiente:

Nro. de campo	Nombre	Descripción
Campo1	host	Nombre de dominio del host cliente o su número de dirección IP.
Campo 2	ident	Identidad del usuario. No disponible se verá un guión (-)
Campo 3	authuser	Si el documento solicitado es un documento protegido con password, este campo registra el nombre de usuario usado en la solicitud.
Campo 4	date	Día y hora de la solicitud.
Campo 5	request	Dirección del documento pedido, entre comillas....
Campo 6	status	Número de tres dígitos que indica el resultado de la solicitud. (Ej: 200 – OK)
Campo 7	bytes	Número de bytes del objeto enviado al cliente.

Una entrada en este archivo se verá de la siguiente forma:

200.2.39.240 - - [01/Aug/2000:00:00:00 +0300] "GET /facu_lg.gif HTTP/1.1" 200 1238

9.2.2 NCSA Extended (or Combined) Log File Format

Este formato es solo una extensión del anterior [12,13]. Los primeros siete campos se mantienen y se agregan tres campos más.

Nro. de campo	Nombre	Descripción
Campo 8	referrer	Url desde la cual el cliente realiza la solicitud.
Campo 9	agent	Navegador web y plataforma utilizada por el cliente.
Campo 10	cookie	Cookie recibida por el cliente, en caso de que reciba una.

Una entrada en este archivo se verá de la siguiente forma:

10.0.0.2 - - [11/Apr/2001:20:22:44 -0300] "GET /manual/dns-caveats.html HTTP/1.1" 200 8415 "http://felipe/manual/mod/core.html" "Mozilla/4.0 (compatible; MSIE 5.0; Windows NT; DigExt)" 10.0.0.2.1969744232639214690

9.3 Algoritmo Vector Quatization

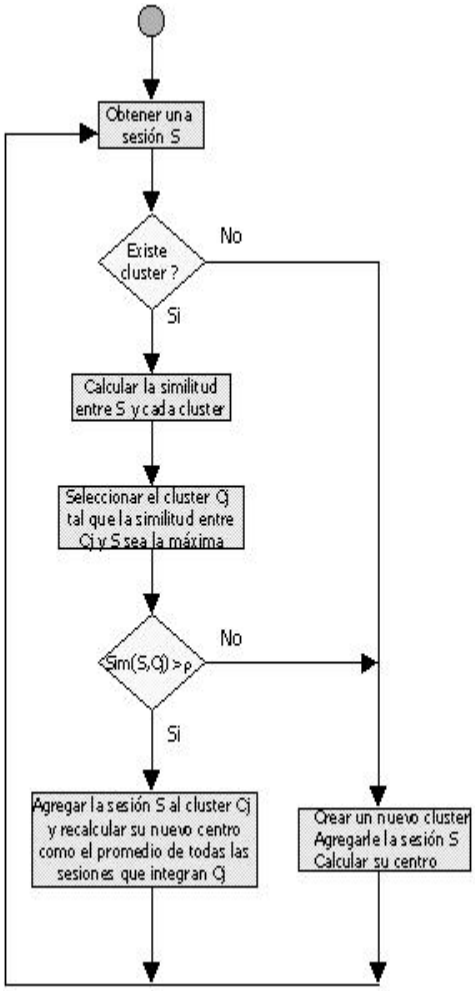
El objetivo de un algoritmo de clustering es dividir un conjunto de elementos en subconjuntos o clusters tal que el grado de asociación sea alto entre miembros de un mismo cluster y bajo entre miembros de diferentes clusters. De esta forma cada cluster determina una clase de elementos a la cual pertenecen los miembros del mismo. Como resultado, los algoritmos de clustering revelan similitudes entre elementos que de otra forma podrían ser imposibles de descubrir.

El algoritmo Vector Quantization, en el contexto de Web Usage Mining, tiene como objetivo separar el conjunto de sesiones de usuario que recibe como entrada, en un número de clusters significativo. Se pretende lograr que las sesiones pertenecientes aun mismo cluster presenten un alto grado de similitud, pero a su vez que el grado de similitud con las sesiones pertenecientes a otros clusters sea bajo.

