UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

MEJORANDO EL CONTENIDO TEXTUAL DE UN SITIO WEB A TRAVÉS DE LA IDENTIFICACIÓN DE SUS WEB SITE KEYWORDS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JOSÉ IGNACIO FERNÁNDEZ

PROFESOR GUÍA:
JUAN D. VELASQUEZ S.

MIEMBROS DE LA COMISIÓN:
CRISTIAN CÉSPEDES

SANTIAGO DE CHILE
ENERO 2010

RESUMEN DE LA MEMORIA
PARA OPTAR AL TITULO DE
INGENIERO CIVIL INDUSTRIAL
POR : JOSÉ IGNACIO FERNÁNDEZ J.
FECHA: 25/01/2010
PROF. GUIA: JUAN D. VELASQUEZ S.

Mejorando el contenido textual de un sitio web a través de la identificación de sus website keywords

En esta memoria se presenta el desarrollo y aplicación de una metodología para la identificación de las palabras más importantes contenidas en un sitio web, desde el punto de vista del usuario que lo visita, también conocidas como "*web site keywords*".

El diseño y construcción de un sitio web, es una tarea no trivial que requiere de la definición del contenido correcto y la estructura correcta del sitio para atraer y/o retener a sus eventuales visitantes.

Asumiendo que la estructura de hipervínculos de un sitio esté relativamente correcta, el problema a abordar es qué contenidos atraerán en mayor medida la atención de sus visitantes, es decir, que imágenes, sonidos, colores, textos, etc. motivarían a un usuario a visitar el sitio web.

Existe una estrecha relación entre la cantidad de visitas de un sitio y su éxito en la Web, por lo que asegurar un correcto contenido, implica mejorar las posibilidades de sobre vivencia del sitio en un mercado altamente competitivo como es el digital.

De todos los contenidos antes mencionados, en el presente trabajo de memoria se desarrolla un set de herramientas que junto con una metodología propuesta, permitirían analizar los textos de las páginas de un sitio web y en particular, las palabras que componen dichos textos con miras a detectar aquellas que atraen la atención del usuario.

El sitio web seleccionado para propósitos de experimentación y validación del trabajo debería ser complejo con respecto a varias características: número de visitas, actualización periódica y ser rico en contenido textual. La página web de un banco virtual Chileno (sin sucursales físicas y donde todas sus transacciones son realizadas electrónicamente) cumplió con dichos criterios.

Se logro finalmente realizar el proceso de identificación de las palabras según las preferencias de navegación de los usuarios, determinando cuales eran más relevantes según la importancia dada por el "peso de la palabra" y por su relación con la industria. Palabras como Crédito, Financiamiento, Ahorro fueron algunas de las detectadas en el proceso y se agruparon en torno usuarios agrupados por la aplicación de algoritmos de web mining.

La validación de las palabras encontradas, y de la metodología planteada, se realizó a través de un test de efectividad de las palabras claves detectadas, donde se consultó a usuarios de Internet y clientes de bancos cual era la relevancia de las palabras clave o *web site keywords* en los textos de las páginas del sitio web, corroborando la relevancia que tiene el uso de estas palabras en el contenido textual de las páginas web.

Se concluye el trabajo mostrando las posibilidades de mejora del sitio web tanto textuales, por la identificación de palabras a través de la metodología con técnicas de web mining, como mejoras estructurales detectadas en el transcurso del trabajo.

INDICE DE CONTENIDO

INDICE DE CUADROS

INDICE DE FIGURAS

INDICE DE ECUACIONES

INDICE DE TABLAS.

INDICE DE GRAFICOS.

CAPITULO 1

1. INTRODUCCIÓN

El desarrollo tecnológico de información y comunicaciones: Internet, Telefonía Móvil, Banda Ancha, Satélites, Wi-fi están produciendo cambios significativos en la estructura social, cultural y económica a nivel mundial. Se habla de un nuevo tipo de sociedad llamada "Sociedad de la información[1]" por los mismos progresos y facilidad de acceso de estas tecnologías.

Las empresas a su vez han estado sujetas a cambios de acuerdo a la evolución de las Tecnologías de Información y Comunicaciones. Diferentes son las áreas que hacen uso de estas donde la información financiera, operaciones, recursos humanos queda a la mano y con una facilidad de acceso y presentación que genera mayor dinamismo en la empresa, así como también la oportunidad de manejar información de mercados internacionales, monedas, eventos económicos y políticos a nivel global. Tal ha sido el crecimiento e importancia de Internet que hay empresas operando a nivel mundial a través solamente de este canal y que basan todo su negocio en este medio (ejemplos: Google, Facebook, Twitter, etc.)

Una de las ventajas que Internet otorga a las empresas es la oportunidad de tener una comunicación directa con sus clientes. A través de este canal se puede presentar productos y servicios, opciones de consultas, sugerencias, reclamos y ofertas, promociones e incluso campañas de fidelización y marketing a través de promociones dirigidas a sus correos electrónicos o equipos de telefonía móvil. Dentro de estos canales directos de información con el cliente, Internet permite a las empresas colocar amplia información acerca de sus actividades, misión, visión, clientes, dirección de oficinas, correos y teléfonos de contacto, productos, servicios y ofertas de interés. Junto con ello complementar dicha información con elementos multimedia como imágenes, sonidos y videos de manera de hacer más atractiva la exposición realizada y cautivar de alguna manera al cliente que visita la página web.

El cómo accede un usuario a un sitio web se explica principalmente a través de uso de buscadores de Internet con un 33% de las preferencias, seguido de un 26% encontrado dentro de la navegación, 17% a través de publicidad online (banners, Google ads), 12% por recomendaciones de amigos y sólo un 4% a través de periódicos[2]. Dentro de este comportamiento del usuario, al momento de tener un resultado de búsqueda, un 23% de usuarios de Internet revisa solo unos pocos resultados de la primera página de su búsqueda, un 39% sólo la primera página, 19% sólo dos páginas de resultados y el restante 19% más de dos páginas. En este mismo concepto, si el usuario dentro de sus preferencias no encuentra en los primeros resultados lo que busca, un 82% reformula la búsqueda en el mismo motor elegido con otros términos[3].

Según lo expuesto, el objetivo del área a cargo del sitio web de una empresa es aparecer dentro de los primeros resultados de un motor de búsqueda como Google o Yahoo. Para ello se realiza la Optimización de Buscadores en Internet (SEO por sus siglas en inglés, Search Engine Optimization) lo cual corresponde a la mejora del volumen o calidad del tráfico de un sitio web

[1] La sociedad de la información es vista como la sucesora de la sociedad industrial cuya economía fue basada en la industria y manufactura en reemplazo del trabajo manual.

