PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

"Web Usage Mining aplicado al estudio del comportamiento de los usuarios en el Sistema de Biblioteca de la PUCV"

ALAN FERNANDO CUEVAS PALMA

Profesor Guía: Pamela Hermosilla Monckton

INFORME FINAL DEL PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

Noviembre, 2010

Este trabajo está dedicado a toda mi familia por brindarme siempre su apoyo incondicional y por estar conmigo en los momentos más difíciles. Agradecer a mis amigos por estar siempre a mi lado brindándome su apoyo y confianza y agradecer a todos los que me apoyaron en la realización de éste trabajo.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Resumen

En el presente informe se presenta el marco general del proyecto, el marco teórico y la aplicación de Web Usage Mining en el estudio del comportamiento de los usuarios en el Sistema de Biblioteca PUCV. Para realizar este tipo de estudios es necesario conocer la forma de gestionar el conocimiento en la web y a partir de esto evaluar las técnicas y herramientas que sirven para apoyar este tipo de estudios y así generar conocimientos a partir de información que se encuentre disponible. En este caso especifico de estudio, se utilizó la información contenida en los logs del servidor donde se encuentra alojado el Sistema de Biblioteca, y se utilizó un Data Warehouse con esquema tipo estrella para mantener la información contenida en ellos de forma estructurada y así facilitar el análisis sobre estos datos mediante la generación de cubos OLAP. A partir del proceso ETL que se llevó a cabo, se logró generar estadísticas generales sobre los accesos al sitio durante el periodo de tiempo considerado en el estudio y posteriormente, se realizó un análisis sobre las sesiones de los usuarios y secuencias de *clickstreams* en donde se pudo obtener resultados interesantes sobre el comportamiento de los usuarios en el sistema. Así, se pudo generar por ejemplo, puntos comunes de salida, duraciones de navegación, cantidad de páginas visitadas por los usuarios, etc. características que son de gran relevancia para las personas encargadas de administrar el sitio. Luego, se utilizó el algoritmo A priori, el cual genera reglas de asociación basadas en una medida de soporte y confianza que permitió observar como acceden los usuarios a las distintas secciones y así establecer posibles formas de reestructuraciones al sistema. Finalmente, se validaron los resultados obtenidos con las personas encargadas y se pudo comprobar que el objetivo principal del sistema se cumple en gran medida.

Palabras clave: Web Mining, Web Usage Mining, Data Mining, Descubrimiento de Patrones, Sesiones de Usuarios, Data Warehouse, Web Logs, Web.

Abstract

This work presents the general context of the project, the theoretical framework and the application of Web Usage Mining in the study of user behavior in PUCV Library System. To perform this kind of study is basic to understand how is managed the knowledge on the web and then evaluate the techniques and tools used to support this kind of study and generate knowledge from the available information. In this specific case of study was used the information contained in the logs' server where is hosted the Library System and, to store the logs information a Data Warehouse with star scheme was used to keep the information structured and thus facilitate the analysis using OLAP Cubes. From the ETL process was performed was possible to generate general statistics about access to the site during the time period considered in the study and subsequently carried out an analysis of user sessions and clickstreams sequences where it could obtain interesting results about user behaviors on the system. Thus, it could be generated for example, common points of departure, navigation durations, number of viewed pages by users, etc. characteristics that are highly relevant to people responsible for managing the site. Then was used the A priori algorithm, which generates association rules based on a measure of support and confidence, that allow to see how users access to various sections and then establish possible ways of restructuring the system. Finally, the results was validated with people in charge of manage the Library System and it was found that the main objective of the system is largely achieved.

Keywords: Web Mining, Web Usage Mining, Data Mining, Pattern Discovery, Users Sessions, Data Warehouse, Web Logs, Web

Glosario de Términos

Business Intelligence: se llama así al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa.

Clickstream: es una secuencia agregada de vistas de páginas ejecutadas por un usuario en particular en la navegación de un sitio.

Cookie: es un fragmento de información que se almacena en el disco duro del visitante de una página web a través de su navegador, a petición del servidor de la página, como por ejemplo, nombres de usuarios y contraseñas.

Data Mining: consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Data Warehouse: es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza

Log: un registro oficial de eventos durante un periodo de tiempo en particular. En este caso es un archivo que informa sobre las conexiones a un servidor.

Online analytical processing*:* es una solución utilizada en el campo de Business Intelligence cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

Referrer: se refiere al sitio desde donde los usuarios acceden al sitio que están visitando.

Web Crawler: es un programa que inspecciona las páginas del World Wide Web de forma metódica y automatizada.

Web Master: es la persona responsable de mantención o programación de un sitio web.

Web Mining: es una metodología de recuperación de la información que usa herramientas de la minería de datos para extraer información tanto del contenido de las páginas, de su estructura de relaciones (enlaces) y de los registro de navegación de los usuarios.

Web Usage Mining: es una de las clasificaciones de Web Mining que está orientada principalmente a capturar y modelar los patrones de comportamiento del usuario en la web.

Lista de Abreviaturas

BI:	Business Inteligence
CSS:	Cascading Style Sheets
ETL:	Extract, Transform and Load
HTML:	HyperText Markup Language
HTTP:	Hypertext Transfer Protocol
IP:	Internet Protocol
ISP:	Internet Service Provider
OLAP:	On-Line Analytical Processing
PUCV:	Pontificia Universidad Católica de Valparaíso
SQL:	Structured Query Language
UPM	User Pageview Matrix
WUM:	Web Usage Mining
XML:	Extensible Markup Language

1 Introducción

Con el explosivo crecimiento de las fuentes de información en el World Wide Web y el rápido incremento de los usuarios en la red ha llegado a convertir Internet en una herramienta indispensable en la vida y comunicación de muchas personas. De esta forma, la web se ha convertido en el repositorio público de datos más grande que existe (más de 20 mil millones de páginas estáticas) y hasta el año 2007 hubo cerca de 120 millones de servidores web (56% Apache, 31% Microsoft) [1]. De lo anterior es posible preguntarse ¿Qué es posible realizar ante tal cantidad de información o qué se pude rescatar a partir de ella, etc.?

En base a lo anterior surge "Web Mining" la cual es una extensión de "Data Mining" y que tiene como objetivo descubrir y analizar información relevante involucrando técnicas y acercamientos basados en "Data Mining" orientados al descubrimiento y extracción automática de información de documentos y servicios de la Web teniendo en consideración el comportamiento y preferencias del usuario [2].

En el presente estudio se dará a conocer la aplicación de "*Web Usage Mining"*, una clasificación específica de Web Mining, en el Sistema de Biblioteca de la *Pontificia Universidad Católica de Valparaíso* para realizar un análisis del comportamiento que poseen los usuarios en este sistema y así ofrecer posibles sugerencias de mejoras y de esta forma, lograr facilitar la navegación del usuario por el sistema.

En este informe se presentará el objetivo general y los objetivos específicos de esta investigación respecto al tema a tratar en este proyecto, para posteriormente explicar el plan de trabajo que se llevará a cabo durante el transcurso de la investigación dando a conocer los alcances y limitaciones que tendrá la misma.

Luego, se expondrá la forma de gestionar el conocimiento en la Web y como el proceso KDD (*Knowledge Discovery in Databases*) puede ser aplicado en este tipo de estudios en conjunto con repositorios de datos como los Data Warehouse. También, se realizará un estudio acabado sobre "Web Usage Mining", dando a conocer las etapas del proceso completo del cual consiste, utilizando como base los *weblogs* del sistema en estudio.

Posteriormente, se expondrá el diseño propuesto para realizar este estudio, el cual consta del diseño de un *Data Warehouse,* el cual mantendrá los logs mencionados anteriormente de una forma estructurada para facilitar la creación de los vectores característicos que serán utilizados en el algoritmo *A priori* para extraer patrones de navegación de los usuarios, así como también facilitar el proceso de análisis de los datos mediante técnicas tipo OLAP (*On-Line Analytical Processing*). También se expondrá el análisis realizado a los *clickstreams* existentes en las sesiones y las conclusiones obtenidas a partir de ellas.

Finalmente, se expondrá el análisis realizado junto con sus respectivas validaciones con las personas encargadas de administrar el sistema y se concluirá sobre lo desarrollado en el proyecto.

2 Marco General del Proyecto

2.1 Objetivos

2.1.1 Objetivo General

Aplicar técnicas y herramientas de Web Usage Mining para analizar el comportamiento que poseen los usuarios en el Sistema de Biblioteca PUCV.

2.1.2 Objetivos Específicos

1. Investigar y analizar técnicas, algoritmos y herramientas de Web Usage Mining para identificar y determinar las que se utilizarán en el estudio.
2. Aplicar y analizar los resultados obtenidos a partir del algoritmo, técnicas y herramientas seleccionadas.
3. Validar resultados obtenidos con la(s) persona(s) encargadas de administrar el sistema.
4. Realizar posibles ajustes al modelo.

2.2 Justificación del Proyecto

El iniciar la investigación en este tema, *Web Mining*, se debe principalmente a la relevancia que ha llegado a tener en estos últimos tiempos en la variedad de aplicaciones web que se han desarrollado en las distintas áreas existentes en el World Wide Web. El poder realizar mejoras en ellas, debido a la alta competitividad que existe actualmente, se ha convertido en un aspecto fundamental para lograr mantenerse en el tiempo y cumplir con las expectativas de los usuarios en la red.

De lo anterior, surge la idea de investigar en profundidad los distintos aspectos que considera Web Mining y enfocarse principalmente, en Web Usage Mining, que es una de las clasificaciones con las que cuenta, y que tiene como objetivo principal capturar y modelar los patrones de comportamiento del usuario en la Web.

2.3 Planteamiento del Problema.

Actualmente, el Sistema de Biblioteca de la PUCV no cuenta con un estudio en detalle sobre el comportamiento que tienen los usuarios en el sitio web (en este caso los estudiantes principalmente y docentes), debido a la falta de personal que existe y también a la carencia de conocimientos específicos sobre cómo realizar este tipo de estudios. Es por esto, que se eligió específicamente este sitio web para realizar el estudio y también por la relevancia que posee en la comunidad estudiantil principalmente.

Por consiguiente, el estudio se centralizará esencialmente en capturar patrones de comportamiento de los usuarios en este sitio y así, obtener conocimientos que puedan dar señales a los administradores del sitio para posibles reestructuraciones de él o sobre las visitas de páginas

que realizan los usuarios, como cuales son las más visitadas por ellos o en las que más gastan tiempo, etc. de tal forma de mejorar la navegación de los usuarios por el sitio.

En este sistema de biblioteca se puede apreciar distintas secciones en las que los usuarios pueden acceder; así, de manera expositiva, se nombra a continuación los ítems o links que posee este sitio junto con algunas capturas de imagen del sistema:

- Catálogos
  - Local (OPAC)
  - Local (Revistas)
  - Catálogo de revistas electrónicas
  - Historia de la Iglesia
  - Derecho Romano
  - Fondo Budge
- Recursos
  - Bibliotecas Virtuales
    - Biblioteca Ágora
    - Océano Diccionarios
    - Océano Administración de Empresas
    - Océano Salud
    - Océano Universitas
    - Biblioteca Virtual Miguel de Cervantes
    - Cybertesis
    - Memoria Chilena
    - Biblioteca Virtual del Bicentenario
    - Fuentes para la Historia de Chile
    - BVS – Psicología
  - Bases de Datos
    - Elsevier Scopus
    - ISI Web of Science
    - MathScinet
    - Naxos
    - OCDE
  - Boletines
    - Boletín de Alerta de Información para Profesores
  - Estándares
    - INN
    - NISO
    - ISO
    - IEEE
  - Revistas Texto Completo
    - BioMed Central
    - BEIC
    - Directory of Open Acces Journals
    - Directorio de Revistas Open Acces con Factor de Impacto
    - Catálogo de Revistas Electrónicas

        - Latindex
        - Nature Biotechnology
        - Nature
        - Pubmed
        - Scielo
        - Science Magazine Online
    - Textos Digitales
- Servicios
    - Multibuscador
    - Guía de Servicios
    - Cubículos de Estudio
    - Videoteca
    - Sala de Música
    - Préstamo en Sala de Lectura
    - Préstamo Fuera de Biblioteca
    - Servicios de Reserva
    - Préstamo Interbibliotecario
    - Conmutación Bibliografica
    - Perfiles Inside
    - Servicios Gratuitos y Pagados
- Herramientas
    - Biblioteca Ágora
    - Manual de Estrategias Didácticas
    - Cómo hacer Citas Bibliográficas
    - Publicando Revistas Electrónicas
    - Aprender a Aprender
    - Guía para las Publicaciones Científicas
    - Webmail de Biblioteca
- Proyectos
    - Proyectos Docentes
    - Proyectos Estudiantiles
    - Electronic Journal of Biotechnology
    - Fondo Budge
    - Fondo Margot Loyola Palacios
    - Poseidon
    - ARPA Archivos Patrimoniales de Valparaíso
    - Archivo Histórico PUCV
- E-Bibliotecario
- El Sistema
    - Bienvenida
    - Misión
    - Plan Ágora 2000-2004
    - Plan Nuevo Ágora 2005-2010
    - Nuestras Bibliotecas
    - Directorio

- Organigrama
- Reglamento de Servicios
- Conferencias, charlas, eventos

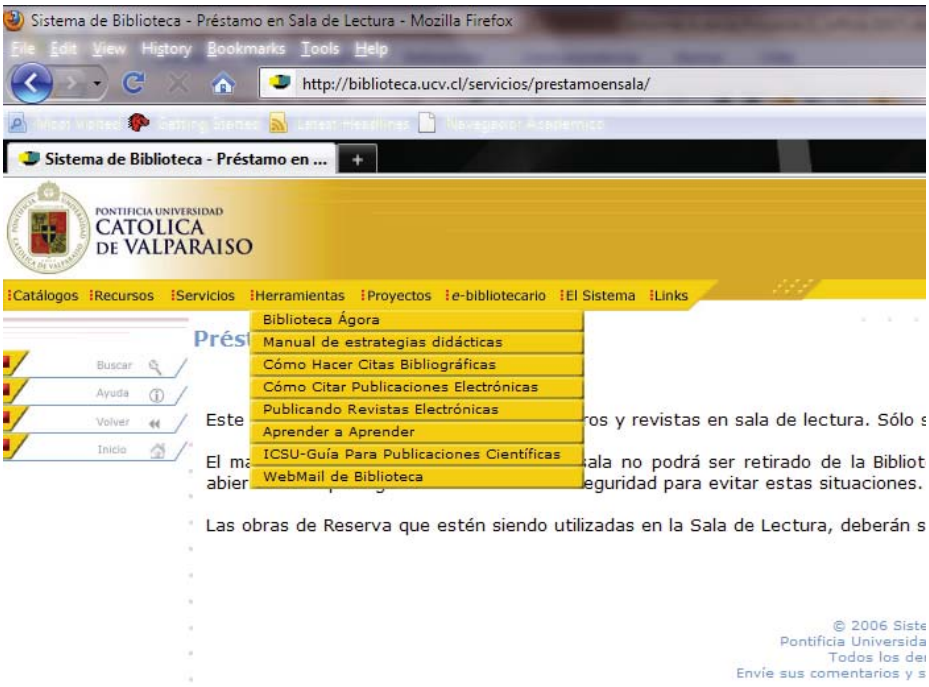


Figura 2.1 Contenido del ítem “Herramientas” del menú principal

Estas son las principales secciones a las que puede acceder el usuario y es por eso que se nombró cada uno, sin embargo, dentro de cada sección existen hipervínculos a otras páginas las cuales también fueron consideradas en el estudio.

2.4 Alcance del Estudio

Considerando que los datos que se pueden extraer desde el servidor donde se encuentra alojado el sitio web son: logs files, cookies, datos de estructura de la información, meta datos y datos de contenido [3], toda ésa información puede ayudar al análisis de un determinado sitio para así generar conocimientos a partir de él y realizar mejoras o incluso tareas de inteligencia de negocios con toda la información analizada.

Sin embargo, un log de un servidor web es una de los recursos fundamentales para realizar los estudios que realiza *Web Usage Mining*, una clasificación de Web Mining que se estudiará en las siguientes secciones, ya que mantiene grabaciones explícitas del comportamiento sobre la navegación de los visitantes del sitio [3]. Es por esto que el análisis de la información se centralizó principalmente en los logs del servidor web donde se encuentra alojado el sitio.

3 Marco Teórico del Proyecto

3.1 Gestión del Conocimiento en la Web

En esta sección se realizará una descripción sobre los procesos que se realizan al momento de gestionar la información presente en la Web para así lograr obtener conocimientos útiles a partir de ella, la cual consta de varias etapas, en donde en cada una de ellas se realiza diversas actividades con el fin de llegar a una estructura apta para realizar un análisis sobre la información y así extraer conocimientos útiles y relevantes en lo que respecta la información Web.

3.1.1 Proceso KDD

El proceso KDD (*Knowledge Discovery in Databases*) ó descubrimiento de conocimiento en bases de datos, en [4] lo definen como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de datos" el cual está implícitamente presente en éstos [5].

Este estudio se realizó en base a la web data (en particular los weblogs) el cual se utilizó el proceso KDD para lograr tomar los logs del servidor web, transformarlos y llevarlos a una estructura que sea más manipulable que en este caso es una base de datos relacional. A continuación se describen las etapas de este proceso:

1. Selección: en esta etapa se realiza la selección y obtención de los datos que se utilizarán en el estudio
2. Pre procesamiento: el formato que poseen los *weblogs* están basados de acuerdo a la estructura que definió el webmaster del sitio, por lo tanto se debe estudiar el contenido que poseen para posteriormente, realizar una limpieza de ellos de tal forma de considerar sólo la información útil para realizar el estudio. Por ejemplo, se debe identificar y eliminar los registros de los *web crawlers*, ya sea mediante el registro **Agent** o bien por la **IP adress** [6]. Además, se debe eliminar todas las peticiones a objetos de las páginas (imágenes, videos, etc.). Una vez que los logs se encuentren limpios, se debe realizar la identificación y construcción de sesiones de usuarios.
3. Transformación: una vez que se encuentren pre procesados y limpios las fuentes de datos (*weblogs*), se procede a reconstruir las sesiones de usuarios agrupando los registros por **IP adress** y **Agent**, utilizando alguna de las dos alternativas de identificación: sesiones de no más de 30 minutos de duración o sesiones sin páginas visitadas más de una vez. [6].
4. Web Mining: todas las etapas que se han nombrado anteriormente se realizan para preparar los datos para ser utilizados como entrada en los algoritmos de Data Mining. Por medio de estos algoritmos, se obtienen patrones de comportamiento de los usuarios y que, mediante una buena interpretación de ellos, dará lugar a la obtención del conocimiento deseado.
5. Evaluación: una vez que se han extraído los patrones de comportamiento de los usuarios, es necesario evaluar con una persona cercana al sitio los resultados que se obtuvieron para lograr identificar las acciones que realizan los usuarios e identificar

tendencias o cambios en las preferencias de los mismos [7]. Para esto, se puede utilizar técnicas de filtrado de información, técnicas de visualización, herramientas de tipo OLAP, y herramientas de minería de datos. Esta evaluación puede dar pie a posibles cambios en las etapas mencionadas anteriormente para así generar conocimientos útiles para la persona experta en el dominio. A pesar que una iteración en el proceso KDD puede dar lugar a no encontrar ninguna información valiosa, esa iteración y sus resultados sirven como base para la siguiente iteración [8].

3.1.2 Data Warehouse y proceso KDD

Mientras que el proceso KDD busca tener los datos procesados y consolidados para extraer patrones mediante Web Mining, los Data Warehouse o Data Webhouse (llamados así también porque almacena *web data* principalmente) ofrecen:

1. Una estructura definida en base a dimensiones que hacen sentido a los tomadores de decisión. Permitiendo la incorporación de herramientas OLAP para responder consultas que se tienen a priori acerca del sitio.
2. Datos limpios y consolidados cuya extracción, transformación y carga es realizada periódicamente, dando lugar a una copia histórica completa de los *web data*, permitiendo la aplicación directa de técnicas de Web Mining.

La Figura 3.1 muestra la relación entre el proceso KDD y los Webhouses para la extracción de conocimiento.

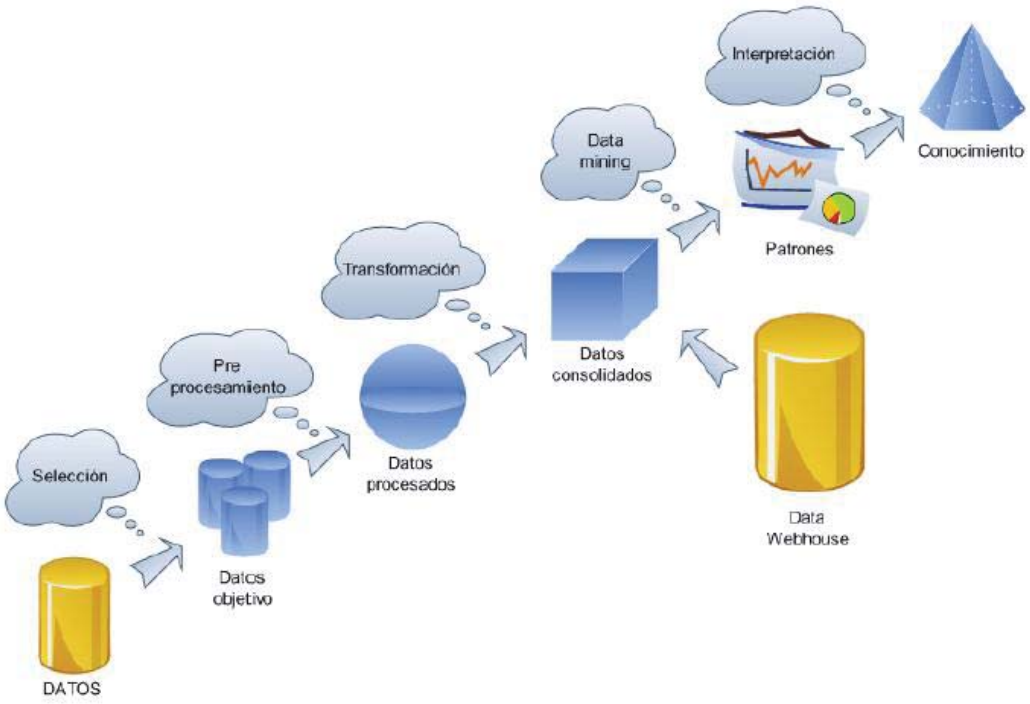


Figura 3.1 Relación entre el proceso KDD y el Data Webhouse [4]

3.1.2.1 **Modelo Estrella**

Tiene una tabla de hechos (o *tabla fact*) que contiene los datos para el análisis, rodeada de las tablas de dimensiones. La relación entre hechos y dimensiones se establece mediante la inclusión de las llaves primarias de las dimensiones como llaves foráneas en la tabla de hechos. De este modo, cada dato de la Tabla Fact está caracterizado por valores en los atributos de las dimensiones a través de estas llaves.