A diferencia de otros algoritmos de clustering, Vector Quantization no requiere conocer a priori el número de clusters que se deben obtener, sino que se crea nuevos clusters dinámicamente según sea necesario. Para su ejecución, solo se necesita el valor de una constante ρ utilizada para determinar cual es el mínimo grado de similitud entre un cluster y una sesión para que la misma pueda pertenecer a él.

El algoritmo comienza con un conjunto vacío de clusters, la primera sesión examinada dará lugar a la creación del primer cluster. Luego para cada sesión de usuario se determina a que cluster debe pertenecer, en función de su grado de similitud con cada uno de ellos. La nueva sesión pertenecerá al cluster para el cuál haya obtenido mayor grado de similitud siempre y cuando sea mayor que la cota mínima ρ. Si el grado de similitud entre esta nueva sesión y los clusters existentes no supera la cota mínima ρ, se crea un nuevo cluster que solo contiene dicha sesión.

Para decidir la pertenencia o no de una sesión a un cluster se utiliza el concepto de centroide de un cluster. El centroide de un cluster es un vector que define la sesión promedio del cluster y se calcula en base a todas las sesiones pertenecientes al mismo. De esta forma cada sesión se compara con el centroide del cluster según el criterio de similitud definido en la sección 4. En base a este criterio se decide cuál es el cluster con mayor grado de similitud con la sesión considerada. Si el grado de similitud entre el cluster y la sesión es mayor que la cota ρ entonces la sesión es agregada al cluster, de lo contrario se crea un nuevo cluster y se la agrega la sesión considerada. Luego de esto se debe recalcular el centroide o sesión promedio del cluster. Si el cluster tiene una sola sesión esta se convierte en el centroide del cluster. El algoritmo termina cuando todas las sesiones de usuario han sido asignadas a un cluster.



Entrada:

- Sn: Conjunto de sesiones extraídas del log de accesos del servidor
- ρ: Valor mínimo de similitud entre clusters

Salida:

- Cs: Conjunto de clusters de sesiones resultante

Algoritmo VQ:

```
Cs = ϕ
Mientras halla sesiones por procesar en Sn
        Obtener una sesión S
        Si no existe ningún cluster entonces
                Crear un nuevo cluster C
                Agregar S a C
                Calcular el centroide de C
                Agregar C al conjunto de clusters resultante Cs
        Obtener el cluster Cj∈Cs / sim(S,Cj) ≥ sim(S,Ci), ∀Ci∈Cs
        Si sim(S,Cj) ≥ ρ entonces
                Agregar S a Cj
                Recalcular el centroide de Cj
        sino
                Crear un nuevo cluster C
                Agregar S a C
                Calcular el centroide de C
                Agregar C al conjunto de clusters resultante Cs
Fin Mientras
```

9.4 Lenguaje de consulta (DQL)

El lenguaje de consulta (DQL) se utiliza para especificar las solicitudes de recomendación de páginas. Cada una de las sentencias del lenguaje se utiliza para realizar una solicitud diferente. A continuación se describe la sintaxis y la semántica del lenguaje.

9.4.1 Sintaxis del lenguaje

digito =
0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
usuario =
digito | digito usuario
grupo =
digito | digito grupo
posicion =
digito | digito posicion
pagina =
digito | digito pagina

consulta =
SELECT RANK (posicion) FROM ALL FOR THISUSER (usuario) |
SELECT RANK (posicion) FROM URLGROUP (grupo) FOR THISUSER (usuario) |
SELECT RANK (posicion) FROM ALL FOR THISGROUP (grupo) |
SELECT RANK (posicion) FROM URLGROUP (grupo) FOR THISGROUP (grupo) |

SELECT RANDOM FROM ALL FOR THISUSER (usuario) |
SELECT RANDOM FROM URLGROUP (grupo) FOR THISUSER (usuario) |
SELECT RANDOM FROM ALL |
SELECT RANDOM FROM URLGROUP (grupo)