[2] Estudio Tendencias Digitales, Julio del 2009. http://www.tendenciasdigitales.com

[3] iProspect Search Engine User Behavior Study, Abril 2006

de forma natural o "no pagada" que ingresa desde la página desde resultados de un buscador de internet. Esta optimización se realiza a través de una serie de chequeos a realizar.

1. Arquitectura del sitio web:
    a. Lenguaje de programación.
    b. Meta tags, tags, énfasis, ennegrecidos, subrayados, formato de textos.
    c. Navegación, sitemaps (mapas del sitio web)
    d. Diseño del sitio web.
2. Contenido:
    a. Análisis de Keywords o palabras claves para determinar las palabras más "lucrativas".
    b. Consolidación del contenido con la estructura del sitio.
    c. Medición del número de páginas para alcanzar el ranking.
    d. Frecuencia de actualización y distribución de contenido.
3. Meta Data y Tags:
    a. Etiquetado de elementos adicionales en el sitio web (tablas, imágenes, videos).
4. Links internos y externos:
    a. Creación de Landing pages (páginas de llegada) apropiadas.
    b. Mapeo de links

La información de uso del sitio web la entregan sistemas de contadores o de análisis que dan resultados estadísticos de las visitas de los usuarios: Cuantas veces visitó la página, en que momento fue visitada, cuanto tiempo estuvo en una página, cual fue la ruta de salida del cliente. Esta información entregada, si bien es una buena fuente de retroalimentación del sitio web a nivel cuantitativo, no permite la identificación de perfiles de navegación, clasificación de los usuarios o si los textos de las diferentes páginas contienen los conceptos que el usuario esta buscando o si estos están inmersos en los intereses inherentes al usuario.

Todos aquellos conceptos que el usuario considera de interés de acuerdo a su búsqueda, están principalmente inmersos en el *contenido* de las páginas del sitio web, es decir, en los textos, palabras, títulos y tablas. Estos contenidos conjugados con los registros de navegación del usuario almacenados en el servidor web (weblogs) sumado a la utilización del proceso de descubrimiento de conocimiento en las bases de datos o KDD[4] aplicado al web mining, se obtiene la clasificación de los usuarios y los contenidos que son de su interés, con lo cual se logra extraer el conjunto de palabras que engloban los conceptos de interés que originaron la búsqueda. Estas palabras se conocen con el nombre de palabras claves del sitio web o comúnmente *website keywords*.

Las *website keywords* [4] son palabras que son relevantes dentro de la industria donde se encuentra la organización y que son capaces de entregar conceptos o que, inmersas en una frase, generan el llamado de atención de los clientes al momento de visitar un sitio web.

El trabajo realizado en esta memoria consiste en identificar las palabras claves de un sitio web y entregarlas junto con un set de recomendaciones contextuales con el fin de que el propietario del sitio web conozca, desde el punto de vista del usuario que ingresa a su *website*, cuáles son esas palabras importantes y sobre que grupo de clientes están inmersas. Con esto

[4] KDD por las siglas en inglés de Knowledge Discovery in Databases
[6] URL: Universal Resource Location.

realizar acciones de mejora textual del sitio web, lo que permitirá tener contenido más ajustado a las necesidades de quienes visitan el sitio web, mayor cantidad de clientes y clientes fidelizados.

1.1. Potencialidades de un sitio web

Los avances realizados en materia de software, hardware y conectividad han logrado un crecimiento exponencial de la Tecnologías de Información y Comunicación (TIC) en la integración con la sociedad. La World Wide Web, como parte de este avance, a evolucionado de hipertextos complementados con contenido multimedia a aplicaciones interactivas, herramientas de apoyo a la gestión, canales de comunicación directa y redes sociales. Adaptándose a esta evolución, la empresa ha debido ajustar constantemente su estructura en beneficio de la operación, finanzas y comunicación. Parte de los beneficios que estos ajustes han generado en el cliente usuario de Internet son: Mayor y nuevos canales de atención (email, chat, Twitter, Facebook), servicios online de auto atención, pago y revisión de cuentas, transferencias electrónicas, compras, integración con cadenas logísticas (despachos) entre otros.

Todo indica para las organizaciones que la interactividad constante con el cliente a través de internet y la oportunidad de una comunicación directa son argumento necesario para considerar una página web como elemento integrador de esta comunicación axial como también una opción de fidelización, retención, adquisición de nuevos clientes, mejoras y seguimientos de gestión.

Como parte de las mejoras posibles en la empresa a partir de una potenciación del canal de Internet se encuentran:

- Cercanía y comunicación directa con el cliente a través de email, chat o de servicios de atención online.
- Posibilidad de marketing dirigido a través de redes sociales o de buscadores temáticos como Google.
- Disminución de atenciones físicas (sucursales) o remotas (call center) por utilización de servicios de auto atención.
- Automatización de sistemas de integración logística, compra o gestión de proveedores.
- Potencial crecimiento de clientes por presencia en Internet.

1.2. La Web en chile.

Por oportunidad de utilización y funcionalidades, la industria bancaria, retail y de servicios ha logrado incluir de forma integra Internet llegando al punto de autoadministración de cuentas por el cliente. Desde la información obtenida por la Cámara de Comercio de Santiago (CCS) de los usuarios activos de Internet, un 40% de la población chilena corresponde a personas de 6 o más años, un 54,8% hace sus compras en tiendas de retail a través de sus sitios web de venta en línea, un 31,4% utiliza la banca digital para realizar transacciones bancarias (transferencias, consulta de saldo, solicitud de talonarios, entre otros) y un 19,8% de la población utiliza este

medio para el pago de cuentas que por lo general esta estrechamente ligado a la banca por la utilización de tarjetas de crédito en el pago electrónico.

La mayor utilización de Internet en Chile es para la obtención de información de empresas, servicios públicos, lectura de diarios y hobbies más que como aplicaciones de utilidad o herramientas complementarias en las labores diarias. Uno de los principales motivos del no uso de Internet para aspectos comerciales como compras o transacciones es la desconfianza o poca claridad a nivel de seguridad, poca información de la empresa u oferta deficiente de los productos y servicios. Esto lo constituye el 18% de los usuarios de Internet.

1.3. Trabajo realizado.

Consiste en la creación de modelos de comportamiento del usuario en la Web y en el desarrollo de algoritmos de *web mining* para la extracción de las *website keywords* presentes en un sitio, las que luego son usadas para el mejoramiento de su contenido textual.

La identificación de las *website keywords* permite al *web master* o administrador encargado del *website*, optimizar su contenido textual. En efecto, del correcto uso de las palabras con que se redacta la información lexicográfica del sitio, dependerá en gran medida su éxito en el mercado digital, es decir, logrará atraer nuevos visitantes y mantener a los clientes que ya frecuentan el sitio.