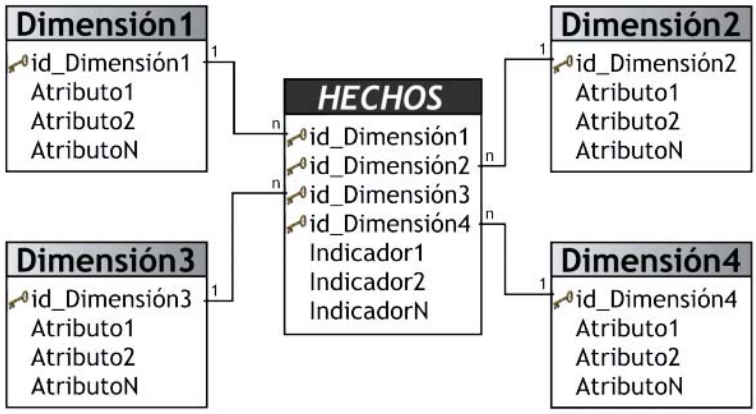


Figura 3.2 Modelo estrella

El modelo estrella es el más simple de interpretar y optimiza los tiempos de respuesta ante las consultas de los usuarios. Este modelo es soportado por casi todas las herramientas de consulta y análisis, y los metadatos son fáciles de documentar y mantener, sin embargo es el más lento para la carga y es el más lento de construir [9].

3.1.2.2 **Modelo Copo de Nieve**

Este modelo representa una extensión del modelo estrella cuando las dimensiones se organizan en jerarquías de dimensiones. Existe una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quiénes a su vez pueden estar relacionadas con una ó más tablas de dimensiones. Este modelo es más cercano a un modelo de entidad relación, que al modelo en estrella, debido a que sus tablas de dimensiones están normalizadas.

Una de los motivos principales de utilizar este tipo de modelo, es la posibilidad de segregar los datos de las dimensiones y proveer un esquema que sustente los requerimientos de diseño. Otra razón es que es muy flexible y puede implementarse después de que se haya desarrollado un esquema en estrella.

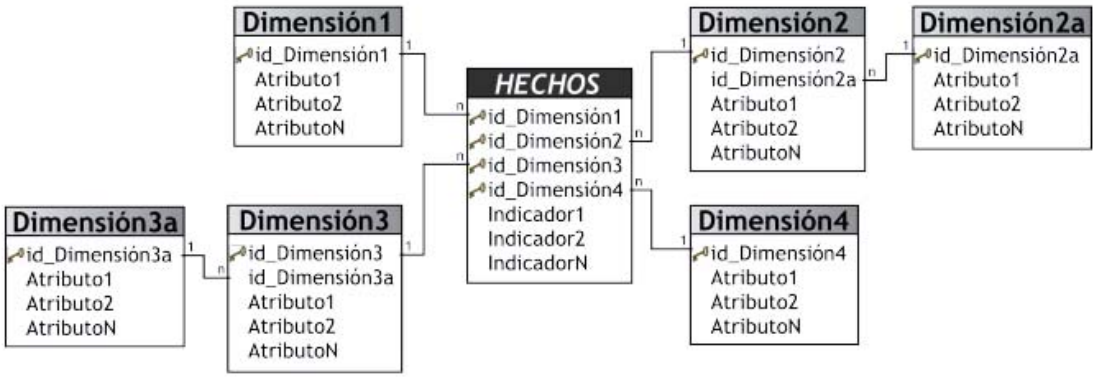


Figura 3.3 Modelo copo de nieve

3.1.2.3 **Modelo Constelación**

Este modelo está compuesto por una serie de esquemas en estrella, y tal como se puede apreciar en la, está formado por una tabla de hechos principal ("HECHOS_A") y por una o más tablas de hechos auxiliares ("HECHOS_B"), las cuales pueden ser resúmenes de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones. No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones.

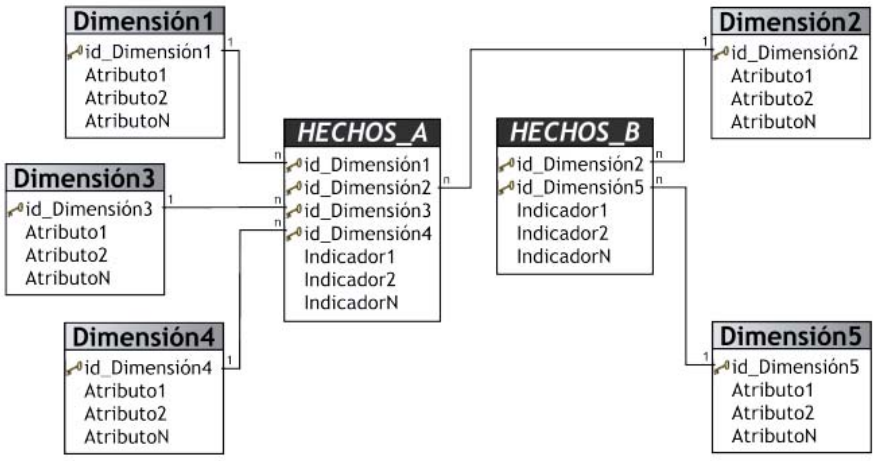


Figura 3.4 Modelo constelación

3.1.3 Representación del Conocimiento

En algunas fuentes proponen mantener un **Repositorio de Patrones** [5] [10] [8] utilizando una arquitectura Data Warehouse en bases de datos relacionales mediante el *modelo estrella* para así almacenar los patrones descubiertos y facilitar el posterior análisis de ellos debido a las ventajas que poseen este tipo de arquitecturas.

En la Figura 3.5 se presenta el modelo de Repositorio de Patrones propuesto por [8], en donde se puede apreciar las distintas dimensiones que considera al momento de almacenar los patrones encontrados. Como se puede observar, el modelo propuesto consta con de una **Tabla Fact**, que da cuenta de los estudios de Web Mining realizados y 4 dimensiones: **Time**, **Text_Preference**, **Browsing_Behavior** y **WMT (Web Mining Technique)**.

La dimensión Tiempo (**Time**), contiene el periodo de tiempo en que se realizó el estudio de Web Mining. La dimensión **Text_Preference** guarda una descripción contextual del estudio y detalles sobre el período de tiempo en que fueron tomados los datos. La dimensión **WMT**, describe la técnica y herramienta de Web Mining utilizadas como clustering utilizando SOFM, K-means, C-means, etc. Por su parte, la dimensión **Browsing_Behavior** contiene los parámetros de navegación encontrados, la **fórmula** de comparación para los vectores de sesiones, el **periodo** de tiempo al cual pertenecen los datos analizados y la **descripción** de los detalles. Por último la **Tabla Fact** contiene las sugerencias de *navegación* inferidas de los patrones descubiertos, *estadísticas* como el porcentaje de visitas en el periodo de estudio y las *keywords* descubiertas.

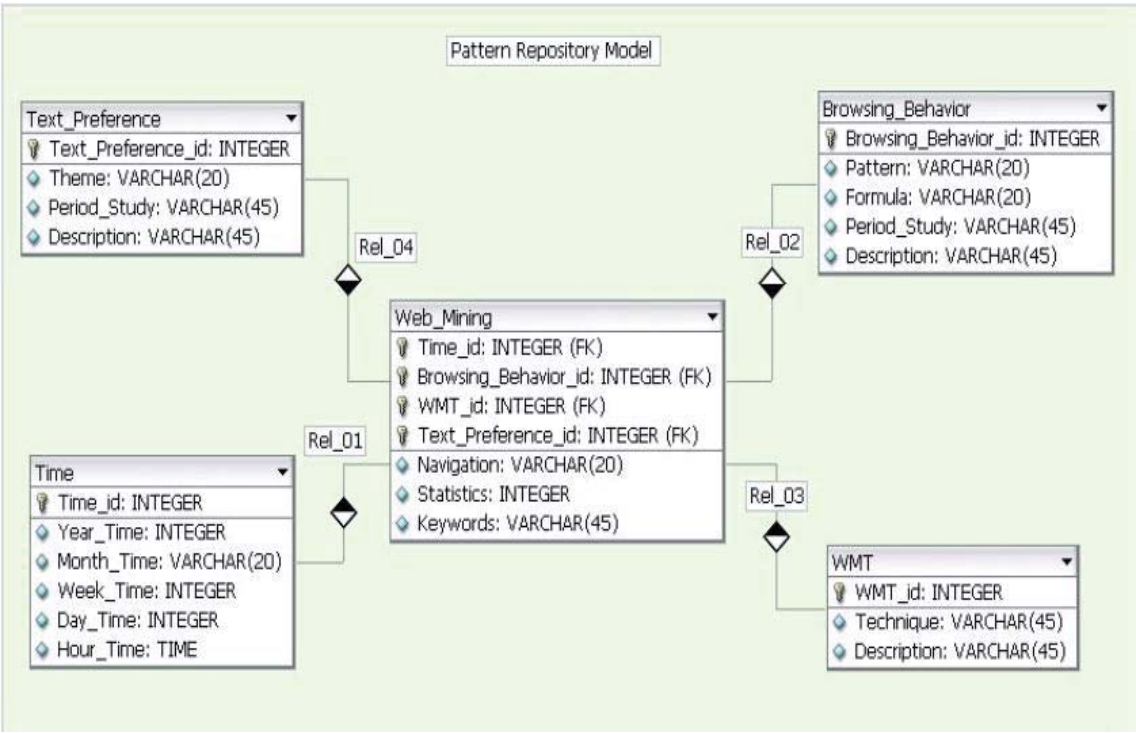


Figura 3.5 Modelo de repositorio de patrones (propuesto por J. Velásquez en [8])

El representar el conocimiento de la forma anteriormente descrita tiene dos aplicaciones: por un lado, proveer al *webmaster* de la orientación necesaria para cambiar el contenido y estructura de un sitio web; y por otro lado, proveer de recomendaciones de navegación online a los usuarios basadas en el conocimiento que se tiene de sus sesiones de navegación [11].

En la literatura se han propuesto distintos modelos multidimensionales enfocados a la construcción de Data Warehouses. Así, en la Figura 3.6 se muestra un modelo estrella genérico para un Data Warehouse propuesto por Juan Velásquez en su libro *Adaptative Websites* [8].

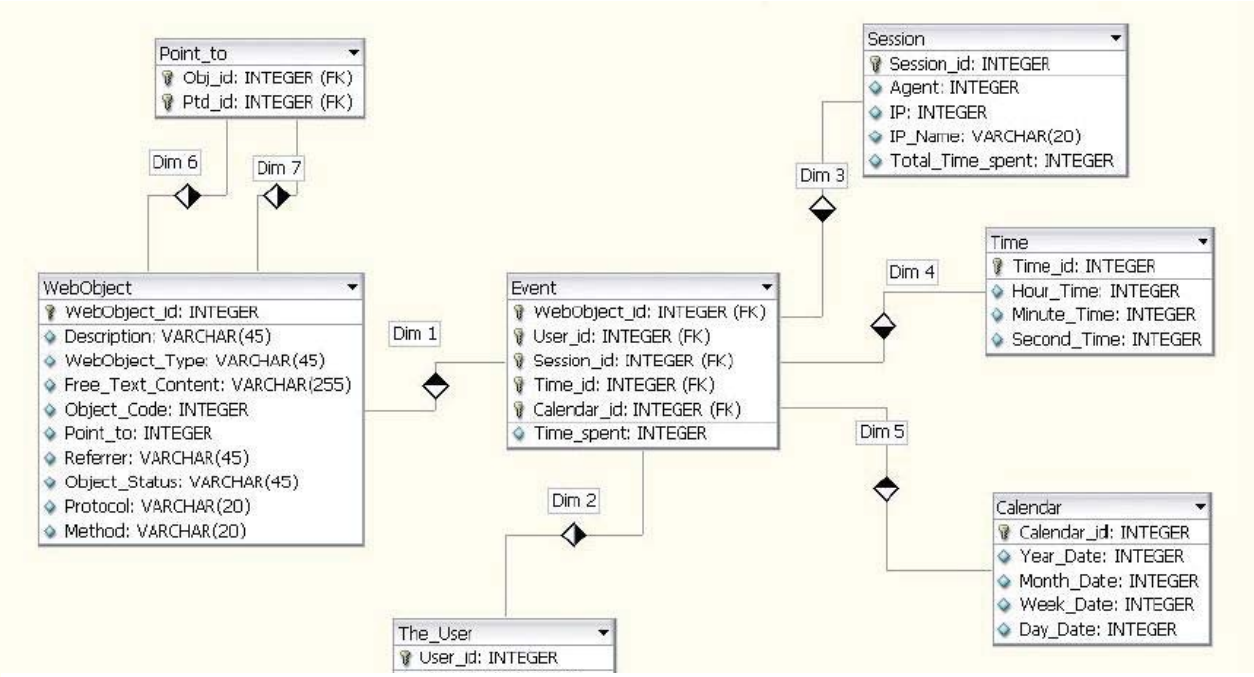


Figura 3.6 Modelo estrella genérico de un Data Warehouse (ó Data Webhouse) (propuesto por J. Velásquez en [8])

La **Tabla Fact** contiene el tiempo gastado en desplegar cada objeto y las dimensiones propuestas incluyen:

1. La dimensión **Session**, que incluye la sesión reconstruida en la cual el objeto fue solicitado.
2. La dimensión Tiempo que fue dividida en **Time** y **Calendar**, de modo de evitar el crecimiento desmedido de la dimensión conjunta, al incluir desde el año hasta el segundo en que fue solicitado un objeto.
3. La dimensión **User** en caso de contar con información acerca del usuario.
4. La dimensión **Web_Object** que caracteriza cada objeto contenido en las páginas y la **Tabla Point_to** que recoge la estructura de hiperlinks del sitio web.

3.1.3.1 **Proceso ETL**

El proceso de extracción, transformación y carga de datos resulta el más costoso en la construcción de un Data Warehouse o Data Webhouse [12]. La razón viene dada por la calidad de los datos, los cuales generalmente vienen con errores de consistencia, formato u omisión.

El proceso ETL, para la construcción de un Data Webhouse o DataWarehouse, tiene directa relación con las etapas del proceso KDD aplicadas a los web data [8]. Así es cómo el ETL debiese abordar la extracción de registros de los web logs y el código HTML de las páginas, eliminar la información irrelevante, transformar los datos en input para los algoritmos de web mining y, de ser necesario, consolidar los formatos de datos.

3.1.3.2 **Análisis Multidimensional**

El modelamiento multidimensional busca diseñar repositorios que respondan los requerimientos de información a partir de cómo los usuarios finales ven el negocio. En otras palabras, la información se almacena en base a dimensiones relevantes para el usuario final, de modo de facilitarles la navegación en el repositorio para que encuentren las respuestas que buscan [12]. Así, relaciona la información de algún fenómeno, como las ventas de una compañía, con los atributos de diversas dimensiones como el Producto, la Ciudad y el Tiempo que hacen sentido al experto del negocio. Esta relación le da sentido a la información almacenada, permitiendo la realización de cruces de información al interior del repositorio, lo que posibilita al usuario a ver más allá de lo que es evidente.

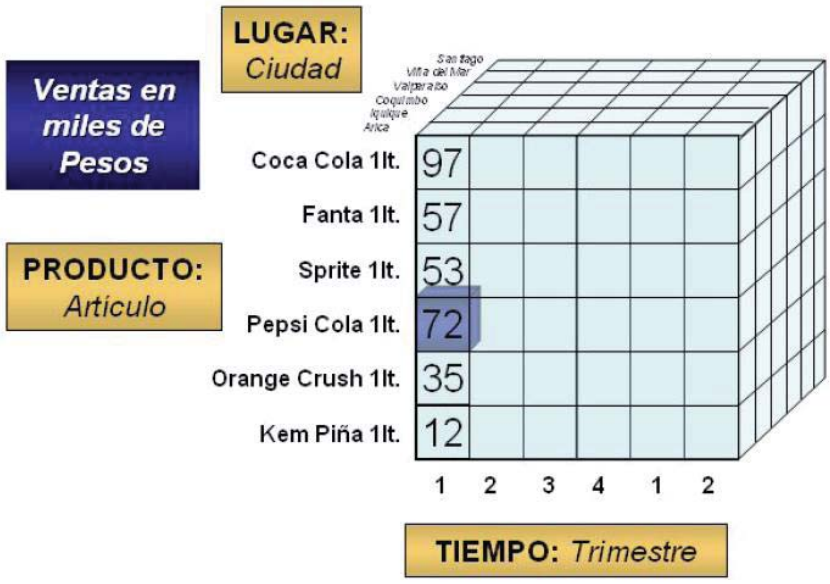


Figura 3.7 Ejemplo de Cubo multidimensional

Así, por ejemplo, el valor 72 mil no dice absolutamente nada, sin embargo, decir que se vendieron 72 mil pesos de Pepsi Cola de 1 litro en Arica durante el primer trimestre del 2006 ofrece todo un contexto a la información que permite tomar alguna decisión.

3.2 Web Mining

En los últimos años el crecimiento y el auge de Internet han aumentado exponencialmente tanto en cantidad de información como de usuarios (personas físicas, empresas, universidades, gobiernos, etc.) debido principalmente a la necesidad de contar con datos para la interrelación del mundo globalizado. De aquí que la web ha llegado a ser el repositorio público de datos más grande que existe, el cual consta con más de 20 mil millones de páginas estáticas [13] y más otra gran cantidad que corresponde a las páginas web que se generan dinámicamente, es decir, aquellas que se generan automáticamente a partir de datos extraídos desde bases de datos.

En base a lo anterior se pueden descubrir una serie de problemas debido al crecimiento exponencial que ha ocurrido en éstos últimos 15 años [13], como por ejemplo el lograr encontrar información relevante, debido principalmente a la baja precisión y escasa cobertura que poseen los motores de búsqueda. La escases de cobertura se debe a que no todos los motores de búsqueda tienen la suficiente capacidad de indexar la web, debido a varios factores; el ancho de banda, el espacio de disco duro, el costo económico, etc.

A partir de lo anterior, surge la inquietud de cómo poder obtener beneficios a partir de esa enorme cantidad de información para así ayudar en la toma de decisiones, búsqueda de información específica, etc. Bueno, para lo anterior surge la Minería Web ó Web Mining (traducida al inglés) que es una extensión de la Minería de Datos y que busca dar solución a los problemas anteriormente nombrados y obtener beneficios de la enorme cantidad de información que se encuentra en Internet.

3.2.1 ¿Qué es Web Mining?

El término de “Web Mining” fue acuñado por O.Etzioni en el año 1996.

Según la referencia [13] Web Mining se define como el descubrimiento de patrones potencialmente útiles y el análisis de información implícita de los artefactos o actividades relacionadas con la Web. Otra definición un poco más detallada la brinda la referencia [14] la cual la describe como:

- Integración de información obtenida mediante los métodos tradicionales de la minería de datos con información recogida sobre la web.
- Descubrir los patrones interesantes en la estructura, contenido y la utilización de los sitios web.

Otra perspectiva más detallada de lo que es Web Mining es que la minería web es un proceso complejo que comprende el análisis de información diversa, como el contenido y estructura de los documentos web (HTML, XML), archivos de texto, bases de datos, bitácoras de acceso de usuarios, bitácoras (logs) de referencias de otros servidores, perfiles de usuarios y

otros, con el fin de encontrar información útil y relevante de acuerdo a las necesidades de un usuario.

En relación a los objetivos que persigue Web Mining, son los siguientes:

- Búsqueda de información relevante o relacionada
- Creación de nueva información a partir de la existente
- Personalización de la información
- Generar conocimientos a partir de los comportamientos de los usuarios web

3.2.2 Fases de Web Mining

En relación a las fases que involucra Web Mining, son las siguientes:

3.2.2.1 **Selección y recopilación de datos:**

En primer lugar decidir qué se quiere estudiar y cuáles son los datos que facilitarán esa información. Posteriormente, se localizan los documentos o archivos a adquirir, de los cuales se capturarán y se almacenarán los datos pertinentes.

3.2.2.2 **Tratamiento previo de los datos**

Se trata de filtrar y limpiar los datos recogidos. Una vez extraída una determinada información a partir de un documento, ya sea HTML, XML, texto, PS, PDF, Látex, etc. se realizan tareas de elección y normalización, eliminando los datos erróneos o incompletos, presentando los restantes de manera ordenada y con los mismos criterios formales hasta conseguir una homogeneidad formal, etc. y demás labores enfocadas a la obtención de unos datos originales listos para su transformación por medios automáticos.

3.2.2.3 **Transformación de los datos**

En esta fase se utilizan algoritmos inteligentes de búsqueda de patrones de comportamiento y detectar asociaciones. Estos algoritmos se elaboran previamente utilizando recursos estadísticos, técnicas procedentes del data mining, etc., para luego proceder a transformar los datos y así obtener como resultado información sobre ellos. Los principales algoritmos se basan en la reunión de grupos homogéneos (ej. Usuarios que visitan más de un número determinado de páginas), reglas de asociación de páginas, seguimiento de rutas o historial de navegación de una persona, etc.

3.2.2.4 **Análisis de las inferencias sobre los datos**

Una vez que los patrones han sido identificados, la parte humana juega un papel importante haciendo uso de herramientas adecuadas para entender, visualizar e interpretar los patrones. Las técnicas más comunes en el análisis de patrones son técnicas de visualización, técnicas de OLAP (On-line Analitycal Processing), consultas de datos y conocimientos y análisis de usabilidad [15].

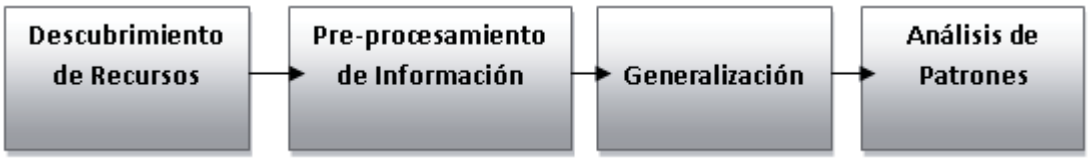


Figura 3.8 Fases de Web Mining

3.2.3 Clasificación de Web Mining

En relación a las distintas áreas o sectores en los que se centra el estudio de Web Mining son: Web Content Mining, Web Structure Mining y Web Usage Mining.