SELECT MOST (posicion) FROM ALL FOR THISUSER (usuario) |
SELECT MOST (posicion) FROM ALL FOR ALLUSERS |
SELECT MOST (posicion) FROM ALL FOR THISGROUP (grupo) |
SELECT MOST (posicion) FROM URLGROUP (grupo) FOR THISUSER (usuario) |
SELECT MOST (posicion) FROM URLGROUP (grupo) FOR ALLUSERS |
SELECT MOST (posicion) FROM URLGROUP (grupo) FOR THISGROUP (grupo) |

VISITED URL (pagina) FOR THISUSER (usuario) |
VISITED URLGROUP (grupo) FOR THISUSER (usuario) |

9.4.2 Semántica del lenguaje

SELECT RANK (n) FROM ALL FOR THISUSER (u)

La consulta requiere que se encuentre la página con el n-ésimo mejor valor de recomendación para la sesión actual del usuario u, tomando en cuenta los patrones de navegación descubiertos dentro del sitio.

SELECT RANK (n) FROM URLGROUP (gp) FOR THISUSER (u)

La consulta requiere que se encuentre la página, perteneciente al grupo de páginas gp, con el n-ésimo mejor valor de recomendación para la sesión actual del usuario u, tomando en cuenta los patrones de navegación descubiertos dentro del sitio.

SELECT RANK (n) FROM ALL FOR THISGROUP (gu)

La consulta requiere que se encuentre la página con el n-ésimo mejor valor de recomendación para la sesión actual del usuario u, tomando en cuenta los patrones de navegación presentados por usuarios pertenecientes al grupo de usuarios gu.

Esta consulta aún no ha sido implementada

SELECT RANK (n) FROM URLGROUP (gp) FOR THISGROUP (gu)

La consulta requiere que se encuentre la página, perteneciente al grupo de páginas gp , con el n-ésimo mejor valor de recomendación para la sesión actual del usuario u, tomando en cuenta los patrones de navegación presentados por usuarios pertenecientes al grupo de usuarios gu.

Esta consulta aún no ha sido implementada

SELECT RANDOM FROM ALL

La consulta requiere que se encuentre cualquier página al azar de entre todas las páginas que actualmente integran el sitio web.

SELECT RANDOM FROM URLGROUP (gp)

La consulta requiere que se encuentre cualquier página al azar de entre todas las páginas que pertenecen al grupo de páginas gp.

SELECT RANDOM FROM ALL FOR THISUSER (u)

La consulta requiere que se seleccione una página al azar, entre todas las páginas que podrían recomendarse al usuario u según su sesión actual y los patrones de navegación descubiertos dentro del sitio.

SELECT RANDOM FROM URLGROUP (gp) FOR THISUSER (u)

La consulta requiere que se seleccione una página al azar, perteneciente al grupo de páginas gp, entre todas las páginas que podrían recomendarse al usuario u, según su sesión actual y los patrones de navegación descubiertos dentro del sitio.

SELECT MOST (n) FROM ALL FOR THISUSER (u)

La consulta requiere que se encuentre la n-ésima página más visitada, por el usuario u.

SELECT MOST (n) FROM ALL FOR ALLUSERS

La consulta requiere que se encuentre la n-ésima página más visitada, por todos los usuarios del sitio.

SELECT MOST (n) FROM ALL FOR THISGROUP (gu)

La consulta requiere que se encuentre la n-ésima página más visitada, por los usuarios pertenecientes al grupo gu.

SELECT MOST (n) FROM URLGROUP (gp) FOR THISUSER (u)

La consulta requiere que se encuentre la n-ésima página, perteneciente al grupo de páginas gp, más visitada, por el usuario u.

SELECT MOST (n) FROM URLGROUP (gp) FOR ALLUSERS

La consulta requiere que se encuentre la n-ésima página, perteneciente al grupo de páginas gp, más visitada por todos los usuarios del sitio.