El presente informe de memoria, fue organizado de la siguiente manera: el capítulo 2 contiene todo el soporte teórico del trabajo a realizar en cuanto a los estudios de *web mining* relevantes en el trabajo. El capítulo 3 muestra la implementación de algoritmos y redes neuronales de fuente propia para la ejecución del trabajo de identificación de palabras claves del sitio web o *website keywords*. El capitulo 4 muestra el caso de aplicación de las herramientas construidas para el caso de un sitio web real lo que deriva en recomendaciones de aplicación y uso mostradas en el capítulo 5. Finalmente el Capítulo 6 muestra las conclusiones del trabajo realizado y hace mención de las posibilidades de mejora así como también otros potenciales trabajos de la misma línea a realizar.

2. OBJETIVOS.

El objetivo principal del trabajo a realizar es:

> *Identificar las palabras más relevantes de las páginas de un sitio web según el punto de vista del usuario.*

Para el cumplimiento de este objetivo, se establecen los siguientes objetivos específicos:

- Diseñar e implementar un módulo de preprocesamiento, transformación de datos y aplicación de algoritmos de web mining para la identificación de palabras relevantes contenidas en las páginas del sitio web.
- Identificar las palabras importantes de los grupos de usuarios de un sitio web.
- Establecer una metodología para la identificación de las palabras relevantes.

CAPITULO 2

3. LA WORLD WIDE WEB

Se entiende por *la Web* al sistema de documentos de hipertexto enlazados y accesibles a través de Internet, la cual es un sistema de interconexión descentralizada de computadores, implementado en un conjunto de protocolos denominado TCP/IP y que garantiza que redes físicas heterogéneas funcionen como una red lógica única. Los orígenes de Internet se remontan al año 1969 cuando se estableció la primera interconexión de computadores conocidos como ARPANET entre universidades de California y Utah en Estados Unidos.

La *Web* es un elemento de Internet mucho más reciente y se establece que en 1990 el inglés *Tim Berners-Lee* en conjunto con el belga *Robert Cailliau* mientras trabajaban en el CERN en Ginebra, Suiza, dio con la creación de este.

3.1. Operación en la Web

La visualización de una página web comienza al momento que el usuario interesado en navegar digita una URL[6] o dirección web en el browser o navegador, o bien mediante el "clickeo" de un enlace existente en algún documento o sitio web que se encuentre visitando. Al entrar la dirección, el siguiente proceso es la transformación de la URL en una dirección IP[7], trabajo que realiza el servidor de nombres de dominio DNS[8], haciendo la traducción respectiva para identificar al computador que alberga la página en Internet y posteriormente al servidor web que la administra.

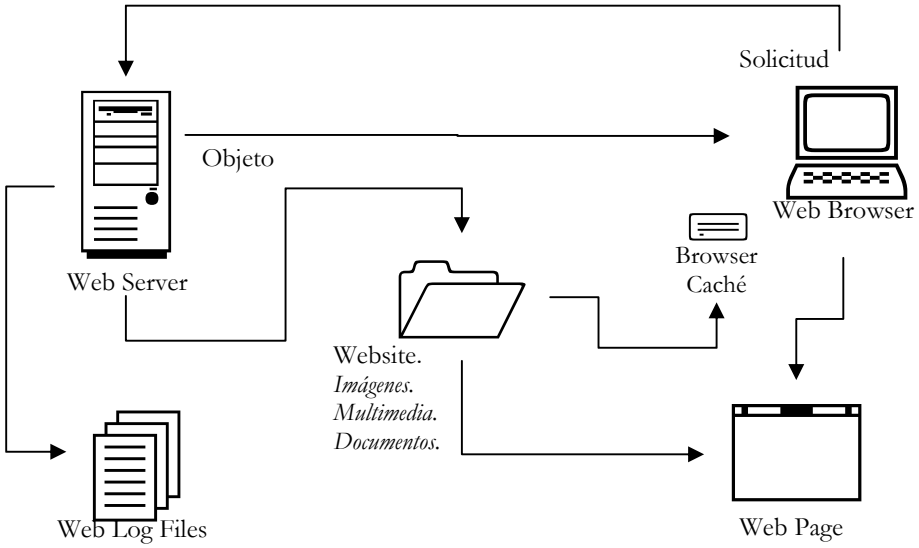
Una vez que se obtiene la dirección IP, se establece la conexión TCP/IP con el equipo donde se encuentra el servidor web que administra la página. Seguido, se hace el requerimiento del recurso mediante el protocolo HTTP[9] cuya solicitud devuelve la página web que en síntesis es un texto con etiquetas que posteriormente será interpretado por el navegador web. Dependiendo de la etiqueta, el navegador web comienza a realizar el "pintado" de la página en la pantalla del usuario. Algunas de estas etiquetas son peticiones a otros objetos que están en el sitio, por ejemplo imágenes, multimedia y documentos anexos, lo cual originará una nueva petición del navegador al servidor web que administra el sitio. De esta forma es que se genera la acción de navegación en la Web por Internet.

La *figura 1* identifica las diferentes componentes de la operación en la web y su interacción en un proceso de navegación.

Durante el proceso de la navegación del usuario en un sitio web, el acceso de cada página web genera un requerimiento que el servidor web devuelve para que sea traducida y desplegada por el navegador. Cada uno de los requerimientos, de la página como sus componentes, quedan registrados con información como IP, fecha, origen del requerimiento, requerimiento devuelto, entre otros en un archivo llamado *web log*. La conjugación entre las páginas web o *web pages* del website es lo que se conoce como *web data* cuyo volumen, en particular el de los *web logs*, puede llegar a ser enorme dependiendo del flujo de requerimientos y cantidad de ellos.

[7] IP: Internet Protocol.
[8] DNS: Domain Name Server.
[9] HTTP: Hyper Text Transfer Protocol.

Figura 1: Esquema de Operación en la Web



3.2. Datos originados en la web

Existen tres tipos de datos que se pueden extraer desde la Web. El primero corresponde al *web log* que es el registro de requerimientos de cada usuario que es almacenado en el servidor web. El segundo de ellos son las *web pages* que corresponden al sitio web como set de documentos, gráficas y multimedia que son las almacenadas y requeridas por los usuarios. Finalmente, una tercera fuente de datos es la estructura de hyperlinks desde donde se puede obtener información de navegación del usuario en el sitio web. A continuación se explican las dos primeras fuentes de web data en detalle que son las más relevantes para análisis de comportamiento y preferencias del usuario [33].

3.2.1. Web Logs.

El principal componente de los web data, son los web logs, puesto que a partir de ellos se puede investigar, analizar y clasificar el comportamiento del usuario en el sitio web. El web log no es un archivo transaccional, lo constituyen líneas de interacción entre el usuario del sitio que está visitándolo y el servidor donde que lo administra. Un ejemplo de líneas de web log son las mostradas en el cuadro 1.