Figura 3.9 Clasificaciones de Web Mining [16]

3.2.3.1 **Web Content Mining**

En la web se puede encontrar una gran cantidad documentos heterogéneos, ya sean hipertextos, documentos de tipo texto, documentos pdf, etc. así como también imágenes, videos, música por lo que se dificulta el proceso de clasificación de información. De esta forma se puede definir Web Content Mining como el enfoque de Web Mining que tiene como objetivo recoger datos y descubrir patrones relativos a los contenidos web y a las búsquedas que se realizan sobre los mismos [2].

De acuerdo con Raymon Kosala y Hendrick Blockeel [17], la minería de contenido puede ser diferenciada desde dos puntos de vista; desde el punto de vista de la Recuperación de Información (IR) y desde la vista de Base de Datos (DB). El proceso de IR consiste en principalmente en asistir en el proceso de recogida de información o mejorar la información encontrada por los usuarios, usualmente basada en las solicitudes hechas por ellos mismos. Desde el punto de vista de DB principalmente trata de modelar los datos e integrarlos en la Web a través de consultas sofisticadas.

a. **Técnicas y Aplicaciones**: Las técnicas que se utilizan en Web Content Mining van a depender del tipo de documento con el que se trabaje, así se pueden nombrar las siguientes técnicas:
   - Text Mining: el cual consta principalmente de procesar información no estructurada, de tal forma que la información que se encuentra contenida en el documento sea accesible por los distintos algoritmos utilizados en Data Mining. Es importante señalar que Text Mining consta con técnicas de recuperación de información principalmente. Alguna de las áreas de aplicaciones en este campo son: indexación de grandes bancos de datos, extracción de información en distintos documentos y generación de conocimientos, entre otros.
   - Hypertext Mining: hace referencia a los enlaces existentes entre documentos y también intro-documentos; para lo anterior se construye un grafo de referencias.
   - Markup Mining: a partir de un documento que está basado en marcas se puede obtener información a partir de ellas, por ejemplo, en HTML (secciones, tablas, negritas: relevancia, cursiva, etc.), XML.
   - Multimedia Mining: se puede obtener conocimientos implícitos a partir de los datos multimedia que se encuentran en los documentos.

3.2.3.2 **Web Structure Mining**

Esta clasificación de Web Mining se centra principalmente en la estructura de los hiperlinks de la web, es decir, se centra en la entrada y salida de los links de las páginas. Los links que apuntan a una página puede sugerir la popularidad de la misma, mientras que los links que salen de la página demuestran los tópicos o riquezas de ella. Otra de las tareas de Web Structure Mining es descubrir la naturaleza de la jerarquía o la red de enlace en los sitios web de un dominio en particular. Esto puede ayudar a generalizar el flujo de información en los sitios web que pueden representar un dominio particular, por lo tanto, el proceso de la consulta será más fácil y más eficiente

a. **Técnicas y Aplicaciones:** Actualmente se utilizan con frecuencia algoritmos como PageRank y los HITS para modelar la topología de la web. En PageRank, cada página web tiene una medida de prestigio que es independiente de cualquier necesidad de información o pregunta, así el prestigio de una página es proporcional a la suma de las páginas que se ligan a él. PageRank es un valor numérico que representa lo importante que es una página en la web, de esta forma, mientras más votos tenga una página, más importante será la página. HITS (Hyperlink.induced topic research) es un algoritmo interactivo que tiene como finalidad excavar el grafo de la Web para identificar “hubs” y “authorities”, donde se entiende como

authorities a las páginas que de acuerdo a un tópico son las que mejor posicionadas están. Los *hubs* son aquellas páginas que hacen liga hacia las *authorities*. El número y el peso de *hubs* apuntando a una página determinan el nivel de posicionamiento de ella.

3.2.3.3 **Web Usage Mining**

Esta clasificación de Web Mining será explicada con más detalles en la próxima sección debido a que es el tema de estudio principal de ésta investigación.

3.3 Web Usage Mining

3.3.1 Introducción

Los servidores web pueden llegar a generar y acumular grandes cantidades de información sobre las interacciones que realizan los usuarios en el sitio web. De aquí, surge la necesidad de minar sobre éstos datos para lograr capturar y modelar los patrones de acceso general del sitio mediante el análisis de los "logs files" del servidor, los cuales contienen información como dirección IP del cliente, identificación del usuario, fecha y hora de acceso, URL de la página accedida, el protocolo utilizado para la transmisión de los datos, un código de error, agente que realizó el requerimiento etc., y los datos asociados a un determinado sitio web. En base a lo anterior se podría dirigir el estudio a varios enfoques de utilización o aplicación del Web Usage Mining (WUM): mejora del sistema, modificación del sitio, personalización y Business Intelligence (BI) [3]. Alguna de las técnicas de "Data Mining" que son utilizados para realizar lo anterior son: Reglas de Asociación, Patrones Secuenciales y Clasificación o Clustering.

3.3.2 Definición y Objetivos

Web Usage Mining (WUM) ó Minería de Uso en la Web, es una de las clasificaciones de Web Mining que está orientada principalmente a capturar y modelar los patrones de comportamiento del usuario en la web [18] la cual utiliza como principal fuente de datos los "logs files" del servidor web. Estos logs del servidor, los cuales serán estudiados en más detalles en los próximos puntos, son de gran relevancia para WUM debido a que mantienen grabaciones explícitas del comportamiento sobre la navegación de los visitantes del sitio.

En base a lo anterior se pueden definir dos objetivos principales en este tipo de minería [19]:

- Extraer patrones generales de uso de un sitio web de manera que pueda reestructurarse para que sea más fácil de utilizar y mejorar el acceso por parte de los usuarios.
- Obtener perfiles de los distintos tipos de usuarios a partir de su comportamiento y navegación para ofrecer una atención más personalizada.

Para lograr realizar los objetivos nombrados anteriormente, WUM utiliza como principales fuentes de datos los archivos de registro del servidor, que incluyen los logs de acceso al servidor

y los registros del servidor de aplicaciones. Otras fuentes adicionales que son también importantes para la preparación de datos y el descubrimiento de patrones son los archivos del sitio y meta-datos, bases de datos operacionales, plantillas de aplicación y conocimiento del dominio [20].

En general, los datos obtenidos mediante las fuentes mencionadas anteriormente, se pueden categorizar en cuatro grupos principales:

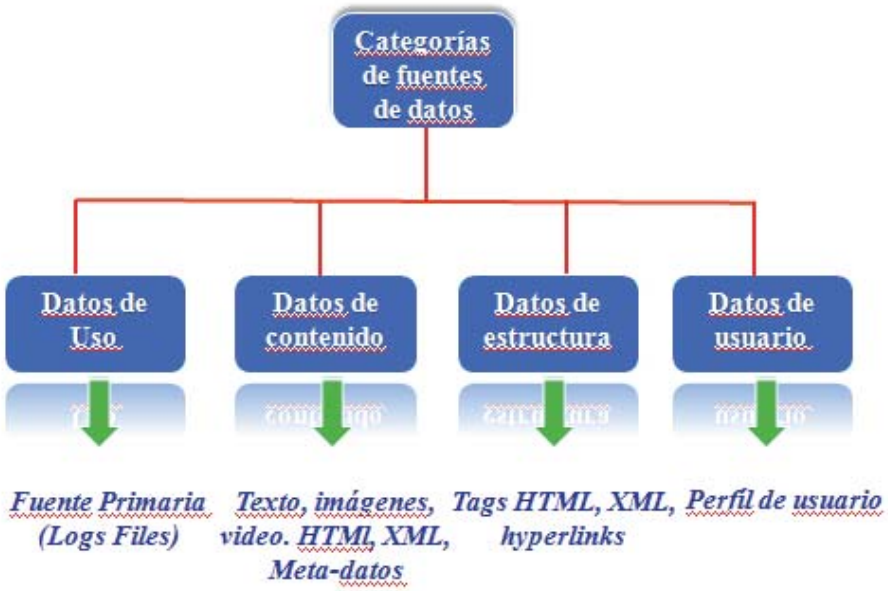


Figura 3.10 Fuentes de Datos para WUM

- Datos de uso: el registro de datos recogidos automáticamente por la Web y los servidores de aplicaciones representa el comportamiento de navegación de grano fino de los visitantes. Es la fuente primaria de datos en la minería de uso de la Web. Cada entrada que se realiza al servidor, corresponde a una petición HTTP, el cual genera una única entrada en los logs de acceso al servidor. Dependiendo del formato del registro, puede contener información como la identificación de la hora y la fecha de solicitud, dirección IP del cliente, el recurso solicitado, el método HTTP utilizado, el agente de usuario (tipo de navegador, sistema operativo y su versión) y si está disponible por el lado del usuario, las cookies que son las que identifican la visita repetida de un usuario. Un ejemplo de un segmento de un log de servidor se muestra en la Figura 3.11. Dependiendo de las metas del análisis, los datos necesitan ser transformados a diferentes niveles de abstracción. En WUM el nivel más básico de abstracción de datos es el de "page view" ó "vista de página" el cual físicamente corresponde a la acción de mostrar algún objeto en el navegador del usuario producto de una simple interacción como un clic. Conceptualmente, una vista de página representa un tipo específico de actividad del usuario en el sitio,

como por ejemplo, leyendo un artículo, agregando un producto al carrito de compras, etc. Desde el punto de vista del usuario, el nivel más básico de abstracción sería una sesión de usuario [20], la cual se puede definir como una secuencia de vistas de páginas durante una única visita.

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Figura 3.11 Segmento de un log de servidor típico

- Datos de contenido: son una colección de objetos y relaciones que son transmitidos al usuario. La gran parte de estos datos son una combinación de material textual e imágenes en donde los orígenes de estos datos pueden incluir páginas estáticas HTML / XML, imágenes, videos y archivos de sonidos. Además, el contenido del sitio incluye meta-datos semánticos o estructurales, palabras claves, atributos del documento, etc. Además el dominio ontológico subyacente también es considerado como datos de contenido (categorías de producto, jerarquías estructuradas, estructuras de directorios, etc.).
- Datos de estructura: la estructura de los datos representa la vista del diseñador en el contenido del sitio. Incluye la estructura de un sitio web como los tags de HTML o XML y la estructura de los hiperlinks.
- Datos del usuario: las bases operacionales del sitio podrían incluir información adicional sobre el perfil del usuario. Ésta, podría incluir datos demográficos sobre los usuarios registrados, relaciones explicitas o implícitas de los intereses de los usuarios, etc.

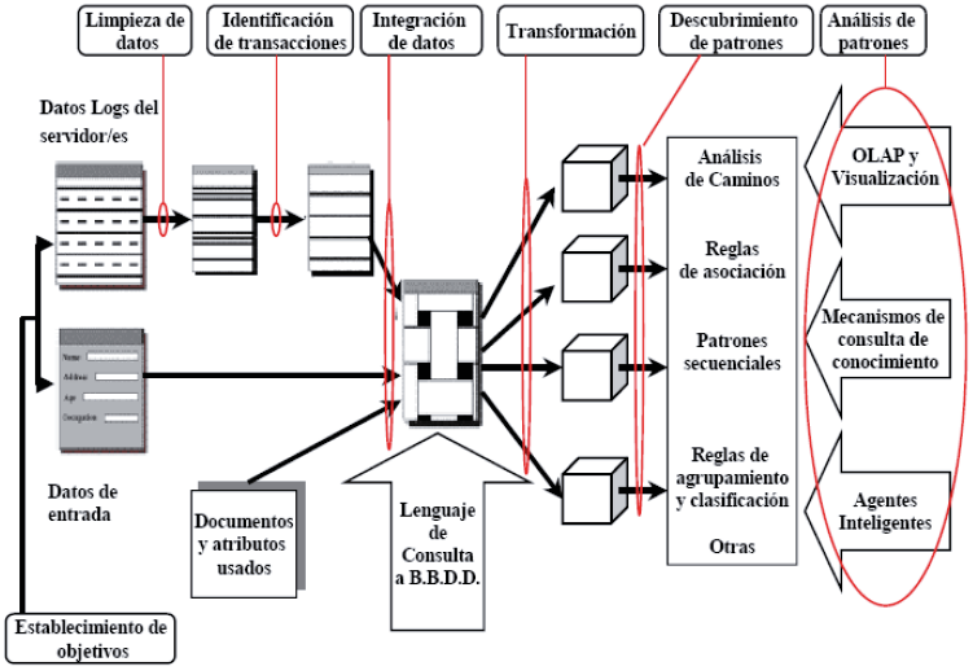


Figura 3.12 Proceso detallado de Web Usage Mining [20]

3.3.3 Proceso de Web Usage Mining

En esta sección se describirán las actividades que han de llevarse a cabo en WUM para lograr comprender el funcionamiento general de este proceso y las técnicas que pueden ser empleadas en cada una de sus etapas.

3.3.3.1 **Pre-procesamiento de los datos**

En esta fase de WUM, el objetivo se centra principalmente en convertir la información de uso, contenido y estructura contenida en varias de fuentes de datos disponibles, a abstracciones de datos necesarias para lograr descubrir patrones. Esta fase es frecuentemente la que utiliza más tiempo y recursos computaciones en el proceso de Web Usage Mining y muy a menudo requiere utilizar algoritmos especiales y heurísticas que no son comúnmente utilizadas en otros dominios [20]. Este proceso es crítico para la extracción exitosa de patrones útiles de los datos.

3.3.3.2 **Pre-procesamiento del uso**

Esta tarea se podría considerar como la más complicada en el proceso de WUM debido a la incompletitud de los datos disponibles. A menos que exista un mecanismo de seguimiento en el lado del cliente, sólo se tiene disponible la dirección IP, el agente y los “click-streams” (una secuencia de page views requeridas) por el lado del servidor para identificar a los usuarios y a las sesiones de servidor. Los problemas típicos que se encuentran en esta etapa son:

i. IP única / sesiones de varios servidores: un único servidor proxy podría tener muchos usuarios accediendo a un sitio web, potencialmente en el mismo período de tiempo.
ii. Múltiples IP / una sesión de servidor: una única sesión de servidor puede tener asociadas múltiples direcciones IP.
iii. Múltiples IP / un único usuario: un usuario que accede a internet de diferentes máquinas tendrán distintas direcciones IP de sesión a sesión. Esto hace que el seguimiento de las visitas de un mismo usuario se dificulte.
iv. Múltiples agentes / único usuario: un usuario que utilice más de un navegador web, incluso en la misma máquina, aparecerá como si fueran múltiples usuarios.

Una vez que cada usuario ha sido identificado, los click-streams de cada usuario deben ser divididos entre de sesiones (treinta minutos son frecuentemente utilizados para terminar una sesión de acuerdo a los resultados basados en [21]).

a. Pre-procesamiento del contenido: este proceso consiste en convertir el texto, imagen, scripts y otros archivos que contiene el sitio web, en formas que sean útiles para el proceso de WUM. En el contexto de WUM, el contenido de un sitio puede ser usado para filtrar la entrada o la salida de un algoritmo de descubrimiento de patrones. Un ejemplo de esto sería clasificar las vistas de página de acuerdo a su uso previsto.
b. Pre-procesamiento de la estructura: la estructura de un sitio es creada por los links de hipertexto entre vistas de páginas. La estructura puede ser obtenida y procesada de la misma manera que el contenido de un sitio. Sin embargo, en los contenidos que son dinámicos, podrían ocurrir problemas debido a que se podría estructurar de manera diferente tras cada sesión de servidor.

3.3.3.3 **Descubrimiento de Patrones**

El descubrimiento de patrones se basa en métodos y algoritmos desarrollados en varios campos, como las estadísticas, minería de datos, aprendizaje automático y reconocimiento de patrones. A continuación se describen algunas de las técnicas más utilizadas en la minería de datos para el descubrimiento de patrones.

- Reglas de asociación: son una técnica de minería de datos que encuentra interesantes asociaciones y/o correlaciones entre un grupo considerable de datos [24] y están asociadas a medidas de soporte y confianza. Uno de los ejemplos más típicos de aplicación de reglas de asociación es el análisis de canasta de mercado, donde básicamente consiste en encontrar relaciones entre los productos que compran los clientes lo cual ayudaría a formular estrategias mercadotécnicas, por ejemplo, un supermercado podría encontrar que de los 1000 clientes que compraron el Miércoles en la noche, 200 compraron pañales, y de aquellos que lo hicieron, 50 compraron cervezas. Por lo tanto, surge la siguiente regla: *"Sí compra pañales,*

entonces compra cerveza", con determinados niveles de soporte y confianza que se presentan a continuación:

Sea A= un cliente compra pañales y B= un cliente compra pañales y cervezas ⇒ P(A)= 200/1000 = 0,2 = 20% se tiene [5]:

$$\text{Soporte} = P(A \cap B) = \frac{Transacciones\ que\ contienen\ A\ y\ B}{Total\ de\ transacciones} = \frac{50}{1000} = 5\%$$

$$\text{Confianza} = P(B|A) = \frac{-P(A \cap B)}{P(A)} = \frac{5\%}{20\%} = 25\%$$

El objetivo principal es encontrar aquellas reglas cuyo soporte y confianza sean altos, sin embargo, si se dispone de una gran cantidad de atributos la combinatoriedad del problema se torna inmensa y para restringir el espacio de solución y hacerlo más manejable surge el *algoritmo A priori.*

- Algoritmo A priori: es la aproximación más popular de las técnicas de Data Mining para encontrar ítems frecuentes en una transacción y derivar reglas de asociación. Así, lo que realiza este algoritmo de modo general es:
    1. Genera todos los ítems sets con un elemento. Usa estos para generarlos de dos elementos, y así sucesivamente. Se toman todos los posibles pares que cumplen con las medidas mínimas de soporte. Esto permite ir eliminando posibles combinaciones ya que no todas se tienen que considerar.
    2. Genera las reglas revisando que cumplan con el criterio mínimo de confianza.
       Una observación importante es que si una conjunción de consecuentes de una regla cumple con los niveles mínimos de soporte y confianza, sus subconjuntos (consecuentes) también los cumple. Por el contrario, sí algún ítem no los cumple, no tiene caso considerar sus súper conjuntos [25].

  De esta forma, los pasos que sigue este algoritmo de modo descriptivo son los siguientes [26]:
    1. En la primera iteración, a partir de las URL iniciales, obtenemos los primeros conjuntos grandes de ítems (ej: si el umbral es 2, se determina qué URLs aparecen en, al menos, 2 sesiones)
    2. En cada paso subsiguiente, se parte del conjunto hallado en la pasada anterior
    3. Se crea un nuevo conjunto de grupos a partir de todas las combinaciones válidas de URLs iniciales (candidatos)
    4. Una combinación es válida en una sesión si todas las URLs pertenecen a la sesión
    5. Se mantienen los candidatos que cumplen con el soporte mínimo y se vuelve a iterar.

  En la Figura 3.13 se puede apreciar el diagrama de flujo correspondiente al funcionamiento del algoritmo para lograr entregar las reglas generadas.

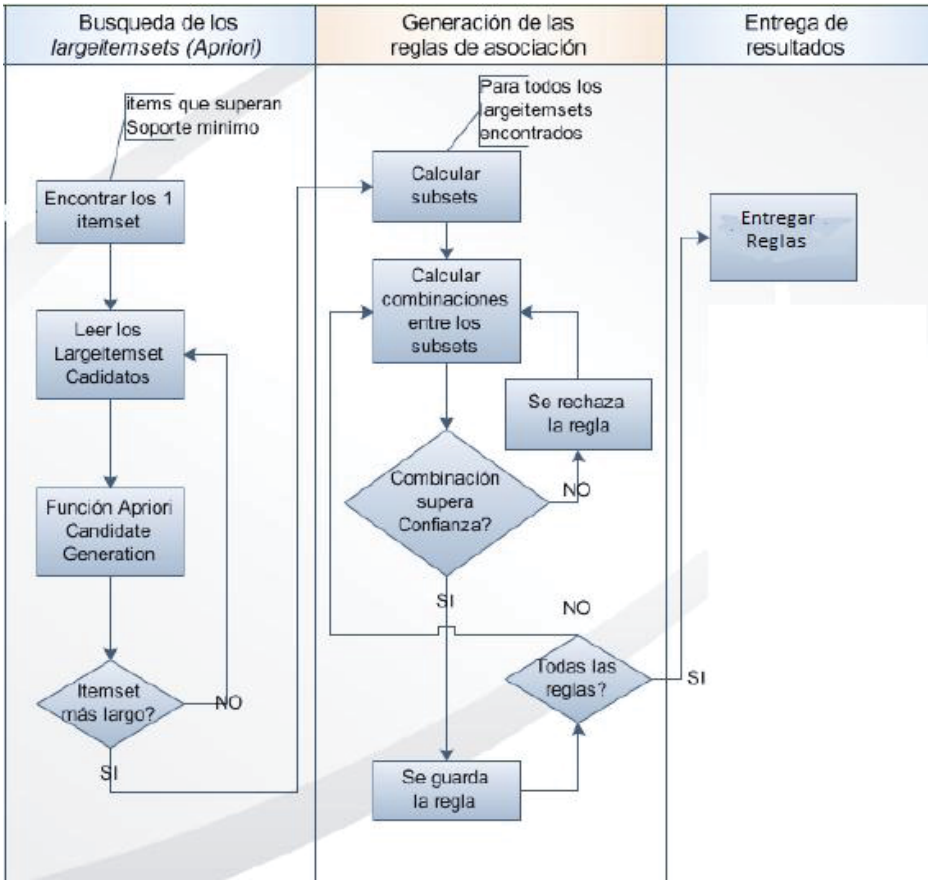


Figura 3.13 Diagrama de flujo del algoritmo A priori

- Análisis estadístico: ésta es la técnica más común para extraer conocimientos sobre los visitantes del sitio web. Analizando los archivos de sesión, se pueden realizar diferentes tipos de análisis descriptivos (frecuencia, media, etc.) en variables como *page views*, tiempo de visita y largo de la ruta de navegación. Estos conocimientos son potencialmente útiles para ayudar al mejoramiento del sistema facilitando las tareas de modificación y suministrando soporte para las tareas de marketing.
- Clasificación: se refiere a la tarea de mapear un ítem de datos dentro de muchas clases predefinidas. Ésta técnica puede ser utilizada basándose en algoritmos supervisados de aprendizaje inductivo, arboles de decisión, máquinas vectoriales, etc. Un ejemplo de utilización de esta técnica sería descubrir reglas como: el 30% de los usuarios que ordenan en línea en /company/products/music se encuentran entre los 18 y 25 años de edad y viven en Santiago.
- Clustering: ésta técnica es utilizada para identificar a los usuarios que comparten características comunes y así agruparlos dentro de conjuntos que mejor representen sus perfiles. Un ejemplo de esto sería: el 50% de los clientes que utilizaron su tarjeta de platino para comprar se encuentran en un grupo entre 25 – 30 años de edad con un ingreso anual entre $400.000 y $800.000.