SELECT MOST (n) FROM URLGROUP (gp) FOR THISGROUP (u)

La consulta requiere que se encuentre la n-ésima página, perteneciente el grupo de páginas gp, más visitada, por los usuarios pertenecientes al grupo gu.

VISITED URL (p) FOR THISUSER (u)

La consulta requiere determinar si el usuario u visito alguna vez la página p durante alguna sesión que haya mantenido en el sitio web.

VISITED URLGROUP (gp) FOR THISUSER (u)

La consulta requiere determinar si el usuario u visitó alguna vez cualquier página del grupo de páginas gp durante alguna sesión que haya mantenido en el sitio web.

9.5 Resultados Obtenidos

En esta sección se muestran parte de los resultados obtenidos en la etapa de evaluación de resultados.

9.5.1 Clusters

A continuación se presenta una tabla conteniendo parte de los clusters obtenidos al ejecutar el algoritmo de clustering, "Vector Quantization", para un valor de $\rho = 0{,}4$.

Cluster	URL	Peso
13	/sysadmin/	1
13	/sysadmin/home.html	1
13	/sysadmin/header.html	1
13	/sysadmin/instructivos/index.html	0,866667
13	/sysadmin/instructivos/ssh/index.html	0,533333
13	/sysadmin/instructivos/news/index.html	0,466667
13	/sysadmin/instructivos/correo/index.html	0,4
13	/sysadmin/instructivos/ssh/inst_ssh.html	0,333333
13	/sysadmin/mapa.html	0,333333
13	/sysadmin/FAQs/index.html	0,333333

Cluster	URL	Peso
17	/~electiva/Semestre1/index.html	1
17	/~electiva/	1
17	/~electiva/Semestre1/CompGrafica.htm	1
17	/~electiva/Semestre1/ProgGenerica.htm	0,888889
17	/~electiva/Semestre1/Interop.htm	0,777778
17	/~electiva/Semestre1/SistInfoGeografica2.html	0,777778
17	/~electiva/Semestre1/GestionSistemasInfo.htm	0,777778
17	/~electiva/Semestre1/GestCalidad.html	0,777778
17	/~electiva/Semestre1/admin.htm	0,777778
17	/~electiva/Semestre1/ProgEntera.htm	0,777778
17	/~electiva/Semestre1/prolog.html	0,555556

17	/~electiva/Semestre1/ProgFunc.htm	0,444444
17	/inco/	0,333333
17	/~electiva/Semestre2/index.html	0,333333

Cluster	URL	Peso
21	/~prog2/curso2000/teorico/indice.htm	1
21	/~prog2/	1
21	/~prog2/curso2000/teorico/cap1/cap1.htm	0,666667
21	/~prog2/curso2000/teorico/cap10/cap10.htm	0,444444
21	/~prog2/curso2000/teorico/cap9/cap9.htm	0,444444
21	/~prog2/curso2000/teorico/cap5/cap5.htm	0,444444
21	/~prog2/curso2000/teorico/cap2/cap2.htm	0,444444
21	/inco/Spanish/ensenanza.html	0,444444
21	/inco/	0,444444
21	/~prog2/curso2000/teorico/cap4/cap4.htm	0,333333
21	/~prog2/curso2000/textos.html	0,333333
21	/~prog2/curso2000/teorico/cap6/cap6.htm	0,333333
21	/~prog2/curso2000/teorico/cap7/cap7.htm	0,333333
21	/~prog2/curso2000/teorico/cap8/cap8.htm	0,333333
21	/~prog2/curso2000/teorico/cap3/cap3.htm	0,333333