Cuadro 1: Líneas de un Web Log

167.106.16.132 - - [03/Mar/2002:22:09:00 +0100] "GET /directorio/16.html HTTP/1.1" 200 - "http://www.dominio.com/directorio/index.html" "Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; Win 9x 4.90)"

200.66.196.132 - - [03/Mar/2002:23:59:20 +0100] "GET /directorio/16.html HTTP/1.1" 304 - "http://www.dominio.com/directorio/page.html" "Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; Win 9x 4.90)"

Si bien el web log fue construido con el fin de almacenar los eventos entre cliente y el servidor, un uso que se puede dar a estos registros es extraer datos estadísticos de la navegación de usuarios en el sitio, como páginas más visitadas, usuarios que han realizado la visita, evolución de las visitas, navegador utilizado, y otros índices de importancia en la revisión y análisis de los accesos y acciones del usuario.

Por la naturaleza de los web logs, también se suma la posibilidad de aplicar técnicas de web mining en búsqueda del comportamiento del usuario como: Contenido de más interés para el usuario, patrón de navegación, fidelidad del cliente en la visita, entre otras más.

La escritura de registros en los web logs, puede ser configurada por el usuario que administra el sitio en el servidor, para que entregue información adicional a los campos del formato estándar llamado *Common Log Format* (CLF) diseñado por CERN Y NCSA.

Los CLF son utilizados por la mayoría de los servidores web, lo que permite que diversas herramientas[10] de análisis estadístico [6] puedan trabajar con los datos almacenados. Los campos que se registran en los archivos log son los siguientes:

- **Host**: Correspondiente a la dirección IP con que el usuario ingresa al sitio web. Como se verá más adelante, esta dirección IP puede no corresponder a la del equipo directamente sino al servidor Proxy o Firewall del proveedor de servicios de Internet (ISP). El fin de esta acción es resguardar la seguridad y privacidad del cliente.

- **User**: Corresponde al nombre de identificación con que el usuario ingresa a la página web. Los sitios web que solicitan autentificación previa antes de mostrar sus páginas, son las que hacen uso de este campo.

- **User Id**: Código de identificación del usuario al ingresar a un sitio web con autentificación previa. El formato de este campo es un número entero.

- **Fecha y Hora**: Es la fecha y hora en que el usuario hace el requerimiento al servidor web y este responde con la ruta del archivo o con el mensaje de error en caso de no encontrarlo. El formato de almacenamiento de la Fecha y Hora es generalmente dd/mmm/aa:hh:mm:ss donde dd es el día mmm es el mes en las tres primeras letras del mes, aa es el año, hh es la hora en formato 24 hrs., mm son los minutos y ss los segundos del requerimiento.

- **Tipo de requerimiento, archivo requerido y protocolo**: Línea con identificación del requerimiento del cliente como tipo de requerimiento como GET o POST, ruta de

[10] Webtrends, Getstats, Analog entre otros

respuesta del archivo solicitado por el cliente y el protocolo http de transferencia (versión).

- **Id Respuesta Servidor**: Es la identificación de la respuesta que da el servidor al requerimiento. Corresponde a un número o código de 3 dígitos indica la respuesta que tiene el servidor al requerimiento. Una respuesta exitosa del requerimiento, es decir, el archivo se encuentra y puede ser entregado al usuario, es el código 200; un error de compilación de una página dinámica devuelve el código 500; la página no encontrada entrega el código 404, etc.

- **Bytes**: Tamaño del documento o archivo que es enviado como respuesta del requerimiento del usuario. Los mayores valores en este campo son dados por imágenes de gran tamaño, documentos o archivos para descarga como programas.

- **Referencia**: Origen del acceso al requerimiento en proceso. Puede ser un link a la página donde se encontró el llamado al requerimiento o puede ser un campo vacío para el caso en que el acceso haya sido directamente por la digitación en la barra de dirección del navegador.

- **Agente**: Identificación de software con que el usuario ingresa y hace los respectivos requerimientos en el sitio web. Corresponde a una cadena que incluye versión del navegador, sistema operativo, plug-ins instalados en el navegador, entre otros. Es este campo también el que permite identificar los casos en que el acceso no es realizado por un usuario humano sino por un robot como Google.

La utilización de los web logs permite analizar e investigar el comportamiento y preferencia que tiene el usuario en ciertas páginas del sitio web. Los registros contenidos, sin embargo, no están libres de errores o de ausencia de información. Algunos aspectos y alcances a considerar previamente a la utilización de estos archivos es:

- Los web logs son archivos de registro de requerimientos no transaccionales, por lo que no existe un orden de aparición de ellos en el archivo. La referencia para esto es la utilización de las fechas.

- No hay identificadores de unicidad de navegación. El host de origen del requerimiento es el más cercano al identificador del usuario, pero dado que este host puede ser un firewall o web Proxy, no se cuenta con la certeza de que sea uno o varios los usuarios que hicieron requerimientos con la misma dirección.

- Los web logs contienen información de todos los requerimientos que tiene una sesión de algún usuario. Luego, dependiendo del análisis e investigación que se realice, los registros pueden no ser relevantes en ellos[11].

- Hay acciones que realiza el usuario durante su sesión que no son registradas en el *web log*. La utilización de los botones en la barra de herramientas como *back*, *forward*, *print* o *save* durante la navegación se ignoran. Lo mismo sucede con la utilización del Mouse en la navegación cuando se hace *scrolling* en la página en búsqueda de contenidos en la parte inferior de la página web. Esta información, si bien puede ser útil en el análisis, ya que muestra el interés que tiene el usuario por

[11] Se hará referencia a este punto en el ítem de análisis y sesionización de los web logs.

los contenidos de la página que navega, no queda almacenada ya que esas acciones se realizan cuando la página ya fue retornada por el servidor y se queda almacenada en el caché del *browser*. Luego, al momento de realizar las acciones mencionadas anteriormente, estas son de origen y resultado local, sin la necesidad de intervención de requerimientos por el servidor.

Los *web logs*, a pesar de haber sido construidos para almacenar la interacción entre el usuario y el servidor, entregan información valiosa para analizar estadísticamente las acciones realizadas en el sitio web. Como se mencionó anteriormente, son muchas las herramientas disponibles, tanto gratuitas como comerciales, que permiten hacer análisis estadístico de estos archivos. Sin embargo, ninguna de estas herramientas realiza data mining sobre los registros, por lo que no se pueden clasificar ni aplicar algoritmos para obtener información relevante a nivel estratégico. Desde los resultados de estas herramientas es posible conocer la cantidad de visitas al sitio web, páginas más frecuentes, tiempos de navegación, página más visitada, pero consultar el porque esta ingresando a una página web o cual fue el comportamiento anterior de un usuario que realiza una compra no es posible.