- Patrones secuenciales: mediante ésta técnica se pretenden relacionar las distintas transacciones efectuadas por el cliente a lo largo del tiempo, así, sí las compañías basadas en tecnologías web pudiesen descubrir los patrones secuenciales de los visitantes, las compañías podrían predecir los patrones de visitas de los usuarios y el ámbito de mercado objetivo de un grupo de usuarios. Un ejemplo de lo anterior sería: el 50% de los clientes que compraron algún ítem en /pcworld/computadores, también compraron una orden en línea en /pcworld/accesorios después de 15 días.
- Modelado de dependencia: la meta en ésta técnica es desarrollar un modelo capaz de representar dependencias significantes a lo largo de varias variables en el dominio web. La información obtenida podría ayudar a desarrollar estrategias para aumentar las ventas de productos ofrecidos en el sitio web o para mejorar la navegación de los usuarios.
- Árboles de decisión: en ésta técnica se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.
- Análisis de clickstream: un *clickstream* consiste en una secuencia de páginas ejecutadas por un usuario en particular en un sitio web, así estos consisten en logs, cookies, metatags, y otros datos web utilizados para transferir páginas web desde un servidor al navegador [23]. Sin embargo, los datos de *clickstream* requieren de un pre procesamiento antes de analizar el comportamiento de los usuarios debido a que en éstos existe información embebida que no aporta información relevante. Una vez que éstos han sido pre procesados, los analistas se pueden realizar las siguientes preguntas: (1) ¿Cuáles son las páginas más comunes visitadas por los usuarios? (2) ¿Qué otros sitios referencian a los usuarios a este sitio web? (3) ¿Cuántas páginas son vistas en una visita típica? (4) ¿Cuántas páginas visita el usuario antes de salir del sitio? (5) ¿Qué tanto dura la visita de un usuario en el sitio? (6) ¿Cuáles son las páginas más comunes de puntos de salida de los usuarios? (7) ¿En qué orden son visitadas las páginas?

Otras técnicas que también se podrían utilizar y resultar muy prácticas en el descubrimiento de patrones en WUM son [15]:

- Conversión de direcciones IP a nombres de dominio: la dirección IP de un visitante puede ser convertida dentro de un nombre de dominio usando el sistema DNS en reversa el cual es llamado “Reverse DNS lookup”. Al convertir la dirección IP a un nombre de dominio se podría obtener algo de conocimiento, como por ejemplo, se podría estimar en qué lugar viven los visitantes del sitio mirando la extensión del nombre del dominio, como .cl (Chile), .ca (Canadá), etc.
- Análisis de ruta: los modelos de grafos son utilizados en este análisis. Un grafo representa una relación definida en las páginas web y cada árbol del grafo representa una página web; las uniones entre los grafos representan los links de ellas. Mediante este análisis se pueden descubrir problemas en la navegación del sitio, por lo que se podría realizar una reestructuración de él para facilitar el acceso a la página web.

- Análisis de cookies: las cookies son independientes de las direcciones IP y trabajan en sitios con un número substancial de visitantes desde los proveedores de servicios de internet (ISP). Sí un sitio web utiliza cookies, el campo de la cookie aparecerá en el log files y podría ser usado en el análisis de tráfico web para hacer un mejor trabajo sobre el seguimiento de los visitantes más recurrentes.

3.3.3.4 **Análisis de Patrones**

El análisis de patrones corresponde al último paso en el proceso de Web Usage Mining descrito en la Figura 3.12.

La motivación principal que existe detrás del análisis de patrones es filtrar o seleccionar los patrones o reglas que sean de interés que se encontraron en la fase de descubrimiento de patrones. La forma más común de análisis de patrones consiste en realizar un mecanismo de consultas como SQL. Otro método consiste en cargar los datos de uso en un cubo de datos para realizar operaciones OLAP (On-Line Analytical Processing) sobre él.

La información de contenido y estructura se pueden utilizar para filtrar patrones que contienen páginas de cierto tipo de uso, tipo de contenido o páginas que coinciden con una estructura de enlaces determinada.

3.3.4 Utilización de Web Usage Mining

En base a todo lo que se ha expuesto en las secciones anteriores, se pueden observar distintas áreas de aplicación de WUM, entre las que se destacan: personalización, mejora del sistema, modificación del sitio e inteligencia de negocios principalmente.

A continuación se describe cada una de las aplicaciones:

3.3.4.1 **Personalización**

Actualmente la personalización ha llegado a tener gran relevancia debido al exponencial crecimiento de información que ha ocurrido en la Web durante los últimos años. Es por esto, que actualmente existe una gran cantidad de estudios y trabajos sobre aproximaciones a la personalización web, utilizando distintos lenguajes y herramientas, los cuales pretenden mejorar la experiencia del usuario en la web.

Un ejemplo de utilización de personalización web se puede apreciar en el *e-commerce*, debido principalmente a que en estos tipos de ambientes web es crucial que el usuario navegue en el sitio de una forma eficiente, y en lo posible, ofreciendo información que sea de relevancia para él utilizando, por ejemplo, recomendaciones dinámicas basadas en el perfil del usuario, etc.

3.3.4.2 **Mejora del Sistema**

La performance y otros atributos de calidad son cruciales para la satisfacción del usuario. Web Usage Mining provee la clave para entender el comportamiento del tráfico en la web, el cual puede ayudar para el desarrollo de políticas de web caching, transmisión en la red, distribución

de datos, etc. Además, WUM puede proveer patrones que son útiles para la detección de intrusos, fraudes electrónicos, entre otros.

Junto con lo anterior, se han desarrollado algoritmos para crear rutas de perfiles de datos contenidos en los logs de los servidores, los cuales son usados para pre-generar páginas HTML dinámicas basadas en los perfiles actuales de los usuarios para reducir los tiempos de latencia de generación de la página.

3.3.4.3 **Modificación del Sitio**

La atractividad de una página web, en términos de contenido y estructura, es crucial para muchas aplicaciones, como por ejemplo, los catálogos de productos de los *e-commerce*.

WUM provee retroalimentación detallada sobre el comportamiento del usuario, la cual es de gran utilidad para el diseñador del sitio para tomar decisiones de rediseño de la página. Junto con esto, se pueden identificar ciertos problemas de estructuración de la página, los cuales podrían afectar a la visualización de ciertos objetos de la misma, por lo que en éstos casos se podría optar por reestructurar el sitio para lograr cumplir con los objetivos de los usuarios.

3.3.4.4 **Inteligencia de Negocios**

En este momento, debido a la gran cantidad de negocios que se manejan por Internet, la gran competencia y la creciente necesidad de mejorar los servicios, el análisis de los datos que se obtienen para convertirlos en información útil se torna imprescindible para poder sobrevivir en este ambiente competitivo.

Para esto, es necesario conocer el comportamiento de los usuarios (potenciales clientes) y brindarles un acceso más fácil y un mejor servicio, así como también saber hacia quien orientar las campañas promocionales mediante la utilización de la información que se encuentra alojada en los servidores web, como por ejemplo, los logs files, los cuales aportan importantísima información y que por medio de Web Usage Mining, se pueden obtener conocimientos que servirán de ayuda a las áreas gerenciales de la empresa para poder tomar decisiones estratégicas.

3.3.5 Herramientas utilizadas en WUM

En esta sección se presentarán algunas herramientas que son utilizadas por Web Usage Mining para analizar logs u otra información relevante, empleando una diversidad de algoritmos que traen integradas cada una de ellas, los cuales serán analizados y estudiados en la segunda fase de esta investigación, de tal forma de realizar la selección de la herramienta y algoritmo(s) lo más óptima posible para lograr dar solución al problema.

3.3.5.1 **WEKA**

Ésta es una potente herramienta desarrollada por la *Universidad de Waikato,* la cual contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario la cual permite acceder fácilmente a sus funcionalidades.

Las principales ventajas que posee esta herramienta son las siguientes:

- Está disponible libremente bajo la licencia pública general de GNU.
- Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Contiene una extensa colección de técnicas para pre-procesamiento de datos y modelado.
- Es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

3.3.5.2 **WebMiner**

WebMiner es un sistema que divide el proceso de WUM en dos partes: la primera parte, incluye el proceso del dominio de la transformación de los datos web en formatos que se ajusten a las transacciones. Esto incluye pre-procesamiento, identificación de transacción e integración de componentes de datos. La segunda parte incluye técnicas de minería de datos y reconocimiento de patrones.

Las principales ventajas que presenta esta herramienta son [27]:

- Presenta modelos de datos y de transacciones para varias tareas del WUM como el descubrimiento por asociación de reglas y patrones secuenciales para los datos Web.
- Aplica técnicas de descubrimiento de conocimiento.
- Define formalmente el registro de las entradas (web logs).
- Define asociación de transacciones Web.
- Hace análisis de patrones secuenciales.
- Propone como trabajo futuro el desarrollo de agentes autónomos que analicen el descubrimiento de reglas de clasificación de tendencias para proveer sugerencias a los usuarios.
- Propone otro trabajo futuro que desarrolle un mecanismo de consultas que pueda ser manipulado en el pre descubrimiento (limpieza de datos, identificación de transacciones).

3.3.5.3 **Clementine**

Clementine es una herramienta de minería de datos de SPSS Inc., una compañía de IBM. Clementine admite la integración con herramientas de modelado y minería de datos disponibles en proveedores de bases de datos como Oracle Data Miner, IBM DB2 Intelligent Miner y Microsoft Analysis Services 2005. Clementine es fácil de aprender a utilizar, su interfaz visual hace que no sea necesario contar con habilidades de programación, reduce la curva de aprendizaje y les proporciona a usuarios expertos y novatos poder analítico. La arquitectura abierta y escalable de Clementine permite que se realicen varios procedimientos en una base de datos central, incluyendo el acceso a algoritmos propios del manejador de la base de datos. Esto puede ser de gran ayuda para maximizar la base de datos para incrementar el desempeño y la velocidad.

3.3.5.4 **Knime**

Es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual el cual está construido bajo la plataforma de Eclipse y programado esencialmente en Java.

Knime contiene una serie de nodos (que encapsulan distintos tipos de algoritmos) y flechas (que representan flujos de datos) que se despliegan y combinan de manera gráfica e interactiva. Los nodos implementan distintos tipos de acciones que pueden ejecutarse sobre una tabla de datos:

- Manipulación de filas, columnas, etc., muestreos, transformaciones, agrupaciones, etc.
- Visualización (histogramas, etc.).
- Creación de modelos estadísticos y de minería de datos, como árboles de decisión, máquinas de vector soporte, regresiones, etc.
- Validación de modelos, como curvas ROC, etc.
- *Scoring* o aplicación de dichos modelos sobre conjuntos nuevos de datos.
- Creación de informes a medida gracias a su integración con BIRT.

El carácter abierto de la herramienta hace posible su extensión mediante la creación de nuevos nodos que implementen algoritmos a la medida del usuario. Además, existe la posibilidad de llamar directa y transparentemente a Weka y/o de incorporar de manera sencilla código desarrollado en R o python/jython.

3.3.5.5 **Pentaho Open BI**

La Open BI Suite de Pentaho, provee un completo espectro de funcionalidades de Business Intelligence (BI, Inteligencia de Negocios), incluyendo reportes, análisis, tableros de control, minería de datos, integración de datos y una plataforma de BI que la han convertido en la suite de código abierto más popular del mundo. Los productos Pentaho son utilizados por organizaciones líderes tales como MySQL, Motorola, Terra Industries, DivX entre otras.

Pentaho Corporation es el patrocinador principal y líder del proyecto Pentaho BI. El proyecto Pentaho BI es una iniciativa en curso de la comunidad *open source* que provee a las organizaciones de las mejores soluciones de su clase para sus necesidades de inteligencia de negocios. Al aprovechar la riqueza de las tecnologías de código abierto y las contribuciones de la comunidad de desarrollo de código abierto, Pentaho es capaz de innovar mucho más rápido que los proveedores comerciales. Como resultado, Pentaho ofrece una alternativa de código abierto que supera a las soluciones de Business Intelligence propietarias en muchas áreas como arquitectura, soporte de estándares, funcionalidad y simplicidad de implantación.

3.3.5.5.1 **Mondrian**

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Mondrian es un servidor OLAP open source que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente.

Es decir, Mondrian actúa como "JDBC para OLAP" el cual permite crear cubos de información para análisis multidimensional.

Dichos cubos se componen de archivos XML y en ellos se definen las Dimensiones y las conexiones de los datos. Los archivos XML por lo general son complejos de realizar manualmente por lo que es común utilizar herramientas gráficas para realizar la edición de estos. Como ejemplo de estas herramientas Open Source Pentaho se tiene a Cube Designer para la Creación de cubos y el Workbench para la edición de los mismos.

4 Diseño de la Solución Propuesta al Sistema de Biblioteca PUCV

4.1 Introducción

En este capítulo se presentará el diseño conceptual de la solución al problema, el cual será la base para la posterior implementación del mismo. Así, se definirán los componentes y características de la solución propuesta.

Para comenzar, se expondrá los requisitos mínimos que se necesitan para realizar el estudio planteado, se realizará una descripción general sobre los componentes que se utilizarán y las relaciones entre ellos a modo de comprender de modo general lo que se profundizará en puntos posteriores.

Luego se describirá las etapas de la cual consta el diseño de la solución y los procesos que se llevarán a cabo en cada una de ellas y que se puede apreciar gráficamente en la Figura 4.1.

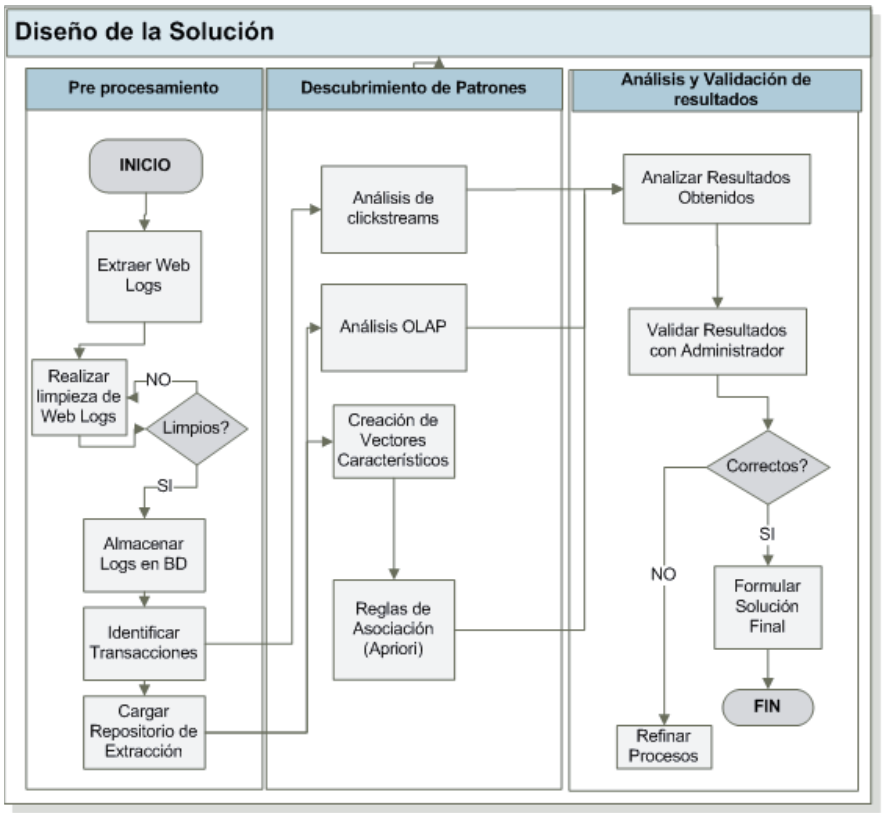


Figura 4.1 Diseño de la Solución

4.2 Requisitos para el estudio

Para lograr llevar a cabo este estudio, es necesario contar con toda la información del sitio web que se encuentre disponible, sin embargo, debido a la criticidad de la información que contiene un sitio (desde el punto de vista del código fuente, posible información de clientes almacenadas en bases de datos, los registros de los *weblogs*, etc.), es difícil conseguir toda la información que se estimase conveniente para realizar un mejor estudio.

Sin embargo, al ser el principal elemento de estudio los *weblogs* del sitio, los modelos que se plantearán en los puntos posteriores estarán basados en ellos principalmente y en información de las URL del mismo sitio, del cual se puede extraer información sobre el contenido de la misma y así poder agruparlas en distintos niveles de abstracción para facilitar el proceso de generación de vectores característicos para los algoritmos de Data Mining que se aplicarán.

Para la etapa de análisis de los patrones de comportamiento descubiertos, así como también análisis estadístico, se necesitará de herramientas que apoyen este proceso, las cuales se utilizarán herramientas que son gratis y otras que no, a modo de realizar comparaciones sobre los resultados obtenidos y llegar a mejores conclusiones.

4.3 Pre procesamiento

4.3.1 Extracción de Web Logs

El primer paso para realizar este estudio es obtener los "web logs" desde el servidor donde se encuentra alojado el Sistema de Biblioteca PUCV, de los cuales se encuentra disponible desde el año 2008 hasta marzo del 2010. Para efectos de este estudio se considerará los últimos 6 meses más actuales, que corresponden a los meses de:

- Octubre (2009)
- Noviembre (2009)
- Diciembre (2009)
- Enero (2010)
- Febrero (2010)
- Marzo (2010)

4.3.2 Limpieza de datos

Teniendo en cuenta que en los logs de los servidores web se almacena prácticamente cada acción que el usuario realiza en el sitio web, muchos de ésos logs van a contener información que no será de relevancia para estudiar los patrones de comportamiento de los usuarios, por lo tanto será necesario eliminar éstos registros, lo que a su vez reducirá considerablemente el tamaño del archivo que contiene los logs y agilizará el proceso de transformación y carga de los mismos en el repositorio de extracción que será detallado en los próximos puntos.

*IP			*TimeStamp	*MET/URL/Protocolo	*Estado	*Bytes	*Referer	*Agent
201.215.55.54	-	-	[17/Apr/2008:15:18:16	GET /CSS/popover.css HTTP/...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:16	GET /CSS/head.css HTTP/1.1	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:16	GET / HTTP/1.1	200	57837	http://www.ucv.cl/si...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.246.42.96	-	-	[17/Apr/2008:15:18:17	GET / HTTP/1.1	200	57837	http://www.ucv.cl/si...	Opera/9.26 (Windows NT 5.1; U; es-es)
158.251.56....	-	-	[17/Apr/2008:15:18:18	GET /cubiculos/css/styles.css ...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
158.251.56....	-	-	[17/Apr/2008:15:18:18	GET /cubiculos/sistema/cubicu...	200	36470	-	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
158.251.56....	-	-	[17/Apr/2008:15:18:18	GET /cubiculos/sistema/menu...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /scripts/mata_frames.js ...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /scripts/precarga_ddrive...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /scripts/fw_menu.js HTT...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /scripts/dhtmltooltip.js H...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /img/tooltip_img/arrow2....	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET / HTTP/1.1	200	57837	-	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /img/arrows.gif HTTP/1.1	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /img/fwmenu1_304x17_...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.215.55.54	-	-	[17/Apr/2008:15:18:19	GET /img/fwmenu6_188x17_...	304	?	http://biblioteca.ucv...	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322
201.246.42.96	-	-	[17/Apr/2008:15:18:19	GET /img/tooltip_img/cincel.jp...	404	351	http://biblioteca.ucv...	Opera/9.26 (Windows NT 5.1; U; es-es)

Figura 4.2 Extracto del log del sitio web en estudio

En la Figura 4.2 se puede apreciar un extracto del log que contiene los accesos que se realizan al sitio en donde se puede apreciar los campos que contiene y que fue definido por el *webmaster* al crear el sitio. Cabe mencionar que se traspasaron los datos a una tabla para presentarlos de mejor forma y para que fuese más legible por el lector.

Observando la figura se puede apreciar que contiene la siguiente información:

- **IP**: dirección IP del usuario
- **TimeStamp**: el tiempo en el cual se responde cada petición
- **MET/URL/Protocolo**: contiene el método de petición de la página, la dirección URL que se solicita y el protocolo utilizado
- **Estado**: el estado en que se encuentra la página solicitada (404: la página no se encuentra disponible)
- **Bytes**: cantidad de bytes retornados en cada petición
- **Referrer**: dirección de origen desde donde se hacen las peticiones
- **Agent**: el web browser utilizado por el usuario

Analizando las URLs solicitadas se pueden observar los objetos que no aportan al descubrimiento de conocimiento y por lo tanto se pueden eliminar. Alguno de estos objetos son [5]:

- Hojas de estilo en cascada CSS
- Imágenes y documentos (.gif, .jpeg, .png, .pdf,.xls, .doc, .jpg, .ico, etc.)
- Videos
- Códigos Javascript

Una vez eliminados estos objetos se obtienen registros más limpios con los que realmente se puede generar conocimientos a partir de ellos, sin embargo, existen unos programas llamados "*Web Crawlers*" o conocidos también como "*spyder robots"* los cuales recorren la Web

guardando copias de las páginas de diversos sitios de forma automatizada. Son frecuentemente usados por motores de búsqueda como Google o Yahoo! para recopilar las últimas versiones de las páginas web e indexarlas en sus servidores, acelerando con ello las búsquedas [23]. Su paso por los sitios deja los *weblogs* con gran cantidad de registros de peticiones en pocos segundos.

Por lo tanto, también es necesario eliminar los registros que dejan estos *spyder robots* para que no altere los resultados finales al momento de realizar análisis sobre los patrones descubiertos, ya que al incluirlos, se estaría considerando por ejemplo, a un usuario que accedió a una gran cantidad de páginas en muy poco tiempo lo cual claramente no sería posible.

Tres heurísticas son descritas en [10] para identificar los registros de los *Web crawlers* para posteriormente realizar la eliminación de ellos. Una de estas heurísticas es calcular una velocidad de navegación como: VN= (número de vistas de páginas) / (duración de la sesión en segundos).

Sí el VN excede un cierto umbral θ_1 (páginas/segundo) y el número de vistas de página de las

visitas excede otro umbral θ_2, entonces el host es considerado como un *Web crawler*. Otra heurística es observar todos los *hosts* que realizan solicitudes al archivo '*Robots.txt'* el cual contiene las reglas de navegación para los *Web crawlers* que indexan el sitio web. Este archivo es la primera solicitud que debería realizar un *Robot* o *Spyder* al sitio web, sin embargo, no es obligación seguir las reglas que posee este archivo y existen muchos Robots que simplemente ignoran este archivo. Y una tercera heurística utilizada es usar una base de datos de W*eb crawlers* la cual contenga la información de cada uno de ellos, y por medio de ésta ir comparando los registros existentes en los Web Logs con los registros de la base de datos e ir filtrando aquellos que estén presentes. En [28] se encuentra una base de datos actualizada con la información de los *web crawlers* existentes y que será la utilizada en este Proyecto.