Cluster	URL	Peso
41	/iq/alimentos/indice.htm	1
41	/iq/alimentos/presentacion.htm	0,983051
41	/iq/alimentos/bienvenido.htm	0,966102
41	/iq/alimentos/	0,788136
41	/iq/alimento.htm	0,610169
41	/iq/alimentos/cursos.htm	0,601695
41	/iq/iq.htm	0,533898
41	/iq/alimentos/ftfd.htm	0,440678
41	/iq/alimentos/ftfd_presentacion.htm	0,440678
41	/iq/alimentos/ftfd_panel1.htm	0,432203
41	/iq/alimentos/ftfd_indice.htm	0,423729
41	/fing/institutos.html	0,389831
41	/iq/alimentos/alimentos.htm	0,330508

Cluster	URL	Peso
55	/imfia/menu.html	1
55	/imfia/present/indice.htm	1
55	/imfia/present/contenido.htm	0,981481
55	/imfia/present/present.htm	0,981481
55	/imfia/imfia.html	0,907407
55	/imfia/present/default.htm	0,888889
55	/imfia/gruptrab/indice.htm	0,37037
55	/imfia/proyin/indice.htm	0,351852
55	/imfia/publica/publica.htm	0,314815
55	/imfia/publica/indice.htm	0,314815
55	/imfia/publica/contenido.htm	0,314815

Cluster	URL	Peso
65	/imfia/imfia.html	1

65	/imfia/menu.html	1
65	/imfia/enseña/enseña.htm	1
65	/imfia/enseña/contenido.htm	1
65	/imfia/enseña/indice.htm	0,9375
65	/imfia/enseña/default.htm	0,9375
65	/imfia/publica/contenido.htm	0,5625
65	/imfia/publica/indice.htm	0,5625
65	/imfia/publica/publica.htm	0,5625
65	/imfia/enseña/grado.htm	0,4375
65	/imfia/present/indice.htm	0,3125
65	/imfia/proyin/indice.htm	0,3125
65	/imfia/gruptrab/indice.htm	0,3125

Cluster	URL	Peso
66	/~csi/Cursos/cursos_preg/avisos.html	0,833333
66	/~csi/Cursos/cursos_preg/teorico/index.html	0,8125
66	/~csi/Cursos/cursos_preg/practicos/index.html	0,729167
66	/~csi/Cursos/cursos_preg/bdatos.html	0,583333
66	/~csi/Cursos/cursos_preg/parciales/index.html	0,416667
66	/~csi/Cursos/cursos_preg/laboratorio/index.html	0,416667
66	/~csi/Cursos/cursos_preg/examenes/index.html	0,395833

9.5.2 Recomendaciones

A continuación se presentan diferentes tablas que muestran cuatro de las sesiones de usuario utilizadas para la evaluación de resultados y las recomendaciones obtenidas, con su correspondiente valor de recomendación. El valor de recomendación mínimo a partir del cuál se ofrecen recomendaciones es 0.3.

- *Primera sesión*

Sesión de usuario paso a paso	Página recomendada	Valor
/iq/alimentos/alimentos.htm	/iq/alimentos/presentacion.htm	0,42
	/iq/alimentos/indice.htm	0,42
	/iq/alimentos/bienvenido.htm	0,42
/iq/alimentos/bienvenido.htm	/iq/alimentos/presentacion.htm	0,65
	/iq/alimentos/indice.htm	0,65
/iq/alimentos/cursos.htm	/iq/alimentos/presentacion.htm	0,53
	/iq/alimentos/indice.htm	0,53
	/iq/alimentos/	0,32
/iq/alimentos/indice.htm	/iq/alimentos/presentacion.htm	0,7
	/iq/alimentos/	0,45
	/iq/alimento.htm	0,36
/iq/alimentos/presentacion.htm	/iq/alimentos/	0,56
	/iq/alimento.htm	0,44
	/iq/alimentos/publicaciones.htm	0,39
/iq/alimentos/publicaciones.htm	/iq/alimentos/	0,51
	/iq/alimento.htm	0,4
	/iq/iq.htm	0,35
	/iq/alimentos/investigacion.htm	0,33
/iq/alimentos/investigacion.htm	/iq/alimentos/	0,47
	/iq/alimento.htm	0,35
	/iq/iq.htm	0,33