3.2.2. Web Pages

El conjunto de documentos e imágenes que conforman el sitio web se conoce como web page o página web. Las páginas web son documentos codificados mediante una estructura de etiquetas o tags que contienen instrucciones de la forma en que será mostrada la información envuelta entre esas etiquetas. El layout que tiene la página al momento de desplegarse se da por las diferentes instrucciones de las etiquetas que están contenidas en el encabezado y cuerpo del documento.

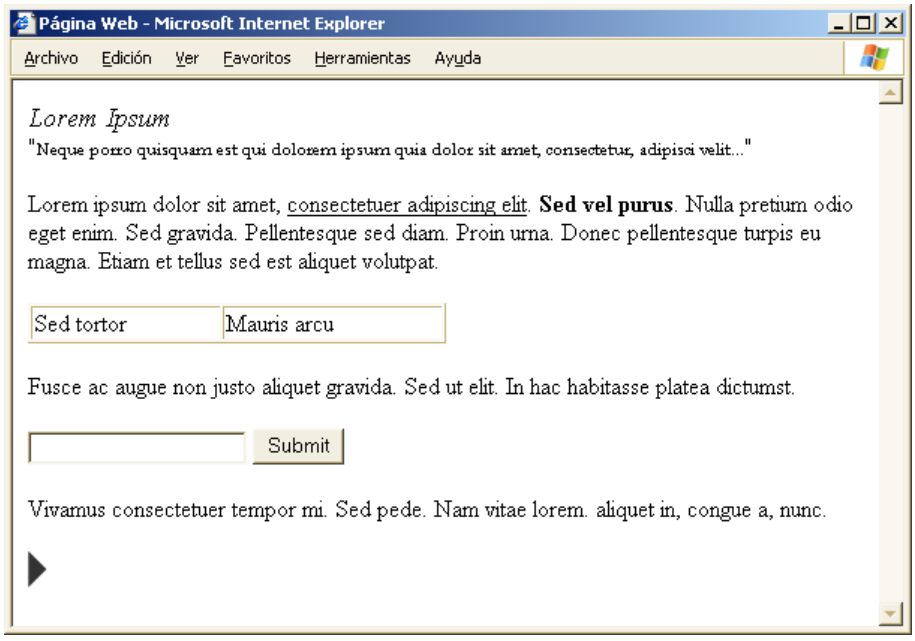
El cuadro 2 muestra la estructura de una página web simple. El orden que se muestra se basa en la estructura de contenidos y etiquetas que darán el formato final de la página., por ejemplo, aplicar ennegrecido sobre un contenido o almacenar palabras en una tabla. Hay etiquetas que son de instrucciones directas como “ennegrecer” (<bold>), párrafo (<p>), etc. Otras etiquetas contienen instrucciones adicionales en su interior que son algunas especificaciones en el formato que se desea dar (por ejemplo, dar un tamaño pequeño de la fuente a mostrar: <font size=1>).

Cuadro 2: Ejemplo formato de etiquetas para una página web.

```
<html>
<head>
<title>Página Web</title>
<meta name="keywords" content="pagina web">
<link href="estilo.css" rel="stylesheet" type="text/css">
<script src="funcion.js"></script>
<script language="JavaScript" type="text/JavaScript">
function date() {print("hola");}
</script>
</head>
<body>
<font face="Times New Roman" size="4"><em>Lorem Ipsum</em></font><br>
&amp;quot;<font face="Garamond" size="2">Neque porro quisquam est qui dolorem ipsum quia
```

```
dolor sit amet, consectetur, adipisci velit...</font>&amp;quot;<br>
        <p>Lorem ipsum dolor sit amet, <u>consectetuer adipiscing elit</u>. <strong>Sed vel purus</strong>. Nulla pretium odio eget enim. Sed gravida. Pellentesque sed diam. Proin urna. Donec pellentesque turpis eu magna. Etiam et tellus sed est aliquet volutpat. Etiam faucibus libero vel lorem. Nunc mattis. Maecenas quam. Nam rhoncus. Aenean eu arcu. Sed urna. Etiam tortor. Vestibulum euismod leo in felis. </p>
        <table width="50%" border="1" cellspacing="1" cellpadding="1">
         <tr>
          <td>Sed tortor</td>
          <td> Mauris arcu</td>
         </tr>
        </table>
        <p>Fusce ac augue non justo aliquet gravida. Sed ut elit. In hac habitasse platea dictumst. </p>
        <form name="formulario" method="post" action="">
         <input type="text" name="textfield">
         <input type="submit" name="Submit" value="Submit">
        </form>
        <p>Vivamus consectetuer tempor mi. Sed pede. Nam vitae lorem. aliquet in, congue a, nunc.</p>
        <img src="image.gif">
        </body>
        </html>
```

Figura 2: Vista de la página del ejemplo anterior desde el navegador



El navegador tomará las instrucciones descritas en el cuadro 2 y las irá interpretando, para ir generando gradualmente la página web. La figura 2 muestra la "interpretación" de la página web que es lo que finalmente verá el usuario.

El listado de etiquetas existentes en la codificación de una página web es muy amplio, pero existen generalmente algunas que son utilizadas obligatoriamente y otras que son de forma alternativa. A continuación se muestra una referencia de las etiquetas más frecuentemente utilizada en documentos.

- **Página Web**: Iniciada con la etiqueta <*html*> y finalizada mediante </*html*>. Dentro de estos marcadores es colocado todo el contenido de las *web pages*. Para el caso de utilización de páginas dinámicas, parte del código que se compila en el servidor puede estar generado antes de comenzar el *tag* de inicio del documento.
- **Encabezado:** Contiene instrucciones de inicio del documento que será cargado por el navegador con las etiquetas <*head*></*head*>:
  - **Titulo**: titulo del documento limitado por la etiqueta <*title*>.
  - **Estilos del documento:** los estilos del documento son una extensión dentro de las etiquetas html que dan formatos especiales a la estructura de la página web, al contenido textual o bien unifican la utilización de los formatos. El estilo de la página web es un archivo que se anexa mediante la etiqueta <*link…*> que contiene la fuente del documento de estilo o css[12] que será cargado junto con la página web; o bien mediante la declaración de configuración de estilos directamente con la etiqueta <*style*> dentro del documento que contendrá el estilo.
  - **Funciones y Utilidades:** Corresponde al anexo de funciones *javascript* y/o *vbscript* que son de utilidad para la utilización de la página web. Estas son limitadas mediante los tags <*script*> y al igual que la hoja de estilo puede estar en un archivo de extensión .js adjunto, pero que en este caso se declara dentro de la misma etiqueta como <*script src=”archivo.js”*></*script*>
- **Cuerpo**: Contiene los contenidos visuales a desplegar en el navegador. El layout que tenga la página al ser desplegada será el resultado de la traducción de las etiquetas según el orden que se encuentren. Esta parte del documento se encuentra demarcada mediante las etiquetas <*body*></*body*> respectivamente. En el cuadro 3 se muestran las etiquetas más comúnmente utilizadas dentro del cuerpo del documento.