Una vez "limpios" los logs, éstos serán almacenados en una tabla relacional en MySql, para poder acceder a ellos de manera eficiente y así realizar las tareas posteriores.

4.3.3 Identificación de Transacciones

Una vez que los *weblogs* se encuentran "limpios" se puede empezar a realizar las tareas previas a la aplicación de los distintos algoritmos de Web Mining, que en este caso sería identificar los usuarios y sus sesiones así como las construcciones de las mismas y luego llevarlas a un repositorio de extracción el cual contendrá la información referente a los usuarios, sus sesiones y las peticiones de cada uno de ellos para así empezar a generar los vectores característicos para los algoritmos específicos que se aplicarán para descubrir patrones de uso de los usuarios.

4.3.3.1 **Identificación de usuarios**

Si un usuario tuviese que ingresar con un nombre de usuario y contraseña cada vez que ingresa a la página entonces se tendría una completa identificación de cada uno de los accesos al sitio, sin embargo, este no es el caso para el sitio web que se está estudiando, ya que no posee ningún tipo de autenticación . Es por esto, que es necesario plantear una heurística de reconocimiento de los usuarios para luego poder construir las sesiones de cada uno de ellos ya

que un mismo usuario puede tener muchas sesiones asociadas a él en diferentes instantes de tiempo.

La forma más común de identificación de usuarios, y que es la que se utilizará en este estudio específico, es mediante el uso del campo IP y Agente, de tal forma que si una misma dirección IP y el mismo Agente aparecen en varios registros de los *weblogs* (que se encuentran limpios), entonces se puede concluir que esos registros pertenecen a un mismo usuario. En la Figura 4.3 se puede apreciar un ejemplo utilizado en [20] para identificar a los usuarios mediante este método.

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Figura 4.3 Ejemplo de identificación de usuario utilizando IP + Agente

4.3.3.2 **Identificación de sesiones de usuarios**

Una sesión de usuario se define como una secuencia de solicitudes hechas por un simple usuario en un cierto período de navegación y un usuario podría tener uno o múltiples sesiones durante un periodo de tiempo [29].

Existen distintos métodos para identificar las sesiones de usuarios, sin embargo, las más comunes o utilizadas son las siguientes [30]:

- Uso de un *timeout:* este método es el más utilizado, en donde se usa 30 minutos de duración por sesión, es decir, se van tomando cada una de las solicitudes en el log y se va realizando una comparación si está o no dentro del umbral definido, de ser así, ésa petición pertenecerá a la sesión *k* de lo contrario pertenecerá a la sesión *k+1*.
- IP/Agente: si se tienen diferentes Agentes para una misma dirección IP entonces se consideran como distintas sesiones.

- Referring Page: si la página *Referrer* de una solicitud no es parte de una sesión abierta, entonces se asume que la solicitud pertenece a una sesión diferente.

Para este caso de estudio específico se utilizará la identificación de sesiones por medio de uso de un *timeout*, utilizando 30 minutos como umbral de referencia. En la Figura 4.4 se puede apreciar visualmente la utilización de este método.

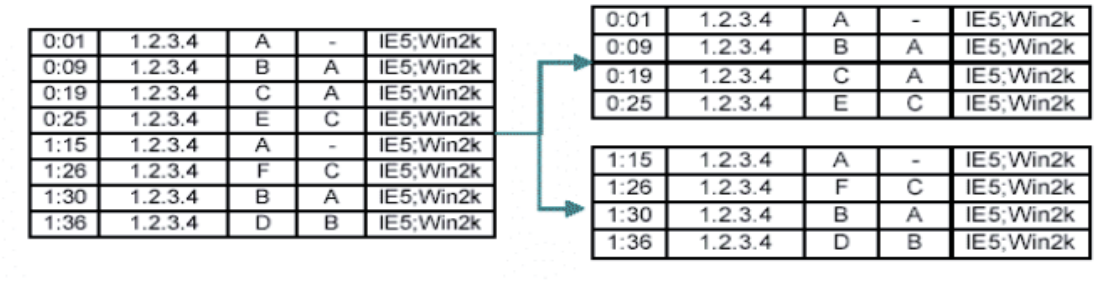


Figura 4.4 Utilización de timeout=30 min. para identificar sesiones

4.3.4 Repositorio de Extracción

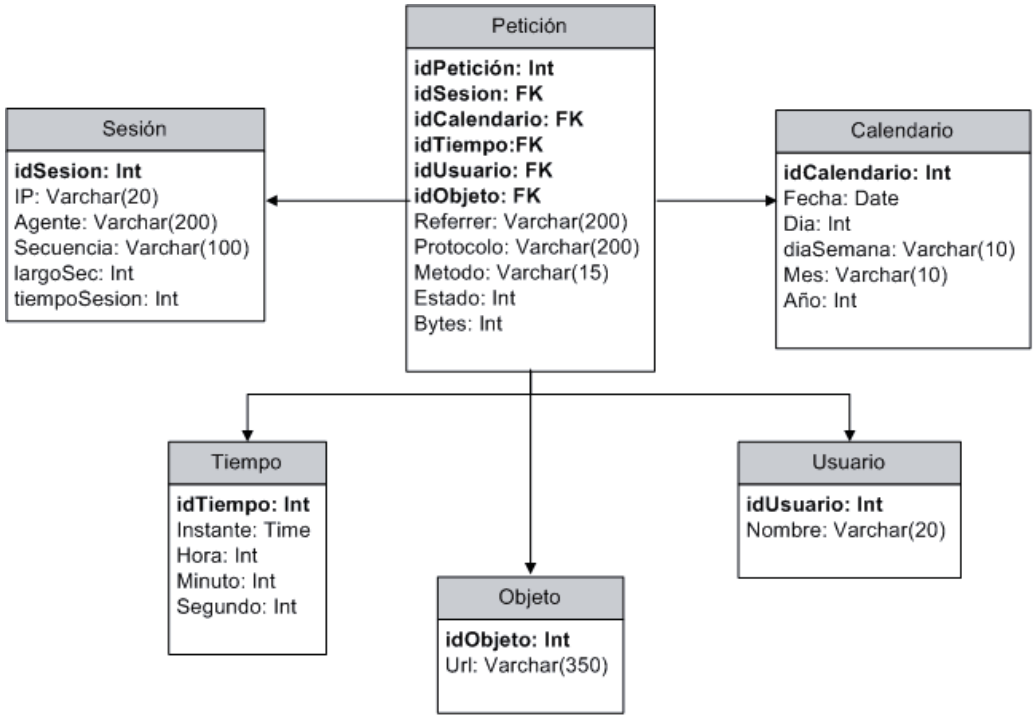


Figura 4.5 Repositorio de extracción

Este repositorio de extracción tiene como fin almacenar cierta información del sitio web en cuestión así como la información que contienen los web logs en una estructura adecuada para Web Mining. Además, a partir de este repositorio se podrá obtener los vectores característicos que serán las entradas para los algoritmos de Data Mining que serán aplicados para analizar el comportamiento que tienen los usuarios en el sitio.

Junto con lo anterior, también será posible realizar un análisis tipo OLAP debido a la estructura que presenta este repositorio, en este caso, es un modelo estrella que es el más característico de los *Data Warehouse* y por lo tanto, será posible obtener fácilmente consultas orientadas al análisis del sitio.

- Dimensión ***Objeto***: registra información de cada objeto que contiene el sitio, que en este caso específico se refiere principalmente a las páginas web que pueden ser accedidas por el usuario.

Tabla 4.1 Descripción de la dimensión *Objeto*

Columna	**Descripción**
Idobjeto	Es la clave primaria en la tabla y sirve como clave foránea en la **tabla Petición**
URL	La dirección URL del objeto, por ejemplo : *http://biblioteca.ucv.cl/proyectos/*

- Dimensión ***Usuario***: registra información de los usuarios que acceden al sitio y acceden los distintos objetos que este contiene.

Tabla 4.2 Descripción de la dimensión *Usuario*

Columna	**Descripción**
Idusuario	Es la clave primaria en la tabla y sirve como clave foránea en la **tabla Petición**
Nombre	Es el nombre del usuario que accede el sitio. Para este caso específico, sólo se tendrá un usuario con nombre Anónimo, debido a que los usuarios no pueden ser identificados.

- Dimensión ***Calendario***: registra información de la fecha en que se realizó la petición del objeto.

Tabla 4.3 Descripción de la dimensión *Calendario*

Columna	Descripción
Idcalendario	Es la clave primaria en la tabla y sirve como clave foránea en la **tabla Petición**
Fecha	Registra la fecha en formato *dd:mm:yyyy*
Día	Registra el día de la semana en que se realizó la petición. Varía de 1 a 7.
diaSemana	Corresponde al nombre del día de la semana en que se realizó la petición. Por ejemplo. *Lunes, Martes,* etc.
Mes	Corresponde al nombre del mes en que se realizó la petición. Por ejemplo. *Marzo, Abril,* etc.
Año	Corresponde al año en que se realizó la petición. Por ejemplo. *2008.*

- Dimensión ***Tiempo***: registra información del instante de tiempo en que se realizó la petición. Los campos de esta tabla se pudieron haber agregado a la tabla anterior, sin embargo se separaron para tratar de evitar el crecimiento desmedido de ella debido a la gran cantidad de logs que se generan en una cierta cantidad de tiempo.

Tabla 4.4 Descripción de la dimensión *Tiempo*

Columna	Descripción
Idtiempo	Es la clave primaria en la tabla y sirve como clave foránea en la **tabla Petición**
Instante	Registra el instante en que se realizó la petición en el formato *hh:mm:ss*
Hora	Corresponde a la hora en que llega la petición. Va desde 0 a 23.
Minuto	Corresponde al minuto en que llega. Va desde 0 a 59.
Segundo	Corresponde al segundo en que llega la petición. Va desde 0 a 59.

- Dimensión ***Sesión***: registra la información referente a las sesiones identificadas en el proceso de sesionización

Tabla 4.5 Descripción de la dimensión *Sesión*

Columna	Descripción
Idsesión	Es la clave primaria en la tabla y sirve como clave foránea en la **tabla Petición**
IP	Corresponde a la dirección IP con la cual se realiza la petición.
Agent	El nombre del web browser con el que se realiza la petición
Secuencia	Corresponde al User Behavior Vector que representa la sesión. En otras palabras la secuencia *s=[(p1),(p2),..(p_n)]* donde p_i es el identificador del objeto solicitado.
largoSec	Corresponde al largo de la secuencia anterior. Por ejemplo, si en la sesión se visitaron 5 sitios, entonces el valor de este campo es 5.
tiempoSesion	Corresponde al tiempo gastado entre la primera y última página visitada en segundos. Si sólo posee una página visitada, el valor de este campo será 0.

- Tabla Fact ***Petición***: cada registro de esta tabla corresponde a una entrada válida de web log.

Tabla 4.6 Descripción de la tabla Fact *Petición*

Columna	Descripción
Claves Foráneas	Los campos *Idobjeto, Idsesion, Idcalendario, Idtiempo e Idusuario* corresponden en conjunto a la clave primaria de esta tabla.
Referrer	Guarda la dirección URL de origen de donde se realiza la petición. Esta dirección puede ser interna al sitio o bien externa.
Protocolo	Guarda el protocolo de comunicación entre el Web Browser y el Web Server. Ejemplo : *HTTP /1.1*
Método	Registra el método usad por el browser para realizar la petición. Por ejemplo: *GET*
Bytes	Registra la cantidad de bytes retornados en cada petición

En base a este repositorio de extracción, basado en un modelo de Data Webhouse, será posible realizar análisis multidimensional mediante la generación de cubos OLAP para obtener

información de manera rápida y sencilla sobre las peticiones realizadas en el Sistema de Biblioteca.

Para la generación de éstos cubos, se utilizará un servidor de consultas OLAP Open Source llamado **Mondrian** el cual permite a los usuarios relacionados al sistema, analizar de manera interactiva y con un gran rendimiento pequeñas o grandes cantidades de información en tiempo real y así sacar conclusiones a partir de ésta. Además, permite explorar los datos de forma dimensional, y debido a que utiliza un modelo relacional permite realizar consultas directas mediante SQL, entre otras características.

Un ejemplo de la información que se podría extraer por medio de consultas OLAP en relación a la información almacenada en el Data Webhouse sería por ejemplo, consultar sobre las cantidades de peticiones realizadas durante los distintos periodos de tiempo considerados en el estudio lo que permite concluir sobre las fechas en que el sistema es utilizado con mayor frecuencia.

4.4 Descubrimiento de Patrones

4.4.1 Análisis OLAP

A partir del repositorio de extracción, se podrá generar consultas OLAP en base a las distintas dimensiones que conforman este modelo debido a las características que posee como tal.

De esta forma, utilizando *Mondrian*, se podrá generar cubos por medio de cruses de información de las distintas dimensiones que conforman el repositorio. Por ejemplo, se podrá obtener la cantidad de peticiones realizadas organizadas por fechas, incluyendo meses, días o incluso instantes en que el sistema fue más solicitado.

4.4.2 Análisis de Clickstrems

A partir de las sesiones encontradas en los logs pre procesados, es posible inferir información de gran utilidad sobre la navegación que realizan los usuarios por el sistema. De esta forma, es posible responder una serie de preguntas que darán ideas sobre posibles problemas en el sistema:

- ¿Cuáles son las páginas más comunes visitadas por los usuarios?
- ¿Qué otros sitios referencian a los usuarios a este sitio web?
- ¿Cuántas páginas son vistas en una visita típica?
- ¿Cuántas páginas visita el usuario antes de salir del sitio?
- ¿Qué tanto dura la visita de un usuario en el sitio?
- ¿Cuáles son los puntos de salida más comunes de los usuarios?

A partir de este análisis será posible observar posibles ideas de reestructuración del sitio de acuerdo a los accesos realizados en las distintas sesiones por los usuarios, también será posible observar si los usuarios navegan o no durante un tiempo considerable en el sistema o simplemente entran a él y salen rápidamente, lo cual podría indicar posibles problemas de diseño

del sitio o verificar el por qué existe un gran número de usuarios que dejan el sistema en una determinada página, etc.

4.4.3 Reglas de Asociación y Algoritmo Apriori

Antes de generar reglas de asociación es necesario crear los vectores característicos que serán la entrada para el algoritmo, así, se puede definir un vector que contiene los objetos que el usuario visitó asociado con un *peso* que representa la importancia de ése objeto o página en ésa sesión.

La forma que tiene este vector es la siguiente:

$$V = [(p_{1,}, w(p_1^t)), (p_{2,}, w(p_2^t)),\ldots, (p_l, w(p_l^t))]$$

Donde:

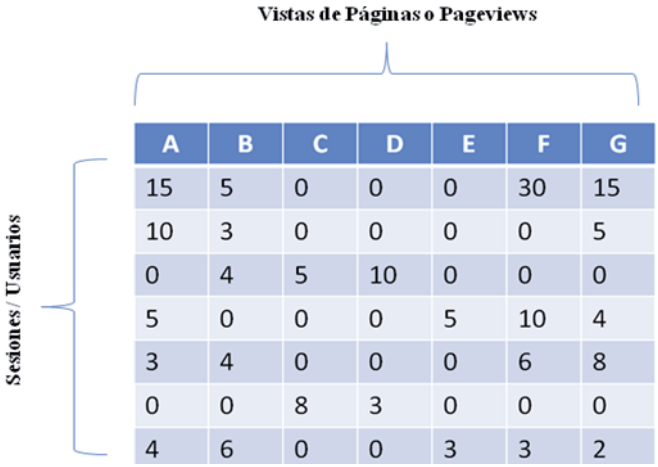
- V: secuencia de largo *l* de pares ordenados que representa las transacciones realizadas por el usuario en la sesión.
- p_i: página visitada por el usuario en la sesión con $0 \leq i < n$, *donde n* es el número de páginas que contiene el sitio web.
- $w(p_i^t)$ es el peso asociado a la vista de página o *page view*.

El peso $w(p_i^t)$, puede ser calculado de diversas formas [20]:

- En filtrado colaborativo, los pesos podrían estar basados en calificaciones de los elementos.
- En la mayoría de las tareas de Web Usage Mining es un valor binario (0,1) que representa la existencia o no de una *page view*, en la transacción.
- O puede ser una función de duración de la *page view* en la sesión del usuario (usualmente el tiempo gastado no se encuentra disponible).

A modo de ejemplo se presenta la matriz de vistas de páginas del usuario o *user-pageview matrix* (UPM) la cual se presentan sólo las siguientes páginas del sitio:

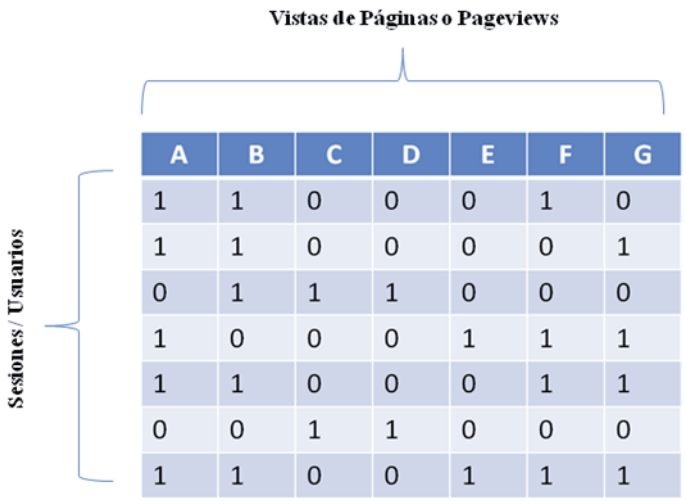
- A=index.html
- B= romano.html
- C=alertapucv.html
- D=guiadeservicios.html
- E=salademusica.html
- F=perfiles.html
- G=misión.html



A	B	C	D	E	F	G
15	5	0	0	0	30	15
10	3	0	0	0	0	5
0	4	5	10	0	0	0
5	0	0	0	5	10	4
3	4	0	0	0	6	8
0	0	8	3	0	0	0
4	6	0	0	3	3	2

Figura 4.6 User Pageview Matrix (UPM) con peso= t(s)

En este caso los pesos $w(p_i^t)$ corresponden al tiempo gastado (en segundos), sin embargo, en la mayoría de las tareas de Web Mining estos pesos suelen ser un valor binario: 0 (si la página no fue visitada) ó 1 (si la página fue visitada), de esta forma, utilizando esta matriz cómo entrada para los algoritmos que se utilizarán, como reglas de asociación por ejemplo, se podrá encontrar importantes relaciones a lo largo de los ítems basados en patrones de navegación de los usuarios en el sitio [20].



A	B	C	D	E	F	G
1	1	0	0	0	1	0
1	1	0	0	0	0	1
0	1	1	1	0	0	0
1	0	0	0	1	1	1
1	1	0	0	0	1	1
0	0	1	1	0	0	0
1	1	0	0	1	1	1

Figura 4.7 User Pageview Matrix (UPM) con peso= valor binario (0,1)

En base a éstos vectores, las reglas de asociación que se generen, darán a conocer las relaciones existentes entre las diferentes páginas y cuales a su vez son accedidas juntas, lo cual dará ideas de realizar posibles reestructuraciones en los contenidos de las páginas para así ayudar a que el usuario navegue más eficientemente por el sitio provocando una mayor satisfacción por parte de él al hacer uso de este servicio. A modo de ejemplo, se podrían obtener reglas de este tipo: el 50% de los visitantes que accedieron a *proyectos /estudiantiles/* y *proyectos/docentes/* también accedieron a *servicios/guiadeservicios/* con un 75% de confianza.

Para lo anterior, se utilizará el algoritmo A priori debido a las ventajas que se mencionaron en la sección 3.3.3.3.

4.5 Análisis y Validación de resultados

Una vez aplicado el diseño expuesto en este capítulo, será necesario validar los resultados obtenidos para comprobar que efectivamente las etapas realizadas durante el desarrollo de este estudio se hicieron correctamente y los resultados obtenidos son satisfactorios. Para esto, será necesario validar los resultados con la persona encargada de administrar el sitio web, debido a que es ella la persona más representativa en relación al manejo y contenido del sitio para realizar posibles ajustes al modelo. Luego, la idea será poder sugerir posibles soluciones a los problemas encontrados en conjunto con el administrador del sistema y así realizar mejoras a este sistema que es de uso de toda la comunidad universitaria.

5 Aplicación en el sitio web en estudio

5.1 Introducción

En este capítulo se detallará la implementación y aplicación del diseño presentado en el capítulo anterior en el ***Sistema de Biblioteca*** de la Pontificia Universidad Católica de Valparaíso. Para esto, se comenzará exponiendo sobre el pre procesamiento realizado a los weblogs del sitio y las estadísticas generales que se pudieron obtener a partir de ellos.

Posteriormente, se detallará los detalles del proceso de construcción del repositorio de extracción expuesto en la Figura 4.5 y el proceso ETL que se realizó en él. A partir de este, se realizarán consultas tipo OLAP para obtener información relevante contenida en los weblogs.

Finalmente, se realizará la aplicación del algoritmo A-priori y se analizará los resultados obtenidos para otorgar recomendaciones de mejoras en el sitio.

5.2 Pre procesamiento

En este estudio se consideraron los 6 meses más actuales que se encontraban disponibles en relación a los weblogs del servidor web donde se encuentra alojado el sitio. Así, los meses considerados en el estudio fueron los siguientes:

- Octubre (2009)
- Noviembre (2009)
- Diciembre (2009)
- Enero (2010)
- Febrero (2010)
- Marzo (2010)

La Figura 5.1 muestra la cantidad de registros realizada en cada uno de los meses en el cual se puede apreciar que el mes que más solicitudes se hicieron al Sistema de Biblioteca fueron los meses de octubre, noviembre y diciembre del año 2009, lo que se puede explicar debido a que ésos meses corresponden al período académico de la Universidad, mientras que los meses de enero y febrero pertenecen al período de vacaciones de la misma.

El total de solicitudes realizadas durante los 6 meses corresponde a 14.383.371, donde:

- En el mes de Octubre (2009) se realizaron 3.132.038 solicitudes que corresponden al 21,78% del total de entradas.
- En el mes de Noviembre (2009) se realizaron 5.442.279 solicitudes que corresponden al 37,84% del total de entradas.
- En el mes de Diciembre (2009) se realizaron 2.711.773 solicitudes que corresponden al 18,85% del total de entradas.