- *Segunda sesión*

Sesión de usuario paso a paso	Página recomendada	Valor
/~electiva/	/~electiva/Semestre1/index.html	0,42
	/~electiva/Semestre1/GestCalidad.html	0,42
	/~electiva/Semestre1/CompGrafica.htm	0,42
/~electiva/Semestre1/admin.htm	/~electiva/Semestre1/SistInfoGeografica2.html	0,4
	/~electiva/Semestre1/prolog.html	0,4
	/~electiva/Semestre1/ProgGenerica.htm	0,4
	/~electiva/Semestre1/ProgEntera.htm	0,4
	/~electiva/Semestre1/Interop.htm	0,4
	/~electiva/Semestre1/index.html	0,4
	/~electiva/Semestre1/GestionSistemasInfo.htm	0,4
	/~electiva/Semestre1/CompGrafica.htm	0,4
	/~electiva/Semestre1/index.html	0,39
	/~electiva/Semestre1/GestCalidad.html	0,39
	/~electiva/Semestre1/ProgFunc.htm	0,32
/~electiva/Semestre1/CompGrafica.htm	/~electiva/Semestre1/index.html	0,57
	/~electiva/Semestre1/GestCalidad.html	0,57
	/~electiva/Semestre1/SistInfoGeografica2.html	0,49
	/~electiva/Semestre1/prolog.html	0,49
	/~electiva/Semestre1/ProgGenerica.htm	0,49
	/~electiva/Semestre1/ProgEntera.htm	0,49
	/~electiva/Semestre1/Interop.htm	0,49
	/~electiva/Semestre1/ProgFunc.htm	0,39
	/~electiva/Semestre1/ProgEntera.htm	0,38
	/~electiva/Semestre1/GestionSistemasInfo.htm	0,38
/~electiva/Semestre1/GestionSistemasInfo.htm	/~electiva/Semestre1/index.html	0,56
	/~electiva/Semestre1/GestCalidad.html	0,56
	/~electiva/Semestre1/SistInfoGeografica2.html	0,55
	/~electiva/Semestre1/prolog.html	0,55
	/~electiva/Semestre1/ProgGenerica.htm	0,55
	/~electiva/Semestre1/ProgEntera.htm	0,55
	/~electiva/Semestre1/Interop.htm	0,55
	/~electiva/Semestre1/ProgFunc.htm	0,44
/~electiva/Semestre1/index.html	/~electiva/Semestre1/GestCalidad.html	0,64
	/~electiva/Semestre1/SistInfoGeografica2.html	0,62
	/~electiva/Semestre1/prolog.html	0,62
	/~electiva/Semestre1/ProgGenerica.htm	0,62
	/~electiva/Semestre1/ProgEntera.htm	0,62
	/~electiva/Semestre1/Interop.htm	0,62
	/~electiva/Semestre1/ProgFunc.htm	0,49
/~electiva/Semestre1/Interop.htm	/~electiva/Semestre1/GestCalidad.html	0,69
	/~electiva/Semestre1/SistInfoGeografica2.html	0,68
	/~electiva/Semestre1/prolog.html	0,68
	/~electiva/Semestre1/ProgGenerica.htm	0,68
	/~electiva/Semestre1/ProgEntera.htm	0,68

	/~electiva/Semestre1/ProgFunc.htm	0,54
/~electiva/Semestre1/ProgEntera.htm	/~electiva/Semestre1/GestCalidad.html	0,79
	/~electiva/Semestre1/SistInfoGeografica2.html	0,74
	/~electiva/Semestre1/prolog.html	0,74
	/~electiva/Semestre1/ProgGenerica.htm	0,74
	/~electiva/Semestre1/ProgFunc.htm	0,74
	/~electiva/Semestre2/index.html	0,59