3.3. Web Mining

Se conoce como *Web Mining* [8, 16, 25] a la aplicación de técnicas de data mining a análisis de los web data, con el fin de estudiar variados aspectos relacionados con un sitio web en particular o con la misma Web. El fin es descubrir patrones, tendencias y comportamientos de los usuarios desde donde se pueden generar mejoras de estructura, contenidos y usabilidad.

Para el descubrimiento de patrones de comportamiento, se hace uso de algoritmos de minería de datos o *Data Mining*, que procesando información existente desde el web data, retornan resultados que mediante técnicas convencionales de análisis no se lograrían obtener. Las técnicas se asocian a grandes cantidades de datos los cuales son procesados y luego representados

[12] CSS: Cascade Style Sheet.

en modelos con el fin de ser analizados. El traslado de este tipo de análisis a la web es lo que se conoce como *Web Mining*.

Cuadro 3: Listado de Etiquetas comúnmente utilizadas en una página web

Contexto	Descripción	
Documento HTML Básico	<html> <head> <title>Nombre el documento</title> </head> <body> Contenido Visible </body> </html>	
Elementos de Encabezado	Mayor Tamaño de Encabezado . . . Menor Tamaño de Encabezado	
Elementos para Texto	Este es un párrafo (quiebre de línea) <hr> (regla horizontal)	
Estilos Lógicos	Texto enfatizado Texto Ennegrecido	
Estilos Físicos	Texto en negrita Texto en cursiva o itálico	
Enlaces	Este es un link Enviar correo	
Listas	No Ordenadas. - Primer Ítem - Siguiente item 	Ordenadas: - Primer Item - Siguiente Item
Tablas	<table border="1"> <tr> <th>encabezado a</th> </tr> <tr> <td>texto a</td> </tr> </table>	
Formularios	<form action="http://www.example.com/test.asp" method="post/get"> <input type="text" name="lastname" value="Nixon"> <input type="password"> <input type="checkbox" checked="checked"> <input type="radio" checked="checked"> <input type="submit"> <select> <option>Apples <option selected>Bananas <option>Cherries </select> <textarea name="Comment" rows="60" cols="20"></textarea>	

Entre las diversas técnicas del *data mining* que se utilizan en *web mining,* es posible

identificar algoritmos de clasificación y agrupamiento (*clustering*), reglas de asociación y sucesos frecuentes, desde donde se puede, por ejemplo, clasificar y agrupar a los usuarios para asignarles patrones de comportamiento, según la reiteración de acciones que se detecten y de acuerdo a ello, brindarles la posibilidad de productos y servicios que se ajustan a estos perfiles.

El *Web Mining* se subdivide según los tipos de web data a procesar:

- *Web Content Mining* (WCM) o Minería de Contenidos de la Web se centra en el análisis de los contenidos desplegados en la web y que pueden ser significativos al tratar de obtener las formas de escribir o palabras que son más atractivas a los usuarios, si los contenidos son de interés del grupo entre otros.
- *Web Structure Mining* (WSM) o Minería de Estructuras de la Web es el análisis de la forma en que se encuentran dispuestos los contenidos en el sitio web. Se puede definir desde este tipo de análisis si es que el usuario encuentra la información que busca, si esta demasiado dispersa o profundiza demasiado. Si los elementos se encuentran en el lugar adecuado o si se entiende la estructura presentada, cuales son las secciones menos visitadas, etc.
- *Web Usage Mining* (WUM) o Minería de Uso de la Web es el descubrimiento de patrones de navegación que realiza el usuario en el sitio web y que permiten generar acciones de cambio para mejorar la experiencia de navegación de los usuarios.

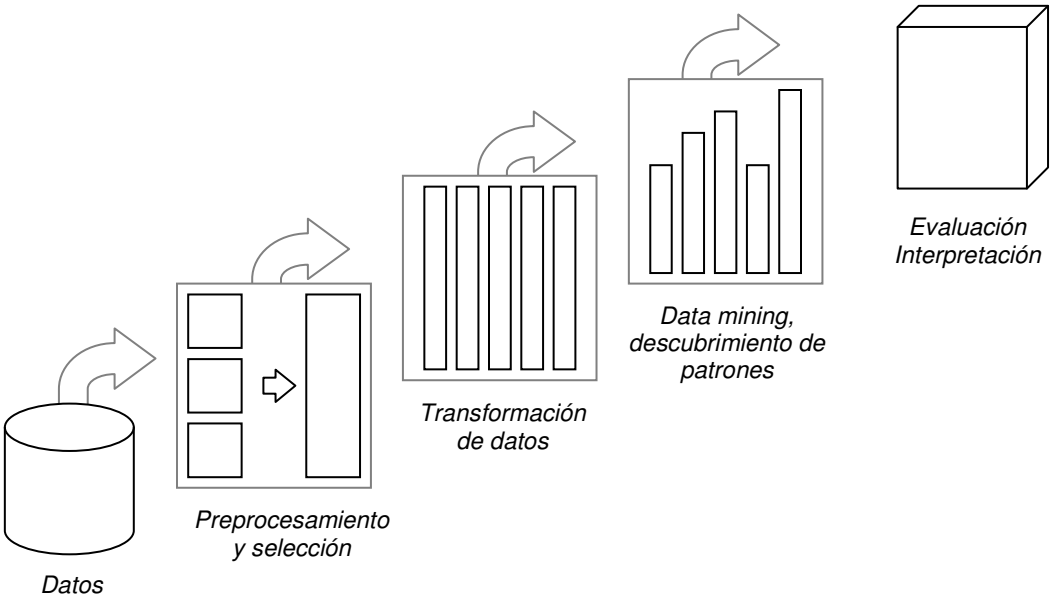
Los análisis anteriores, no son independientes ni mutuamente excluyentes. La disponibilidad de los web data permiten efectuar análisis pasando por los diferentes subgrupos de *web mining* y en algunas ocasiones combinarlos para obtener los resultados deseados.

En el trabajo realizado se hace la combinación de *Web Usage Mining* para el análisis de los patrones de navegación y comportamiento de los usuarios según los registros almacenados en los *web logs*; se hace uso también del *Web Content Mining* para el análisis de contenidos y palabras que se encuentran en la *web page*. El fin del trabajo es combinar la información de ambos orígenes de datos para obtener el comportamiento del usuario, sus preferencias de navegación y las palabras claves o *website keywords* que identifican esas preferencias.

3.4. El proceso de *knowledge discovery on database* (KDD) aplicado a la Web.

El descubrimiento de patrones en un website es una de las etapas importantes en el proceso KDD que se debe aplicar para la obtención de las *website keywords.* La figura 3 muestra las etapas de este proceso que son detalladas a continuación bajo un análisis orientado a la web [27].