- En el mes de Enero (2010) se realizaron 1.213.052 solicitudes que corresponden al 8,43% del total de entradas.
- En el mes de Febrero (2010) se realizaron 640.602 solicitudes que corresponden al 4,45% del total de entradas.
- En el mes de Marzo (2010) se realizaron 1.243.628 solicitudes que corresponden al 8,65% del total de entradas.

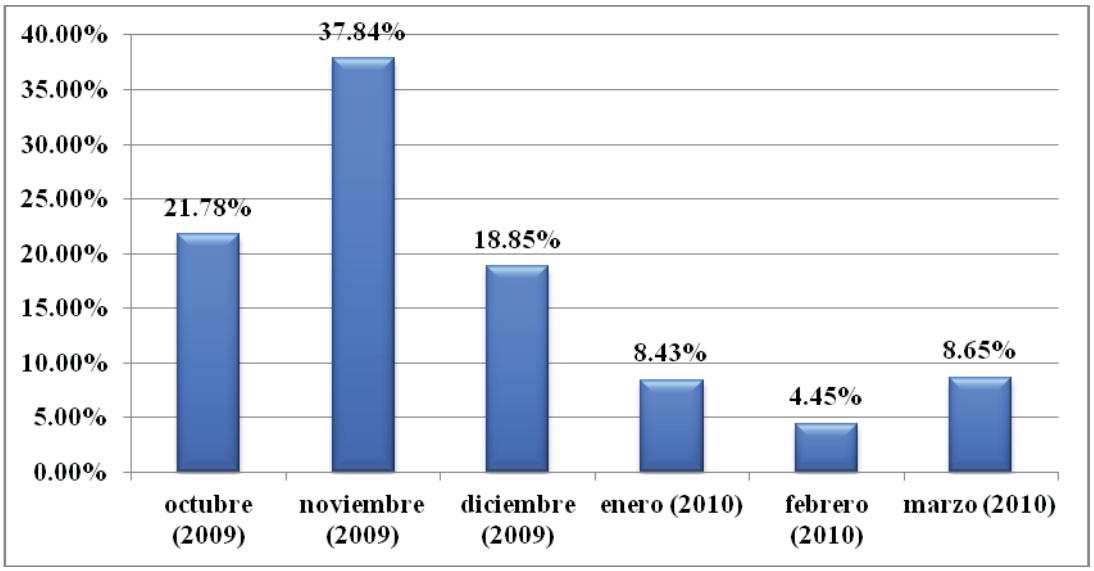


Figura 5.1 Entradas totales en el log por mes

En relación al número total de solicitudes realizadas al servidor, se puede observar que existen registros de distintos tipos que son almacenados en el servidor web al momento que se realiza una visita al sitio. Estos registros contienen:

- Imágenes (jpg,gif,ico,png)
- Códigos (javascript, css)
- Archivos (pdf, txt,xls,doc)
- Páginas (html,jsp,php)
- Entradas de *Web crawlers*
- Peticiones incorrectas al servidor (Sin URL por ejemplo)

De esta forma, fue necesario realizar una limpieza de los registros y dejar sólo aquellos que sean útiles para generar conocimientos. En la Figura 5.2 se pueda apreciar el porcentaje de elementos que tuvieron que ser removidos de los web logs, en donde se puede observar que la mayor parte de éstos elementos corresponden a peticiones de imágenes y archivos (correspondientes a un 86,95% del total de objetos inservibles), seguido por peticiones a códigos JS y CSS con un 11,47%, luego por peticiones que no contenían URL por lo que debían ser eliminadas también (1,2%), y por último con sólo un 0,39% del total de peticiones corresponden a entradas de *web crawlers*.

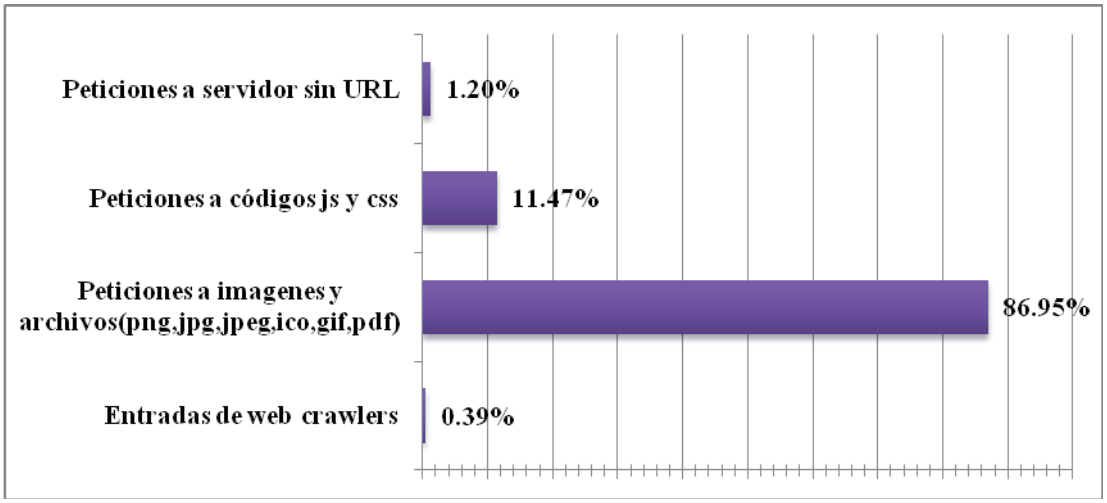


Figura 5.2 Porcentaje de objetos removidos en los web logs (total de objetos: 13.469.089)

Junto con esto, de acuerdo al estado de resolución de cada petición (campo *Status Code* o *Estado*), la gran mayoría de estas fue correctamente respondida. En la Tabla 5.1 se puede observar que el 99,2% de las entradas tuvo una resolución positiva (estados 200, 302 y 304).

Tabla 5.1 Web logs según *Status Code* o *Estado*

Código	Significado	#	% del Total
200	OK	5.692.611	39.578%
204	No Content	0	0.000%
206	Partial Content	34.735	0.241%
301	Moved Permanentely	3.144	0.022%
302	Found	3.303	0.023%
304	Not Modified	8.573.422	59.542%
400	Bad Request	143	0.001%
401	Unauthorized	175	0.001%
404	Not Found	75.706	0.526%
405	Method Not Allowed	132	0.001%
500	Internal Server Error	0	0.000%
		14.383.371	**100%**

A continuación se detalla el significado de cada uno de los *códigos de estado*:

- *200 - OK*: La petición fue exitosamente recibida, entendida y aceptada
- *204 - No Content* : La petición fue exitosamente respondida, pero no implicó enviar contenido al web browser del usuario
- *206 - Partial Content* : La petición ha sido parcialmente satisfecha. Por ejemplo: descarga no completada
- *301 - Moved Permanently* : El recurso solicitado tiene una nueva URL permanente, así que cualquier referencia a este en el futuro debería ser hecha a esta URL

- *302 - Found*: El recurso solicitado ha cambiado provisionalmente de localización. El servidor ofrece al cliente la nueva localización
- *304 - Not Modified* : El servidor entiende que existe una copia en caché, e informa de que el documento no ha cambiado: sigue siendo el mismo que el guardado en la memoria caché.
- *400 - Bad Request* : La sintaxis de la solicitud es incorrecta.
- *401 - Unautherized* : La petición necesita autorización para ser respondida.
- *404 - Not Found*: La sintaxis de la solicitud es correcta, pero el servidor no encuentra ningún recurso cuya URI se corresponda con la de la solicitud.
- *405 - Method Not Allowed* : La solicitud se ha hecho empleando un método (GET, POST, HEAD...) que no está permitido.
- *500 - Internal Server Error*: Se ha producido un fallo en el servidor, que no ha podido resolver la solicitud.

En conclusión, las entradas que se considerarán para cargar el Data Webhouse serán las que poseen las siguientes características:

- La entrada pertenece a un "usuario humano", es decir, no corresponde a una entrada de un *web crawler*
- Aquellas que solicitan objetos relevantes para el análisis
- Aquellas cuyo estado es 200,302 ó 304.

En base a esto, se obtiene un total de 142.008 entradas válidas para realizar el estudio. En otras palabras, se tiene que sólo un 1,095% de los datos totales son útiles para el análisis y extracción de conocimiento lo cual se explica principalmente por el alto número de solicitudes a objetos no relevantes.

5.3 Carga de Repositorio de Extracción

5.3.1 Proceso de Extracción

Este proceso corresponde al proceso de la selección de los web logs del servidor donde se encuentra alojado el Sistema de Biblioteca de la PUCV y la carga de los mismos en una tabla de una base de datos relacional, en este caso MySql, para realizar posteriormente la transformación y carga de éstos en el Data Webhouse.

La tabla que contiene los logs "limpios" fue llamada "Logs" la cual contiene cada registro válido (en el sentido que es útil para el estudio) que se encontraba en el archivo "biblioteca-ucv.log". Para realizar la extracción de los logs desde la fuente de origen y obtener cada uno de los campos de este, se utilizó la herramienta *Knime* los cuales se encuentran separados por un espacio y así posteriormente procesarlos utilizando Java y llevarlos a la tabla en MySql.

Figura 5.3 Extracción y pre-procesamiento de logs

Una vez procesados los logs, lo que se obtiene es una tabla relacional con los logs en forma adecuada para empezar a realizar el proceso de transformación y carga de ellos en el Data Webhouse. La Figura 5.4 muestra un fragmento de la tabla obtenida con cada una de sus columnas la cual contiene un total de 142.008 registros para realizar el estudio.

id	ip	timestamp	metodo	url	protocolo	estado	bytes	referrer	agente
1	201.252.154.116	2009-10-15 03:37:20	GET	/herramientas/citaselectronicas/iso690-2/is...	HTTP/...	200	1590...	-	Mozilla/5.0 (X11;.
2	190.232.204.127	2009-10-15 03:39:41	GET	/herramientas/citasbibliograficas/iso690/iso...	HTTP/...	304	0	http://www.google.com...	Mozilla/4.0 (com..
3	186.82.147.58	2009-10-15 03:56:23	GET	/catalogoderecho/Revistas/Rev_Derecho...	HTTP/...	304	0	http://www.google.com...	Mozilla/4.0 (com..
5	186.82.147.58	2009-10-15 03:56:33	GET	/catalogoderecho/Revistas/Rev_Derecho...	HTTP/...	200	3465	http://biblioteca.ucv.cl/...	Mozilla/4.0 (com..
6	66.249.67.111	2009-10-15 04:03:20	GET	/proyectos/archiveros/	HTTP/...	200	37518	-	SAMSUNG-SGH.
7	217.126.33.236	2009-10-15 04:03:23	GET	/herramientas/citasbibliograficas/iso690/iso...	HTTP/...	200	1118...	http://www.google.es/...	Mozilla/5.0 (Maci.
8	81.33.11.197	2009-10-15 04:09:15	GET	/herramientas/citasbibliograficas/iso690/iso...	HTTP/...	200	1118...	http://www.google.es/...	Mozilla/4.0 (com..
9	81.33.11.197	2009-10-15 04:10:05	GET	/herramientas/citaselectronicas/iso690-2/is...	HTTP/...	200	1590...	http://www.google.es/...	Mozilla/4.0 (com..
10	85.49.202.78	2009-10-15 04:22:37	GET	/e-bibliotecario/normas/estad%CDstica.asp	HTTP/...	200	21850	http://www.google.es/...	Mozilla/4.0 (com..
11	82.130.246.23	2009-10-15 04:46:34	GET	/poseidon/	HTTP/...	200	5971	http://www.google.es/...	Mozilla/4.0 (com..
12	82.130.246.23	2009-10-15 04:47:31	GET	/poseidon/index2.html	HTTP/...	200	7024	http://www.google.es/...	Mozilla/4.0 (com..
13	82.130.246.23	2009-10-15 04:53:51	GET	/poseidon/index2.html	HTTP/...	200	7024	http://www.google.es/...	Mozilla/4.0 (com..
14	85.55.136.78	2009-10-15 04:56:15	GET	/herramientas/citasbibliograficas/	HTTP/...	200	35538	http://www.google.com...	Mozilla/4.0 (com..
15	82.130.246.23	2009-10-15 04:58:13	GET	/poseidon/	HTTP/...	200	5971	http://www.google.es/...	Mozilla/4.0 (com..
16	80.34.155.134	2009-10-15 05:00:47	GET	/herramientas/citasbibliograficas/iso690/iso...	HTTP/...	200	1118...	http://odas.es/portal/re...	Mozilla/5.0 (Win..
17	150.214.20.199	2009-10-15 05:05:10	GET	/poseidon/index2.html	HTTP/...	200	7024	http://www.google.es/...	Mozilla/4.0 (com..

Figura 5.4 Extracto de logs de la tabla relacional “Logs”

5.3.2 Proceso de Transformación y Carga

A continuación se detalla la implementación del proceso de Transformación y Carga de los web data desde la tabla donde se encuentran alojados los logs hacia el Repositorio de Extracción. La Figura 5.5 muestra las etapas de las cual consta este proceso:

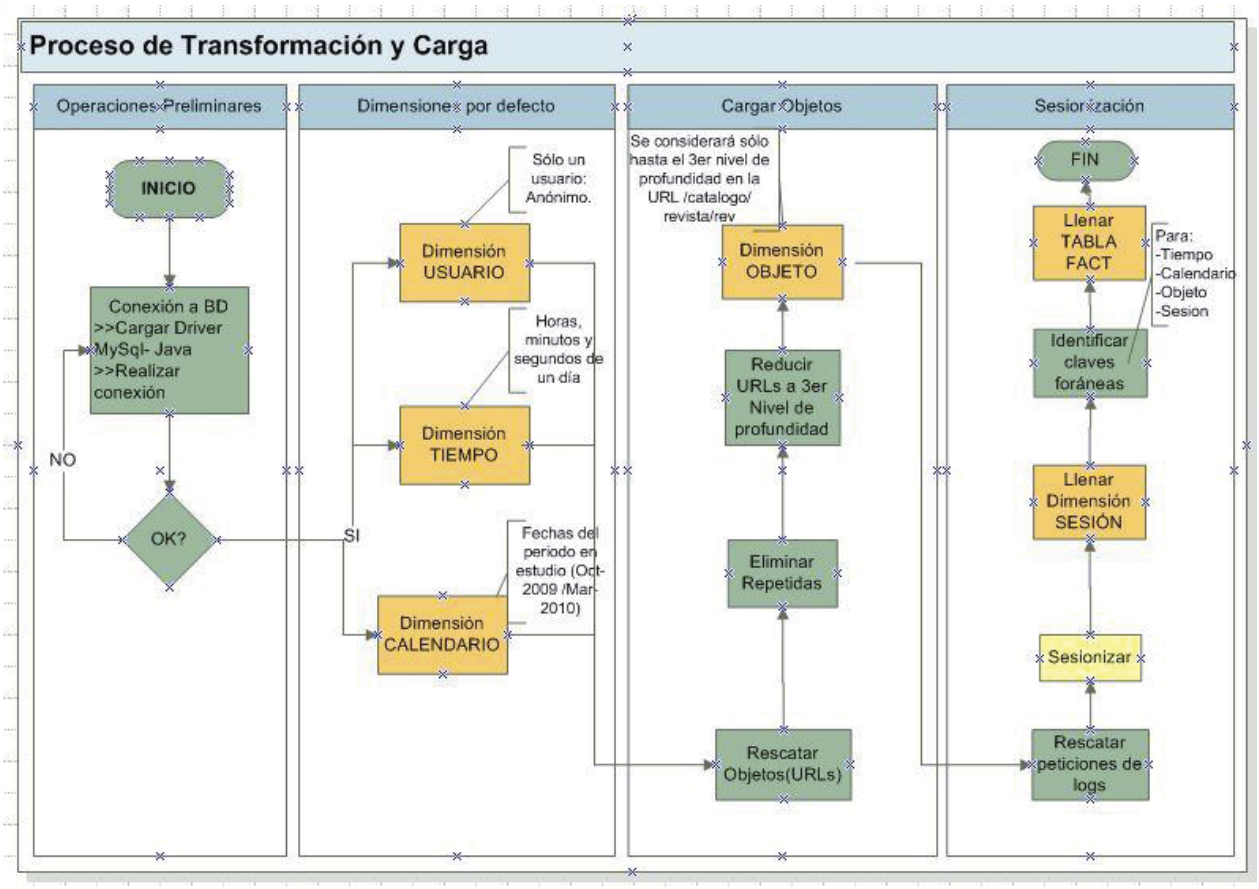


Figura 5.5 Proceso de Transformación y Carga

a) Operaciones preliminares: corresponde a las acciones que se deben realizar previamente antes del procesamiento y carga de los logs. En esta etapa se realiza principalmente la carga de drivers de MySql con Java y se realiza la conexión a la base de datos donde se alojarán los logs.
b) Llenado de dimensiones por defecto: en el repositorio de extracción existen 3 tablas que se deben llenar con valores por defecto debido a que no cambian sus valores, estas tablas son: *Usuario*, *Calendario* y *Tiempo*.
    - La dimensión *Usuario* contendrá sólo un registro, el cual será "Usuario Anónimo"

- La dimensión *Calendario* contendrá 182 registros correspondientes a las fechas del período de estudio que se está considerando, es decir, desde el 1 de octubre de 2009 hasta el 31 de marzo de 2010.
- La dimensión *Tiempo* tendrá un registro por cada segundo que compone un día cualquiera, es decir, 84.600 registros.

c) Carga de Objetos: para realizar la carga de los objetos, es decir, las URL´s que componen el sitio, se consideró sólo hasta el tercer nivel de profundidad debido al aumento explosivo del número de URL´s válidas que se podrían obtener al considerar cada una de ellas por separado. Por ejemplo, la dirección del siguiente objeto: */herramientas/citaselectronicas/iso690-2/iso690-2.html* será considerada sólo hasta */herramientas/citaselectronicas/iso690-2/*. De esta forma, se logró cargar esta dimensión con un total de 514 registros.

d) Sesionización: en esta etapa se realiza el proceso de sesionización a partir de la lectura de cada log contenido en la tabla "Logs". A medida que se va realizando la lectura de cada log, se va aplicando el algoritmo de sesionización que se encuentra implementado en Java, y así se va llenando la dimensión *Sesión* a medida que se cree una nueva sesión. Junto a esto, se va identificando las claves foráneas de cada dimensión para cargar el registro que se está leyendo en la *Tabla Fact Petición.* Finalmente se obtuvo un total de 69.090 sesiones y la tabla petición fue llenada con un total de 142.008 registros. La Figura 5.6 muestra el flujo seguido para realizar el proceso de sesionización.

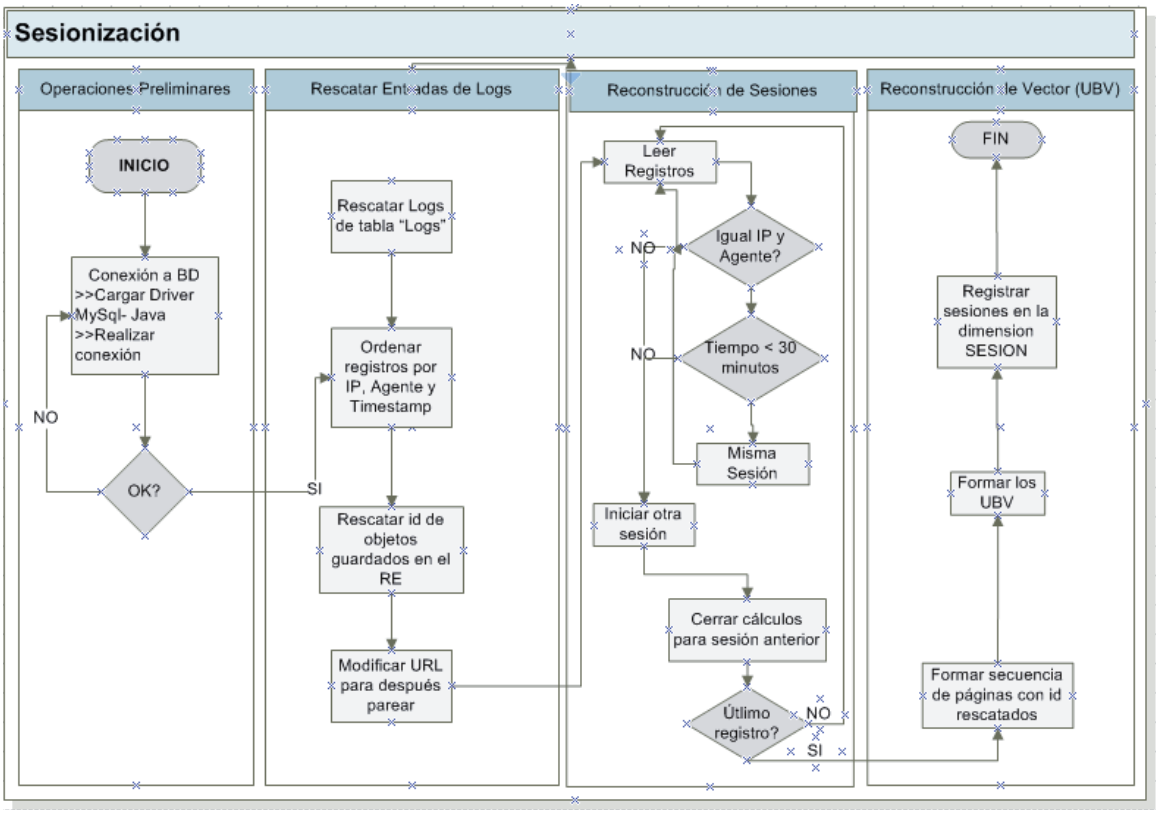


Figura 5.6 Flujo del programa Sesionizador

5.4 Descubrimiento de Patrones

5.4.1 Análisis OLAP

Gracias a la estructura que posee el *Repositorio de Extracción*, modelo tipo estrella, es posible extraer información de manera sencilla utilizando herramientas tipo OLAP. En este caso, se utilizará un servidor OLAP de software libre ***Mondrian***, el cual permitirá obtener información del *Data Webhouse* a partir de los datos registrados en este.

A. Considerando la dimensión *Calendario* del Repositorio de Extracción, se puede realizar un análisis respecto a la cantidad de solicitudes realizadas en las distintas fechas que se consideró en este estudio. La Figura 5.7 indica la cantidad de solicitudes realizadas en los distintos períodos de tiempo.