- *Tercera sesión*

Sesión de usuario paso a paso	Página recomendada	Valor
/imfia/gruptrab/default.htm	/imfia/menu.html	0,53
	/imfia/gruptrab/indice.htm	0,53
	/imfia/imfia.html	0,32
/imfia/gruptrab/indice.htm	/imfia/menu.html	0,75
	/imfia/imfia.html	0,46
/imfia/imfia.html	/imfia/menu.html	0,8
	/fing/institutos.html	0,31
/imfia/menu.html	/imfia/proyin/indice.htm	0,48
	/imfia/proyin/default.htm	0,48
	/imfia/present/indice.htm	0,46
	/imfia/present/present.htm	0,45
	/imfia/present/contenido.htm	0,45
	/imfia/enseña/enseña.htm	0,43
	/imfia/enseña/contenido.htm	0,43
	/imfia/present/default.htm	0,4
	/imfia/enseña/indice.htm	0,4
	/imfia/enseña/default.htm	0,4
	/fing/institutos.html	0,37
/imfia/present/contenido.htm	/imfia/present/indice.htm	0,59
	/imfia/present/present.htm	0,57
	/imfia/present/default.htm	0,52
	/imfia/proyin/indice.htm	0,43
	/imfia/proyin/default.htm	0,43
	/imfia/enseña/enseña.htm	0,38
	/imfia/enseña/contenido.htm	0,38
	/imfia/enseña/indice.htm	0,36
	/imfia/enseña/default.htm	0,36
	/fing/institutos.html	0,33
/imfia/present/default.htm	/imfia/present/indice.htm	0,68
	/imfia/present/present.htm	0,67
	/imfia/proyin/indice.htm	0,39
	/imfia/enseña/enseña.htm	0,35
	/imfia/enseña/contenido.htm	0,35
	/imfia/enseña/indice.htm	0,33
	/imfia/enseña/default.htm	0,33

	/fing/institutos.html	0,3
/imfia/present/indice.htm	/imfia/present/present.htm	0,77
	/imfia/proyin/indice.htm	0,37
	/imfia/proyin/default.htm	0,37
	/imfia/enseña/enseña.htm	0,36
	/imfia/enseña/contenido.htm	0,36
	/imfia/enseña/indice.htm	0,34
	/imfia/enseña/default.htm	0,34
/imfia/present/present.htm	/imfia/proyin/indice.htm	0,37
	/imfia/proyin/default.htm	0,37
	/imfia/enseña/enseña.htm	0,36
	/imfia/enseña/contenido.htm	0,36
	/imfia/enseña/indice.htm	0,34
	/imfia/enseña/default.htm	0,34
	/imfia/proyin/indice.htm	0,33

- *Cuarta sesión*

Sesión de usuario paso a paso	Página recomendada	Valor
/sysadmin/	/sysadmin/home.html	0,45
	/sysadmin/header.html	0,45
	/sysadmin/instructivos/index.html	0,39
/sysadmin/FAQs/index.html	/sysadmin/home.html	0,43
	/sysadmin/header.html	0,43
	/sysadmin/instructivos/index.html	0,37
/sysadmin/header.html	/sysadmin/home.html	0,61
	/sysadmin/instructivos/index.html	0,53
	/sysadmin/instructivos/ssh/index.html	0,32
/sysadmin/home.html	/sysadmin/instructivos/index.html	0,66
	/sysadmin/instructivos/ssh/index.html	0,4
	/sysadmin/instructivos/news/index.html	0,35
	/sysadmin/instructivos/correo/index.html	0,3
/sysadmin/instructivos/news/index.html	/sysadmin/instructivos/index.html	0,67
	/sysadmin/instructivos/ssh/index.html	0,41
	/sysadmin/instructivos/correo/index.html	0,31
/sysadmin/instructivos/ssh/index.html	/sysadmin/instructivos/index.html	0,7
	/sysadmin/instructivos/correo/index.html	0,32
/sysadmin/mapa.html	/sysadmin/instructivos/index.html	0,7
	/sysadmin/instructivos/correo/index.html	0,32