Figura 3: Proceso de KDD



3.4.1. Obtención de Datos.

Corresponde a la obtención de los *web data,* que será la información de origen en la investigación. Incluye los archivos que se desean analizar desde sus distintas fuentes de origen y las acotaciones sobre ellos, por ejemplo, intervalo de fechas de los *web logs*, conjunto de *web pages*, etc.

Con el objeto de certificar consistencia en los resultados y en el análisis a realizar, se establece una fecha de estudio y se saca en ese período una "imagen" o copia del sitio web para replicar las páginas que se muestran durante este periodo. Con esto, se logra que los registros de uso del sitio web (web logs) sean consistentes con la información y textos libres contenidos en las páginas del sitio web.

Se busca que la asociación de contenidos páginas del website sea consistente con las líneas del *web log* analizado. Una de las propiedades de las páginas web es que estas son fácil y rápidamente modificables en estructura. Puede tratarse incluso de una página web dinámica donde, mediante administradores, se puede agregar en línea información, la que se despliega en el

sitio según criterios previamente definidos. Un ejemplo de este tipo de acceso es utilizar una variable aleatoria o a través del criterio de la fecha de acceso al sitio web.

En resumen, la determinación de fecha inicio y término del análisis de los web data, tiene como finalidad la adquisición páginas web y registros de logs correctamente relacionados.. Debido a que los web logs pueden ser extensos también la utilización de fechas límite permite limitar y acotar la cantidad de datos en el estudio.

3.4.2. Preprocesamiento y Selección.

Consiste en la depuración y limpieza de datos para obtener la información de análisis clara y sin componentes que puedan causar error en interpretación o procesamiento.

3.4.2.1. Preprocesamiento y Selección de web logs.

Los registros presentes en los web logs, pueden ocasionar una sobre dimensión de la verdadera interacción del usuario con el sito, por lo cual los registros deben ser agrupados para caracterizar la sesión del usuario. El ejemplo del cuadro 4 muestra el requerimiento de una página web seguida por las imágenes y archivos anexos que componen el requerimiento:

Cuadro 4: Ejemplo de requerimiento de una página web con imágenes y archivos referenciados

```
164.77.100.19 [01/Jan/2003:09:33:55 -0400] "GET /servicios/plan/index.html HTTP/1.1" 200 14648
164.77.100.19 [01/Jan/2003:09:33:56 -0400] "GET /estilos.css HTTP/1.1" 200 8078
164.77.100.19 [01/Jan/2003:09:33:56 -0400] "GET /javascript/funcion.js HTTP/1.1" 200 19394
164.77.100.19 [01/Jan/2003:09:33:57 -0400] "GET /imagen/background.gif HTTP/1.1" 200 233
164.77.100.19 [01/Jan/2003:09:33:58 -0400] "GET /imagen/logo.gif HTTP/1.1" 200 2768
164.77.100.19 [01/Jan/2003:09:33:58 -0400] "GET /imagen/adsbanner.gif HTTP/1.1" 200 1569
164.77.100.19 [01/Jan/2003:09:34:00 -0400] "GET /imagen/separador.gif HTTP/1.1" 200 49
164.77.100.19 [01/Jan/2003:09:34:00 -0400] "GET /imagen/tabulador.gif HTTP/1.1" 200 89
164.77.100.19 [01/Jan/2003:09:34:00 -0400] "GET /imagen/espaciador.gif HTTP/1.1" 200 179
164.77.100.19 [01/Jan/2003:09:34:01 -0400] "GET /medios/banner/ads.jpg HTTP/1.1" 200 23818
164.77.100.19 [01/Jan/2003:09:34:02 -0400] "GET /medios/banner/contrate.gif HTTP/1.1" 200 3458
164.77.100.19 [01/Jan/2003:09:34:02 -0400] "GET /imagen/flecha.gif HTTP/1.1" 200 49
164.77.100.19 [01/Jan/2003:09:34:02 -0400] "GET /medios/botones/click.gif HTTP/1.1" 200 823
164.77.100.19 [01/Jan/2003:09:34:04 -0400] "GET /imagen/pixel.gif HTTP/1.1" 200 49
164.77.100.19 [01/Jan/2003:09:34:04 -0400] "GET /medios/banner/ads2.jpg HTTP/1.1" 200 14527
```

En el ejemplo citado en el cuadro 4, la página e imágenes registradas provienen de un requerimiento del cliente con IP 164.77.100.19.

El inicio de los registros muestra la llamada a un documento web correspondiente al *index.html*. Esta página pudo ser accedida por referencia desde otra página a través de un link o bien directamente por digitación en la barra de dirección del navegador. La carga siguiente es un archivo adjuntado al documento anterior correspondiente a una hoja de estilos que contiene información de los formatos según el tipo de etiqueta que dará a la página - si por ejemplo el formato para una etiqueta de tipo td (celda de una columna) es "*font: 12px*" indica que para todas

las celdas de cualquier tabla se aplicara una fuente de tamaño 12 pixeles. Seguido se agrega la carga de un archivo JS que corresponde a códigos *javascript* que le dan funcionalidad a la página o modificaciones de forma, donde por ejemplo se encuentran eventos del Mouse como clicks o scroll para cambiar imágenes, o puede tratarse de validación de formularios, por ejemplo rut o campo no vacío. Cargada la hoja de estilos y funciones comienzan a cargarse las imágenes que aparecerán en el despliegue del *website* y que complementan la estructura final que será vista por el usuario. Estas imágenes se cargarán de acuerdo a la aparición y traducción de etiquetas de imagen cuando el navegador comienza a procesar y desplegar el documento html.

En el ejemplo mostrado anteriormente se observa que para un requerimiento se hizo el llamado a 15 objetos que componen la página, con lo que finalmente se registraron 15 líneas de log. Por lo tanto, si se toma un alto tráfico de clientes sobre un sitio web de muchas páginas con alto apoyo de imágenes, sonidos y/o videos, se obtiene una gran cantidad de registros para cualquier período de estudio.

Como se mencionó en la sub sección 2.2, la importancia de los contenidos en el análisis de este estudio se encuentra en los registros de los *web logs* y los contenidos de cada una de las *web pages* del *website.* Luego, los registros de archivos anexos como hojas de estilo (CSS) o funciones como *javascripts* y/o *vbscripts* (JS), además de las imágenes que contiene el archivo (JPG, GIF, PNG, ICO) y posibles archivos de multimedia como videos o sonidos (MPG, AVI, MP3, MIDI, WAV) no son parte esencial de la investigación de contenidos textuales y búsqueda de las *website keywords*. Luego, cada registro de estos, como se muestra más adelante dentro del proceso de sesionización, es identificada y depurada.