	Measures
Calendario	Cantidad Solicitudes
All Calendarios	142,008
2009	89,928
December	20,717
November	46,240
October	22,971
2010	52,080
February	17,566
January	16,934
March	17,580

Figura 5.7 Consulta OLAP: Peticiones según dimensión CALENDARIO

Observando los resultados obtenidos de esta consulta OLAP, se puede apreciar que durante los últimos meses del año 2009, el Sistema de Biblioteca posee una cantidad de acceso superior a la de los primeros meses del año 2010. Esto se puede explicar debido a que en los meses de Octubre, Noviembre y Diciembre, pertenecen al período académico de los estudiantes de la Universidad, y por lo tanto, el número de accesos es considerablemente mayor al de los primeros meses (Enero, Febrero y Marzo) del año 2010. Así, también se puede apreciar que el mes de Noviembre fue el mes que más visitas recibió el sistema seguido por el mes de Octubre. Esta herramienta, además permite generar gráficos a partir de los resultados obtenidos. La Figura 5.8 muestra el gráfico asociado a la figura anteriormente descrita.

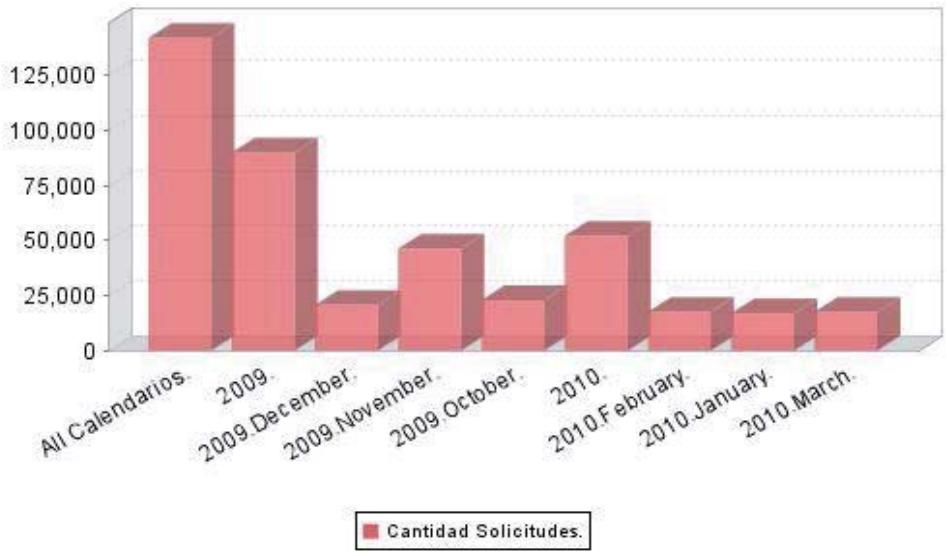


Figura 5.8 Consulta OLAP: gráfico de Peticiones según dimensión CALENDARIO

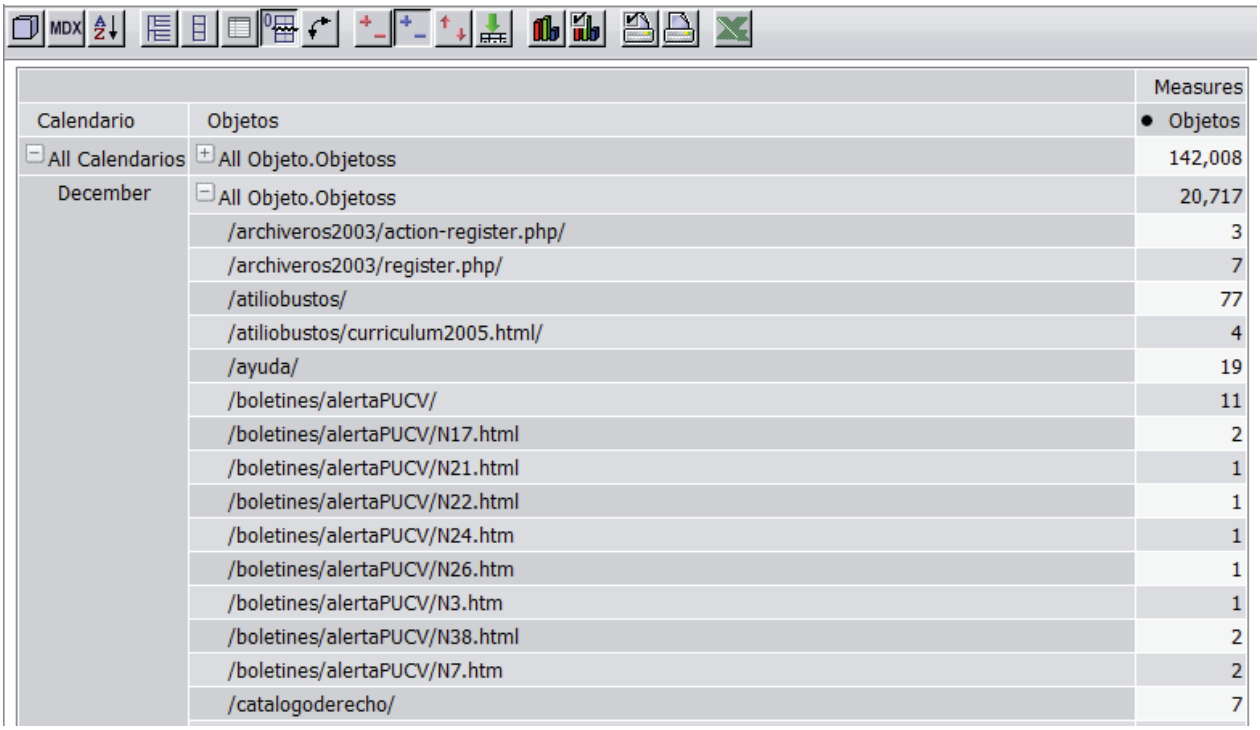
Si se sigue expandiendo los contenidos obtenidos a partir de la consulta anterior, se puede observar, además, las peticiones realizadas en cada día de los meses en estudio. La Figura 5.9 muestra las peticiones realizadas durante el mes de Noviembre los distintos días de la semana.

	Measures
Calendario	● Cantidad Solicitudes
⊟ All Calendarios	142,008
⊟ 2009	89,928
⊞ December	20,717
⊟ November	46,240
Friday	6,131
Monday	10,355
Saturday	2,920
Sunday	4,153
Thursday	7,058
Tuesday	8,146
Wednesday	7,477
⊞ October	22,971
⊞ 2010	52,080

Figura 5.9 Consulta OLAP: Peticiones en días según dimensión CALENDARIO

En base a la información contenida en la consulta anterior, se puede apreciar que los días que más visitas se realizó en el Sistema de Biblioteca durante el mes de Noviembre fueron los días Lunes y Martes con 10.355 y 8.146 visitas respectivamente.
De esta forma se puede concluir a partir de esta información, que el Sistema de Biblioteca tuvo su mayor actividad durante el mes de Noviembre del 2009, específicamente los días Lunes y Martes.

B. El otro cruce de dimensiones realizadas fue respecto a la dimensión *Calendario* y la dimensión *Objeto*, en donde se puede apreciar la cantidad de solicitudes a los distintos objetos realizadas durante los distintos meses considerados en el estudio.



		Measures
Calendario	Objetos	Objetos
All Calendarios	All Objeto.Objetoss	142,008
December	All Objeto.Objetoss	20,717
	/archiveros2003/action-register.php/	3
	/archiveros2003/register.php/	7
	/atiliobustos/	77
	/atiliobustos/curriculum2005.html/	4
	/ayuda/	19
	/boletines/alertaPUCV/	11
	/boletines/alertaPUCV/N17.html	2
	/boletines/alertaPUCV/N21.html	1
	/boletines/alertaPUCV/N22.html	1
	/boletines/alertaPUCV/N24.htm	1
	/boletines/alertaPUCV/N26.htm	1
	/boletines/alertaPUCV/N3.htm	1
	/boletines/alertaPUCV/N38.html	2
	/boletines/alertaPUCV/N7.htm	2
	/catalogoderecho/	7

Figura 5.10 Consulta OLAP: Peticiones a objetos de acuerdo a dimensión OBJETO y CALENDARIO

En la Figura 5.10 se puede apreciar los distintos objetos solicitados durante el mes de Diciembre del año 2009. En este fragmento, se puede observar que el objeto que más visitas tuvo fue */atiliobustos/* seguido de */ayuda/* lo que indica el interés que poseen los visitantes del Sistema durante este periodo de tiempo.
A modo de resumen, se presentan los objetos con más solicitudes realizadas en el sistema en la Tabla 5.2:

Tabla 5.2 Objetos con mayor cantidad de solicitudes

Objeto	Cantidad Solic.
/cubiculos/sistema/cubiculo.php	15,051
/opac/index.html	13,169
/opac/top.html	13,107
/herramientas/citasbibliograficas/iso690/iso690.htm	12,000
/poseidon/index2.html	6,515
/herramientas/citaselectronicas/iso690-2/iso690-2.html	5,930
/herramientas/citasbibliograficas/	5,913
/e-bibliotecario/referencia/diccionariosytraductores.asp	3,776
/search/	1,587
/guidelines/guidelines.htm	1,271
/recursos/beic/	1,143
/recursos/derecho/index.php	1,130
/poseidon/libros/libro1/60.html	1,117

5.4.2 Análisis de clickstreams

En base al proceso de construcción de sesiones, es posible estudiar en más detalle estas secuencias de páginas o *clickstreams* para poder obtener información de gran importancia y que se encuentra implícita en ellas. Así, los siguientes puntos muestran la información que se pudo extraer a partir de ellas:

A. En la Tabla 5.3 se puede apreciar los puntos de salida más comunes en las sesiones de los usuarios, es decir, las páginas donde la mayor parte de los usuarios abandonan el sistema cuando navegan. Esto entrega ideas al administrador y/o diseñador del sitio de poder observar las posibles causas y poder optar por posibles soluciones al problema de tal forma de que el usuario no abandone el sistema una vez que entra a dicha página.

Tabla 5.3 Puntos de salidas más comunes durante la navegación en el sistema

Cod. Pág.	URL	Cantidad
2	/herramientas/citasbibliograficas/iso690/iso690.htm	2719
12	/opac/index.html	2282
1	/herramientas/citaselectronicas/iso690-2/iso690-2.html	1597
43	/search/	535
9	/herramientas/citasbibliograficas/	512
8	/poseidon/index2.html	325
149	/poseidon/libros/libro3/40.html	235
47	/recursos/beic/	222
31	/recursos/derecho/index.php	200
155	/poseidon/libros/libro3/presentacion.html	184
129	/poseidon/libros/libro2/20.html	174
42	/servicios/multibuscador/texto.html	170
110	/poseidon/libros/libro2/50.html	154
57	/elsistema/bibliotecas/mayorfilosofiayeducacion/	138
21	/poseidon/libros/libro2/30.html	136
74	/poseidon/libros/libro1/60.html	126
40	/e-bibliotecario/referencia/diccionariosytraductores.asp	125
176	/poseidon/libros/libro1/30.html	123
114	/herramientas/citaselectronicas/	123
193	/e-bibliotecario/index.html	120
154	/poseidon/libros/libro3/indice2.html	113
109	/poseidon/libros/libro2/i5.html	112
28	/catalogoderecho/CATALOGO.HTM	108
150	/e-bibliotecario/biblioteca_virtual/bibliotecas_digitales.asp	102

En la Tabla 5.3 la columna "*Cód. Pág.*" se refiere al código de la página, "*URL*" a la dirección url de la misma y "*Cantidad*" a la cantidad de sesiones en las que el usuario abandonó el sistema en dicha página. A partir de esto, se puede observar que existe un número considerable de sesiones en que los usuarios abandonan el sistema en las primeras tres páginas por lo que es muy probable que exista algún factor predominante que explique la causa de esto.

B. Respecto a los *referrers* más comunes desde donde el Sistema de Biblioteca es visitado se muestran en la Figura 5.11.

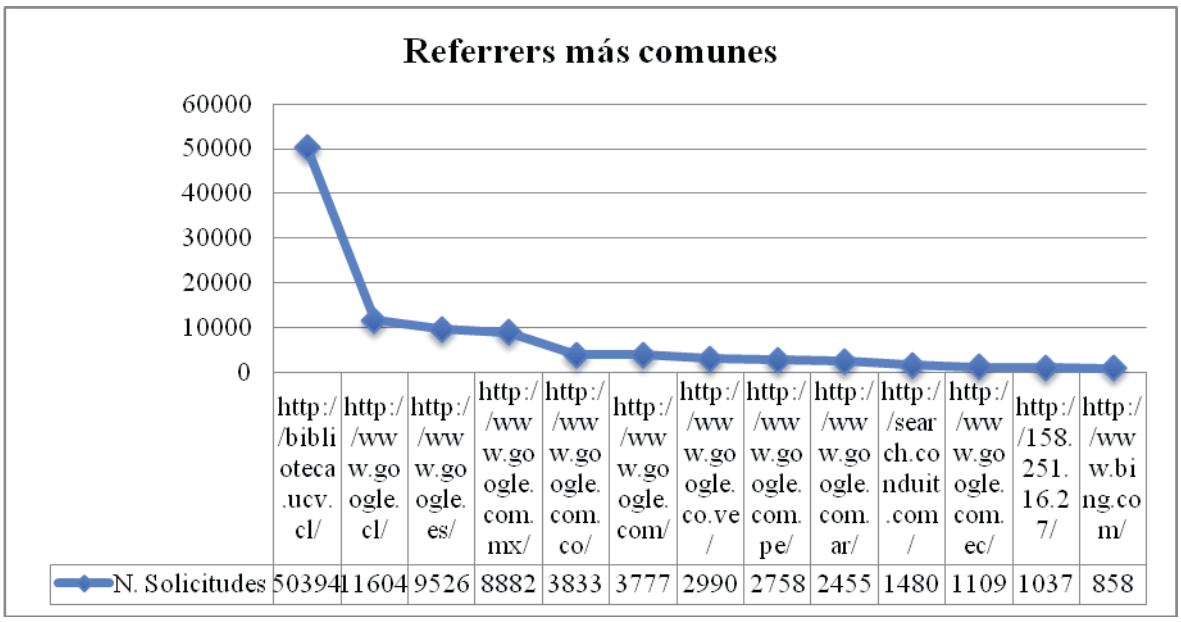


Figura 5.11 Páginas desde donde el Sistema de Biblioteca PUCV es accedido

Se puede observar que la mayor parte de las solicitudes realizadas al sistema provienen desde el mismo Sistema de Biblioteca con 50.394 solicitudes realizadas, seguida por *Google (.cl, .es, etc.)*, lo cual refleja que también gran parte de la información que contiene el sistema es accedida por medio de la ayuda de buscadores externos al sitio.

C. En relación a la cantidad de páginas que son visitadas en las diferentes sesiones de los usuarios, se puede apreciar que en promedio los usuarios acceden a 2 páginas en el sistema antes de abandonarlo, aunque considerando un indicador estadístico más representativo, como la mediana por ejemplo, se obtiene que los usuarios acceden sólo a 1 sitio y luego lo abandonan.

D. Considerando el tiempo de navegación de los usuarios, se obtuvo un tiempo promedio de **250** segundos por sesión correspondientes a 4,2 minutos aproximadamente, sin embargo considerando como indicador estadístico la mediana, se obtuvo sólo **46** segundos de navegación por sesión.

5.4.3 Reglas de Asociación y Algoritmo Apriori

A continuación se presenta el formato en que se encuentran almacenados los accesos a las páginas en las distintas sesiones y que deberán ser transformadas para crear los vectores de comportamiento de los usuarios los cuales serán utilizados para realizar un estudio de las distintas preferencias que poseen mediante reglas de asociación utilizando el algoritmo Apriori. En la Figura 5.12 se puede observar el ID de la sesión junto con la secuencia de páginas visitadas en ella. Así, se puede observar por ejemplo, que en la sesión 30.992 se accedió al objeto con ID=413 y posteriormente al objeto con ID=411.

idSesion	secuencia
30985	150 150 150 150
30986	150
30987	411
30988	411
30989	522
30990	409
30991	150 150
30992	413 411
30993	411
30994	411 411
30995	472
30996	411
30997	409 411
30998	36 58
30999	42 51 50 58 42
31000	163
31001	215 210
31002	215 210
31003	215 210
31004	215 210
31005	55 550 570 575 575 579 573 580 577 575 575 575 575
31006	215 210
31007	215 210
31008	210 215
31009	592 58 557 210 215
31010	215 210 215 210 210 215
31011	215 210 215 210
31012	215
31013	215 210

Figura 5.12 Extracto de vectores de comportamiento del usuario (UBV)

Para crear los UBV fue necesario tomar las secuencias de páginas contenidas en cada sesión y llevarlas a un formato transaccional para posteriormente realizar la conversión a la matriz que representa los UBV y que será la entrada para A priori. Así, utilizando Java se tomó cada secuencia de páginas y se separaron cada código de ellas con su respectivo ID de sesión quedando de la siguiente forma:

Tabla 5.4 Formato transaccional

IdSesión	CodPag
10001	28
10001	45
10010	50
10011	46
10012	50
10014	67
10018	1

A partir del archivo creado con este formato se procedió a formar la UPM (*user pageview matrix*) utilizando distintos nodos que contiene *Knime*. En la Figura 5.13 se observan los nodos y sus conexiones.

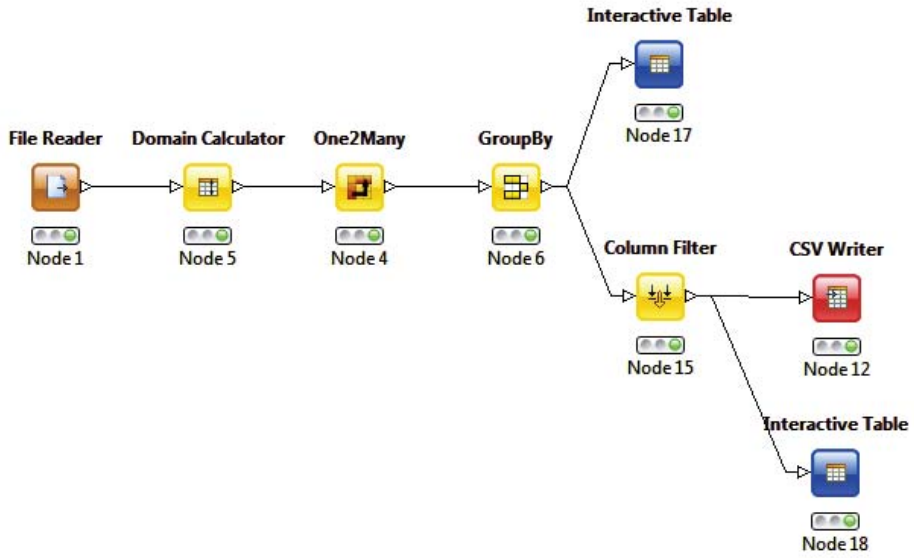


Figura 5.13 Utilización de *Knime* para generar la UPM

Lo que se obtuvo fue una matriz de la cantidad de objetos existentes en la dimensión *objeto* del repositorio de extracción por la cantidad de sesiones y que contiene 0 y 1 de acuerdo a si el objeto se encuentra en la transacción o no. La Figura 5.14 muestra una porción de dicha matriz.

tran	Max(1)	Max(15)	Max(30)	Max(16)	Max(28)
10001	0	0	0	0	1
10010	0	0	0	0	0
10011	0	0	0	0	0
10012	0	0	0	0	0
10014	0	0	0	0	0
10018	1	0	0	0	0
10020	0	0	0	0	0
10026	0	0	1	0	0
10032	0	0	0	0	0
10033	0	0	0	0	0
10036	0	0	0	0	0
10038	1	0	0	0	0
10042	1	0	0	0	0
10048	0	0	0	0	0
10053	0	0	1	0	0
10059	0	0	0	0	0
10068	0	0	0	0	0
10073	0	0	0	0	0
10079	0	0	0	0	0
10084	0	0	0	0	0
10086	1	0	0	0	0
10087	0	1	1	0	1
10088	0	0	0	0	0
10091	0	0	0	0	0
10092	0	0	0	0	0

Figura 5.14 Extracto de *User Pageview Matrix*

Para aplicar el algoritmo A priori se utilizó la herramienta *Clementine* en donde se utilizaron distintos nodos para cargar la matriz y adecuar los datos contenidos en ella para realizar la correcta aplicación del algoritmo. En primera instancia se obtuvieron muy pocas reglas y de poca relevancia, incluso disminuyendo los parámetros de soporte y confianza por lo que se procedió a aumentar aún más el nivel de abstracción de las url´s para verificar si de esta forma se conseguía un número mayor de reglas y de más relevancia. Así, se consideró objetos de la forma: /e-bibliotecario/, /herramientas/, /proyectos/, etc. reduciendo considerablemente el número de objetos y por lo tanto el tamaño inicial de la matriz.

El modelo utilizado en *Clementine* para lograr utilizar Apriori se presenta en la Figura 5.15.