10 Referencias

[1] Analog. http://www.statslab.cam.ac.uk/~sret1/analog/

[2] Wusage7. http://www.boutell.com/wusage/

[3] Openwebscope. http://www.openwebscope.com/

[4] Webalizer. http://www.mrunix.net/webalizer/

[5] Persistent Client State HTTP Cookies.
http://www.netscape.com/newsref/std/cookie_spec.html

[6] Todo sobre las cookies. http://www.iec.csic.es/criptomicon/cookies/

[7] Internet cookies. http://www.ciac.org/ciac/bulletins/i-034.shtml

[8] World wide web committee web usage characterization activity. http://www.w3.org/WCA

[9] GetStats. http://www.eit.com/software/getstats/getstats.html

[10] WWWStats. http://www.ics.uci.edu/pub/websoft/wwwstat/

[11] Tools and Utilities. http://www.nisto.com/mac/tool/logs.html

[12] http://archive.ncsa.uiuc.edu/edu/trg/webstats/

[13] http://www.bacuslabs.com/WsvlCLF.html

[CMS99] Cooley, R., Mobasher, B., y Srivastava, J., Data preparation for mining World Wide Web browsing patterns. En *Journal of Knowledge and Information Systems*, (1) 1, 1999.

[KMM97] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., y Riedl, J. GroupLens: applying collaborative filtering to usenet news. *Communications of the ACM*, Agosto 1999

[NFJK99] Nasraoui, O., Frigui, H., Joshi, A., y Krishnapuram, R., Mining Web access logs using relational competitive fuzzy clustering. En *Proceedings of the Eight International Fuzzy Systems Association World Congress*, Agosto 1999.

[Microsoft2001] Microsoft, Diccionario de Informática e Internet. McGraw-Hill Interamericana, 2001

[MCS99] Mobasher, B., Cooley, C., y Srivastaba, J. Creating adaptative web sites through usage-based clustering of urls. En *IEEE Knowledge and Data Engineering Workshop* (KDEX'99), 1999

[MCS2000] Mobasher, B., Cooley, C., y Srivastaba, J. Automatic Personalization Based On Web Usage Mining. *Communication of ACM*, (43) 2, Agosto, 2000

[MJHS97] Mobasher, B., Jain, N., Han, E-H., y Srivastava J. Web Mining: Pattern Discovery from World Wide Web Transactions. En *Proceedings of the ninth IEEE International Conference on Tools with AI* (IACTAI, 97), 1997

[PM96] Pandya A. y Macy R. Pattern Recognition with Neural Networks in C++. *CRC Press*, 1996

[PE97] Perkowitz, M. y Etzioni, O., Adaptive web sites: an AI challenge. En *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997

[PE98] Perkowitz, M. y Etzioni, O., Adaptive Web sites: Automatically synthesizing Web pages. En *Proceedings of Fifteenth National Conference on Artificial Intelligence*, Madison, WI, 1998

[PHMZ2000] Pei, J., Han, J., Mortazavi-asl, B. y Zhu, H. Mining Access Patterns Efficiently from Web Log. *PAKDD 2000*, 396-407

[SPF99] Spliopoulou, M., Pohle, C., y Faulstich, L. C., Improving the effectiveness of a Web Site with Web usage mining. En *Workshop on Web Usage Analysis and User Profiling* (WebKKD99), San Diego, Agosto 1999

[SCDT2000] Srivastava, J., Cooley, R., Deshpande, M., y Tan, P-T. Web Usage Mining: Discovery and Apliccations of Usage Patterns from Web Data. *SIGKDD Explorations*, (1) 2, 2000.

[SZAS97] Shahabi, C., Zarkesh, A. M., Adibi, J. y Shah, V. Knowledge discovery from users Web-page navigation. En *Proceedings of Workshop on Research Issues Data Engineering*, Birmingham. Inglatera, 1997.

[YJGD96] Yan, T., Jacobsen, M., Garcia-Molina, H., Dayal, U., From user access patterns to dynamic hypertext linking. En *Proceedings of the 5th International World Wide Web Conference*, Paris, Francia, 1996.