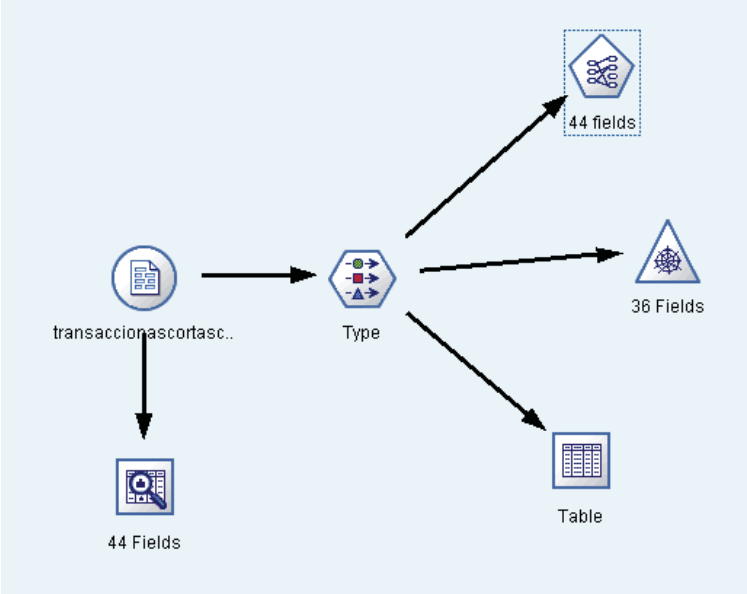


Figura 5.15 Aplicación de algoritmo Apriori en *Clementine*

Nuevamente, considerando valores altos de confianza y soporte (superiores a 70%) no se encontraron reglas de asociación, por lo tanto se puede empezar a deducir que existen pocas relaciones entre los accesos de las distintas sesiones. Sin embargo, disminuyendo los niveles de soporte y confianza al mínimo (5%) como se muestra en la Figura 5.16 se lograron obtener 13 reglas que se indican en la Tabla 5.5:

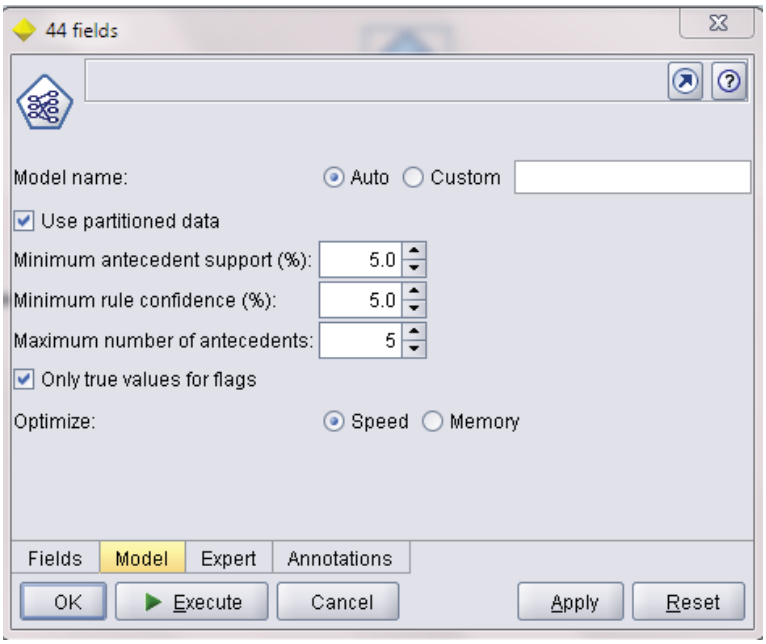


Figura 5.16 Parámetros de Apriori en Clementine

Tabla 5.5 Reglas obtenidas mediante algoritmo Apriori

Antecedente	Consecuente	Soporte	Confianza
/elsistema/	/opac/	7,80%	47,92 %
/recursos/	/opac/	5,10%	46,23%
/recursos/	/search/	5,10%	20,03%
/elsistema/	/search/	7,80%	18,80%
/elsistema/	/servicios/	7,80%	14,20%
/elsistema/	/e-bibliotecario/	7,80%	12,70%
/recursos/	/e-bibliotecario/	5,10%	8,09%
/opac/	/elsistema/	46,3%	8,10%
/elsistema/	/catalogos/	7,80%	7,93%
/recursos/	/elsistema/	5,10%	6,51%
/catalogoderecho/	/search/	5,08%	5,87%
/recursos/	/servicios/	5,10%	5,87%
/opac/	/recursos/	46,3%	5,12%

A partir de las reglas presentadas anteriormente, se puede deducir por ejemplo que:

- Las reglas cuyo consecuente es la página */opac/* son aquellas con mayores niveles de confianza. Por ejemplo, si se ha visitado la página */elsistema/* ó */recursos/* existe alrededor de un 47% de probabilidad que también se haya visitado la página */opac/* en un 7,8% y 5,10% de las sesiones respectivamente.
- Si se ha visitado la página */recursos/* existe alrededor de un 20% de probabilidad que también se haya visitado la página */search/* en un 5,10% de las sesiones.
- Si se ha visitado la página */elsistema/* existe un 18,8% de probabilidad que también se haya visitado la página */search/* en un 7,8% de las sesiones.
- Si se ha visitado la página */elsistema/* existe un 14,2% de probabilidad que también se haya visitado la página */servicios/* en un 7,8% de las sesiones.
- Si se ha visitado la página */elsistema/* existe un 12,7% de probabilidad que también se haya visitado la página */e-bibliotecario/* en un 7,8% de las sesiones.
- Si se ha visitado la página */recursos/* existe un 8,09% de probabilidad que también se haya visitado la página */e-bibliotecario/* en un 5,10% de las sesiones.
- Si se ha visitado la página */opac/* existe un 8,10% de probabilidad que también se haya visitado la página */elsistema /* en un 8,10% de las sesiones.

- Si se ha visitado la página */elsistema/* existe un 7,93% de probabilidad que también se haya visitado la página */catalogos /* en un 7,93% de las sesiones.
- Si se ha visitado la página */recursos/* existe un 6,51% de probabilidad que también se haya visitado la página */elsistema /* en un 5,10% de las sesiones.
- Si se ha visitado la página / catalogoderecho / existe un 5,87% de probabilidad que también se haya visitado la página / search / en un 5,08% de las sesiones.
- Si se ha visitado la página / *recursos* / existe un 5,87% de probabilidad que también se haya visitado la página / servicios / en un 5,10% de las sesiones.
- Si se ha visitado la página / opac / existe un 5,12% de probabilidad que también se haya visitado la página / servicios / en un 46,3% de las sesiones.

Si bien es cierto que las reglas obtenidas no poseen grados de soporte y confianza muy altos, éstos de igual forma pueden ser validados con los administradores del sistema y generar algún conocimiento en cuanto a la forma en que acceden los usuarios a las distintas secciones del sistema.

6 Análisis y Validación de Resultados

A continuación, se presenta el análisis de los resultados obtenidos en la sección anterior y la correspondiente validación con el administrador del Sistema de Biblioteca PUCV.

De modo general se puede deducir que a pesar de que se incluyeron 6 meses en el estudio, existe una cantidad reducida de visitas al sistema considerando que este se encuentra disponible para toda la comunidad universitaria en todo momento. En relación a esto, se podría pensar que:

- El sistema no contiene información relevante para la comunidad universitaria (estudiantes principalmente) ó bien, no se ha dado a conocer de manera masiva las funcionalidades y/o ventajas que posee este sistema.
- El uso que se da al sistema es para cosas puntuales, como por ejemplo, buscar un determinado libro, y no se accede ó probablemente no posee otras funcionalidades que resulten de mayor utilidad para el usuario
- Resulta dificultoso navegar por el sistema y/o encontrar las secciones deseadas por los usuarios por lo que provoca un rechazo por parte de ellos en el uso del sistema
- Posee un diseño poco atractivo para los usuarios por lo que también podría ser motivo de rechazo

Respecto a la validación de los cuatro puntos nombrados anteriormente se ha podido comprobar que efectivamente no existe una utilización masiva del Sistema de Biblioteca PUCV y que además, este sitio está pensado en ser una especie de directorio en donde los usuarios puedan acceder a lo que buscan en la página principal del sistema más que un sitio donde los usuarios puedan navegar por todas sus secciones en búsqueda de determinados contenidos, y por lo mismo, el usuario lo que realiza principalmente es a lo más un par de clicks y luego abandona el sitio, ya sea por búsquedas de libros ó bien búsquedas de links externos al Sistema de Biblioteca PUCV.

En relación a los períodos en estudio donde existieron más accesos fue en los meses de octubre (2009), noviembre (2009), y diciembre (2009) lo cual se explica principalmente porque corresponden a los períodos académicos intensos de la Universidad. En base a esto, se puede apreciar que el servidor ha respondido correctamente casi a la totalidad de las solicitudes realizadas por los usuarios pese a que existió un porcentaje mínimo de solicitudes que fueron efectuadas y que se obtuvo como respuesta un error 404 (*Not Found*), error que claramente genera un descontento considerable por parte de los usuarios al querer solicitar cierto contenido y este simplemente no es encontrado en el servidor. Este punto fue tratado con el administrador y se llegó a la conclusión que existen links que se encuentran rotos debido a que éstos han sido re direccionados a otra URL, sin embargo, los motores de búsquedas como Google, al no tener la posibilidad de acceder a éstos links para actualizar sus índices, las personas siguen accediendo por medio de este a los links rotos encontrándose como respuesta del sistema un error 404. Esto a su vez, empieza a aumentar el ranking que posee la página en los buscadores, y por ende, empieza a aparecer en las primeras posiciones de búsquedas y aumenta el acceso de los usuarios a éstos links rotos produciendo una mayor cantidad de registros en los logs que resultan inservibles para un estudio de este tipo. Otro factor que influye en éstos estados, es que los

browsers de los usuarios mantienen en el caché de este las páginas que se encuentran rotas y por lo tanto, los usuarios siguen accediendo a él en medida que el *browser* los sugiere.

Otro aspecto que fue observado, es que existía un reducido número de registros de *web crawlers* pese a que en la literatura se señala que corresponden a uno de los principales agentes en dejar una cantidad considerable de registros en los logs de los servidores al momento en que éstos refrescan sus índices, lo cual se explica debido a que se encuentra bloqueado el acceso de los *web crawlers* a gran parte del sitio en el archivo *robots.txt* existente en el directorio principal donde se encuentra alojado el sistema. Esto fue corroborado con el administrador, y efectivamente ellos sólo permiten el acceso de *web crawlers* a páginas específicas debido principalmente a que sus visitas aumentan considerablemente el tamaño de los logs.

En base a las consultas OLAP realizadas se pudo apreciar que existe una tendencia a aumentar los accesos al sistema en los días lunes y martes durante el periodo de estudio que se consideró, además se pudo apreciar que en general existen pocos accesos a cada objeto existente en el sistema aunque en algunos específicos se ve una clara tendencia a acceder más, y por ende podría ser preferible considerar éstos objetos y hacerlos más visibles al momento de ingresar en el sistema para facilitar y agilizar el acceso a los mismos. Efectivamente, los objetos que más son accedidos son los que caracterizan al Sistema de Biblioteca, que en este caso son la búsqueda de libros, la entrega y/o reserva de cubículos y otra que resulta más particular es la página que contiene la norma ISO-690, en donde consultando con el administrador, se pudo concluir que al ser la única universidad que contiene esta norma traducida al español al menos a nivel nacional, se genera una gran cantidad de visitas a esta por usuarios internos y externos a la universidad, incluso desde fuera del país puesto que al realizar la búsqueda en Google es una de las primeras 10 direcciones en aparecer en el buscador.

Considerando los accesos realizados en las distintas sesiones de usuarios, se pudo observar que existen puntos comunes en los que los distintos usuarios abandonan el sistema, lo cual puede dar a entender posibles problemas que puedan existir en dichas páginas, de hecho si se accede a la sección */herramientas/citasbibliograficas/iso690/iso690.htm* la cual corresponde al punto de quiebre más común en las distintas sesiones, se podrá apreciar lo presentado en la Figura 6.1:

Figura 6.1 Sección *Norma Internacional ISO-690 del Sistema de Biblioteca PUCV*

Como se puede observar, se despliega un contenido que pareciera estar completamente fuera del Sistema de Biblioteca ya que se pierde absolutamente toda referencia a este y no existe una manera explícita y rápida de poder regresar al inicio del sistema a menos que el usuario dé click en el botón "Atrás" ó "Volver a la página anterior" del *browser* algo que no resulta muy natural considerando que se está navegando dentro del sitio y por ende los usuarios tienden a cerrar la página y abandonar el sistema. Algo similar ocurre cuando se accede a */herramientas/citaselectronicas/iso690-2/iso690-2.html* el cual también es uno de los puntos principales de términos de sesión. Consultando al administrador por la lista de puntos comunes de términos de sesión, se pudo comprobar que los distintos links que se encontraron corresponden precisamente a secciones que el usuario normalmente busca y por lo tanto, después de acceder a éstas, ellos abandonan el sistema.

Con respecto a los *referrers* más comunes, se puede apreciar que la mayor parte de ellos proviene desde el mismo Sistema de Biblioteca, aunque si se considera en conjunto el resto de los *referrers* se puede observar que realmente la mayor parte de las solicitudes realizadas a las secciones u objetos del sitio corresponden a buscadores externos al sitio lo que podría dar indicios de un problema en las búsquedas que se realizan dentro del mismo sistema. Esto fue discutido con el administrador, y dio a entender que el problema principal de esto, es que existen muchos usuarios que no acceden directamente al Sistema de Biblioteca por medio de la URL en la barra de direcciones del navegador, sino que acceden indirectamente tipiando la URL en el buscador (en este caso específico Google) y por lo tanto, tiene a existir una mayor cantidad de *referrers* relacionados a buscadores. Respecto al buscador propio que contiene el sistema, este posee serias falencias las cuales fueron admitidas por el administrador, ya que realmente no entrega resultados relevantes al utilizarlo y también porque no ha existido un trabajo mayor para mejorarlo debido a que no han considerado que sea de gran utilidad a los usuarios que navegan por el sistema debido al objetivo mismo que este persigue.

Considerando el largo de las sesiones, o más bien, la cantidad de páginas que los usuarios acceden al sistema antes de salirse del mismo, se puede apreciar que en promedio acceden a 2 páginas, sin embargo, considerando otro indicador estadístico como la mediana, se observa que los usuarios acceden a sólo 1 sitio y luego abandonan el sistema. Éstos son resultados no muy alentadores respecto al uso que se le da al sistema, pues los usuarios prácticamente lo que realizan es entrar a una página y posiblemente visitan otra, para posteriormente abandonar el sistema. Si se considera el tiempo de duración de las sesiones, se podrá observar que en promedio los usuarios navegan 250 segundos, es decir, 4,2 minutos aproximadamente en el sistema, aunque considerando la mediana como el otro indicador estadístico, se obtuvo sólo 46 segundos por sesión lo cual indica que la navegación por el sistema es muy baja en comparación con los valores entregados por la media. Pese a que éstos resultados no resultan muy alentadores respecto al uso que se le da al sistema, sucede que es precisamente el objetivo que posee este Sistema de Biblioteca, ya que según lo aclarado por el administrador y la sub directora de este sitio, el objetivo principal es que los usuarios accedan a él y encuentren rápidamente lo que buscan sin necesidad de navegar por las otras secciones del sitio, y es por lo mismo que los principales contenidos del sistema se encuentran en la página de inicio del mismo. Así, una real preocupación de los encargados del sistema sería que los usuarios pasaran mucho tiempo navegando por el sistema ya que esto indicaría que ellos no logran encontrar lo que buscan.

Y por último, respecto a las relaciones encontradas mediante el algoritmo A priori, las reglas fueron validadas en conjunto con el administrador, señalando que eran completamente consistentes con la navegación que normalmente realizan los usuarios y que los niveles de soporte y confianza resultan ser bajos debido a la escasa navegación que realizan en sí los usuarios, por lo tanto las relaciones que se pueden extraer a partir de esto son relativamente bajas aunque interesantes de analizarlas.

Con este análisis se ha logrado dar un estudio más profundo sobre la forma en que los usuarios se comportan en el sistema contrastándolo con la visión que posee las personas que se encuentran directamente relacionadas con la administración del sistema, por lo tanto, se pudo validar cada resultado obtenido los cuales comprueban que se cumple en gran medida uno de los principales objetivos que posee el sistema como tal, el cual es que los usuarios accedan rápidamente a lo que buscan y no naveguen más de lo necesario por el sistema.

7 Recomendaciones para mejorar el sitio web

En base a lo estudiado y analizado, es posible sugerir algunas mejoras que se podrían realizar en el sistema:

- Sería interesante realizar una encuesta en la comunidad universitaria sobre la percepción que poseen del sistema en cuanto a diseño y si realmente se cumple el objetivo que persigue, y en base a esto, tomar decisiones sobre posibles cambios, reestructuraciones o mejoras en general al sistema con el objetivo de satisfacer en mayor medida las necesidades de los usuarios.
- Mejorar el buscador interno que posee el sistema para dar la opción al usuario de utilizarlo en caso de necesitarlo y así no recurrir a buscadores externos para realizar este tipo de tareas.
- Sería adecuado, quizás, habilitar cada cierto tiempo el acceso a los *web crawlers* a las secciones que se encuentran bloqueados para que ingresen al sistema y actualicen los índices que poseen y que hacen referencia a elementos que ya no se encuentran disponibles o fueron re direccionados a otra sección. De esta forma, se evitaría gran porcentaje de los errores 404 encontrados en este estudio y así se reduciría los registros creados en el log que no aportan mayores conocimientos y se mejoraría la percepción del usuario hacia el sistema.
- Respecto a los documentos que posee el sistema, como la norma ISO-690 por ejemplo, sería adecuado manipularlos como archivos descargables y no como texto plano como se muestra en la Figura 6.1. de tal forma de facilitar al usuario el acceso a este tipo de documentos.
- Sería adecuado considerar en el Sistema de Biblioteca sólo aquellos contenidos que tienen directa relación con la misma, debido a que existen secciones a las cuales se accede y lo que se encuentra es un sistema completamente aparte al de Biblioteca y por ende, se generan logs que no pertenecen precisamente a la navegación que realizan los usuarios por ella y puede distorsionar un poco este tipo de estudios al considerar los mismos como parte de este sistema.

Estas son algunas recomendaciones que se podrían tomar en cuenta para poder mejorar ciertos aspectos del Sistema de Biblioteca PUCV y así realizar mejoras en un servicio que es utilizado por toda la comunidad universitaria e incluso por gente externa a ella.

8 Conclusiones

A partir de lo expuesto en este proyecto, se puede concluir que el estudio se ha llevado a cabo satisfactoriamente debido a que se han cumplido todos los objetivos propuestos en el inicio del mismo. Así, se logró:

- Estudiar y analizar las distintas técnicas, herramientas y algoritmos de Web Usage Mining para realizar un estudio del comportamiento de los usuarios en el Sistema de Biblioteca PUCV, seleccionando las más adecuadas para el estudio.
- Aplicar y analizar correctamente los resultados obtenidos a partir de las técnicas, herramientas y algoritmos seleccionados para realizar el estudio.
- Validar los resultados con las personas encargadas de administrar el sitio.
- Y por último, realizar todos los ajustes necesarios para obtener resultados más representativos y generar un mejor análisis a partir de estos.

Si bien es cierto que durante el desarrollo del proyecto se tuvo que enfrentar a diversos problemas con respecto a las herramientas utilizadas principalmente, debido a la falta de documentación de las mismas, éstos se pudieron solucionar exitosamente y seguir con el curso del proyecto hasta finalizarlo.

Respecto a los resultados obtenidos, se puede concluir que fueron de gran relevancia para las personas que se encargan de administrar el sistema, ya que pudieron validar o confirmar una de los objetivos principales que ellos tuvieron al momento de implantar este Sistema de Biblioteca, el cual está orientado a que los usuarios naveguen lo menos posible por el sistema para encontrar lo que buscan, ya que su idea es actuar más como un directorio que como un sitio web donde los usuarios navegan e interactúan en mayor proporción.

Como trabajo futuro, se podría incorporar un análisis de forma permanente sobre la forma en que los usuarios navegan e interactúan con el sistema mediante la implantación de una plataforma que permita realizar el análisis propuesto en este proyecto, pero de una manera más persistente para saber cómo actúan los usuarios durante el transcurso del tiempo y así realizar mejoras constantes y adecuadas a los posibles cambios de navegación en los usuarios de este importante recurso en la comunidad universitaria.

Bibliografía

[1] **Yates, Ricardo and Poblete, Bárbara.** *Un modelo de minería de consultas para el diseño del contenido y la estructura de un sitio Web.* Valencia, España : Revista Iberoamericana de Inteligencia Artificial, 2007.

[2] **Camacho, Francisco.** *Web Mining: fundamentos básicos.* 2005.

[3] **Sristava, Jaideep, Cooley, Robert y Deshpand, Mukund.** *Web Usage Mining for Discovery User Behavior.* s.l. : Communications in Computer and Information Science, 2000.

[4] **Fayyad, Usama, Piatetsky - Shapiro, Gregory and Smyth, Padhraic.** *The KDD Process for Extracting Useful Knowledge.* s.l. : Proceedings of the ACM CHI Conference, 1996.

[5] **Rebolledo Lorca, Victor.** *Plataforma para la extracción de los Web data.* 2009.

[6] **Cooley, Robert, Mobasher, Bamshad and Srivastava, Jaideep.** *Data Preparation for Mining World Wide Web Browsing Patterns.* s.l. : Journal of Knowledge and Information Systems, 1999.

[7] **Antonio, González.** *Minería web y personalización: Revisión bibliográfica y propuesta de un marco de referencia.* 2007.

[8] **Vásquez, Juan and Palade, Victor.** *Adaptive Web site: A Knowledge Extraction from Web Data Approach.* s.l. : Frontiers in Artificial Intelligence and Applications, 2008.

[9] **Bernabeu, Ricardo Dario.** *Data Warehousing: investigación y sistematización de conceptos.* Córdoba, Argentina : s.n., 2007.

[10] **Tanasa, Dorus.** *Web Usage Mining:Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support.* 2005.

[11] **Velásquez, Juan, et al.** *Mining Web Data to Create Online Navigation Recommendations.* s.l. : Lecture Notes in Computer Science,Springer, 2004.

[12] **Fayyad, Usama.** *Knowledge discovery in databases- An overview.* s.l. : Lecture Notes in Computer Science, 1997.

[13] **Sánchez, Enrique and Heider, Hisayas.** *Aplicación en minería de datos.* 2008.

[14] **Clemente, Alvaro, López, Pablo and Penado, José.** Grupo de Sistemas Inteligentes. [Online] 2008. http://www.gsi.dit.upm.es/~ gfer/ssii/trabajos2005/Mineria de Datos.

[15] **Jinguang Liu, Roopa Datla.** Docstoc. [Online] Febrero 24, 2010. http://www.docstoc.com/docs/26387002/Web-Usage-Mining/.

[16] **Espinoza, Evelyn and Andaur, Pamela.** *Análisis del comportamiento del usuario en la web para optimizar la estructura de navegación de un sitio usando algoritmos genéticos.* 2010.

[17] **Kosala, Raimond.** *Web Mining Research: A Survey.* 2000.

[18] **Scime, Anthony.** *Web Mining: Applications and Techniques.* New York : s.n., 2004.

[19] **Fuentes, Sady and Ruiz, Marina.** *Web Mining: a necessary resource for the information professional.* 2007.

[20] **Mobasher, Bamshad.** *Web Usage Mining.* 2008.

[21] **Pitkow, James.** *Characterizing browsing behaviors on the world wide web.* s.l. : Elsevier Science Publishers B. V., 1995.

[22] **P, Castaño P. Andres.** *Minería de Uso para la Identificación de Patrones.* 2009.

[23] **Markov, Zdravko and Larose, Daniel.** *Data Mining the Web.* 2007.

[24] **Chang, Horng-Jinh, Hung, Lun-Ping and Ho, Chia-Ling.** *An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis.* 2006.

[25] **Armas, Arazay.** *Herramientas para la detección de errores usando reglas de asociación.* 2009.

[26] **Botía Blaya, Juan.** *Introducción a la Minería de Uso Web.* 2006.

[27] **Ovalle, Alexander and Arias, Demetrio.** *Web Usage Mining : revisión del estado del arte.*

[28] **Andreas Staeding.** User-Agents.org. [Online] 2002-2010. http://www.user-agents.org/.

[29] **Pallis, George, Angelis, Lefteris and Vakali, Athena.** *Validation and interpretation of Web users' sessions clusters.* 2006.

[30] **Velásquez, Juan.** *Mining web data: Techniques for understanding the user behavior in the Web.* 2006.