Un campo que entrega información que permite también depurar información es el "id de respuesta del servidor". El campo muestra, junto con la información del requerimiento, el mensaje adjunto del servidor que puede corresponder a una respuesta exitosa o bien, error en el requerimiento. Como indicador de la calidad de la estructura de los sitios web, este mensaje del servidor permite identificar requerimientos de documentos y/o archivos que no existen o links erróneos o "rotos". A nivel de mantención y administración del servidor, esta información es valiosa pues permite mantener todos los archivos y links bien conformados, pero para propósitos de la investigación, estos registros no prestan utilidad pues no es el resultado del requerimiento del usuario, sino un intento infructuoso de obtener algún contenido.

Existen programas automatizados que analizan la web llamados *web spiders* y *web crawlers* cuya tarea es indexar páginas en alguna base de datos o motor de búsqueda, validar códigos HTML, validar links, monitorear cambios de contenido y en algunas ocasiones generar de respaldo de páginas o *mirroring*. La aplicaciones a cargo de este proceso generan un llamado a la página de inicio o *home page* del website y extraen el contenido para indexarlo o almacenarlo. Junto con lo anterior, identifican links a otras páginas del website dentro del mismo contenido y se repite el proceso. Estos constantes requerimientos de las aplicaciones mencionadas son registrados en los web logs siendo el campo agente el que contiene el string que permitirá identificar posteriormente cual fue el robot que hizo dicho análisis el cual es generalmente comparado con un listado de robots disponible[13] en la red. Este listado es un documento que se encuentra disponible de forma gratuita y que se basa en un estándar de exclusión de robots. Se consensuó la construcción de este estándar entre 1993 y 1994 por varios propietarios de robots

[13] El listado de robots disponibles se puede descargar desde la url *http://www.robotstxt.org/wc/active/all.txt*

con el fin de contar con un listado validado de estos motores evitando que se haga el impedimento de analizar las páginas web y, por otra parte filtrar aquellos robots cuyo funcionamiento de forma maliciosa pretende dar de baja páginas por ataques de múltiples requerimientos, o por envío de códigos de requerimientos que causan la alteración o caída del servidor.

El paso final del preprocesamiento es la identificación de las sesiones reales "humanas" que se identifican en el *web log*. Este proceso se llama ***sesionización*** y corresponde al proceso de reconstrucción de sesiones desde los *web logs*.

3.4.2.1.1. Detalle de Componentes del Web Log.

Para un correcto preprocesamiento de un *web log* es importante que se encuentren los campos mínimos necesarios para realizar la sesionización por tiempo de navegación. Estos campos permiten dar una consistencia a la reconstrucción de sesiones y además contienen información relevante para los siguientes pasos del estudio. Se indican a continuación los campos que componen el *web log* (Campos mínimos) con la respectiva relevancia que tiene cada uno de ellos en la reconstrucción de las sesiones.

- **IP del host**: Este campo almacena la dirección IP del HOST o usuario que hace el requerimiento al *website*. Este campo puede permitir en algunas ocasiones identificar a un usuario único que accede al website. Esto sucede generalmente con los casos de usuarios que tienen una IP fija, pero que actualmente es poco frecuente por la utilización de IPs dinámicas por parte de los ISP y por lo tanto, poco probable la utilización de este dato como identificador en el estudio de las sesiones. En algunos casos la ip del host que genera el evento de navegación corresponde a la IP de un *web proxy* o de un *firewall*, los cuales habitualmente pertenecen a empresas o ISP que utilizan estas herramientas para proteger la privacidad de sus empleados y usuarios respectivamente.

- **User y User id**: Estos campos son para identificación del usuario que ingresa a un *website* que se encuentra con un proceso de autentificación. Como se trata de un análisis de sesiones con la heurística de tiempo, este campo no se hace necesario en el análisis. También, este campo no se utiliza con mucha frecuencia ya que el usuario evita aquellos sitios web que requieren de una autentificación previa para ver sus contenidos. La utilización de este campo es principalmente para casos de *websites* que requieren un nivel de seguridad adicional en sus páginas, por ejemplo bancos, intranets, etc.

- **Fecha y hora**: Este campo es uno de los principales dentro de los registros pues es con este que se generan las diferencias de tiempo de navegación entre páginas, así como también permite identificar cuanto es el total de navegación y en que momento dejo de navegar el usuario. El campo es de formato dd/mm/aaaa:hh:mm:ss gmt y contiene la fecha en que se hace efectivo el requerimiento del usuario.

- **Formato de solicitud, ruta y versión de la respuesta**: El campo principal en esta tupla de información es la ruta del documento que entrega el servidor como respuesta al requerimiento del usuario. Se conoce a priori que un usuario hace una solicitud de un sitio web el cual esta compuesto por varios objetos dentro de los que se encuentran: textos, imágenes, animaciones, multimedia, documentos por nombrar los principales. Por lo tanto, es posible con este campo identificar cuales son solamente las páginas navegadas por el

usuario y cuales archivos se pueden descartar, por ejemplo imágenes (gif, jpg, bmp, tif), sonidos (wav, midi, mp3), video (mov, mp4, mpeg, avi, asx), documentos (pdf, doc, xls, ppt, txt,wri) y scripts (js, css, vbs) y dejar solamente las páginas web navegadas por el usuario (html, htm, asp, php, aspx, dhtml, shtml).

- **Id de respuesta del servidor**: La respuesta del servidor, es un código de resultado respecto del éxito de la solicitud recibida por el servidor ante un requerimiento del usuario. Este código permite obtener cuál es la respuesta final e identificar si la solicitud fue exitosa o bien si se obtuvo un error. La forma de respuesta de este campo es un código que puede ser interpretado como un mensaje resultante del requerimiento. Los principales códigos y grupos de mensajes son:
    - 200: respuesta exitosa del requerimiento del usuario. Quiere decir que el requerimiento dio como resultado un documento y que éste esta siendo devuelto para despliegue en el *browser.*
    - 3XX: son respuestas que retornan un documento pero que es necesario redireccionarlo, es decir, cambiarlo de la ruta original a una nueva donde se encuentra el documento requerido.
    - 4XX: Errores de acceso al sitio web. Algunos casos frecuentes son:
        - 403: Página o acceso prohibido: ocurre cuando un sitio Web solicitado requiere autentificación previa o bien se trata de ingresar a una carpeta de acceso restringido.
        - 404: No se encontró la página web. Respuesta del servidor a un requerimiento en que no se encuentra la página Web o se digita de forma incorrecta en el navegador.
    - 5XX: Respuesta de error del servidor cuando se producen problemas de compilación de código contenido en *web pages*. Es más frecuente la aparición de este error cuando se utilizan páginas dinámicas como asp, php o jsp.

- **Bytes**: Componente que permite identificar los bytes enviados en respuesta por el servidor ante el requerimiento. Corresponde al tamaño del documento que se envía como respuesta. Este campo permite identificar los volúmenes en tamaño de las transacciones de los usuarios.

- **Página desde la que se hace el requerimiento:** El requerimiento realizado de un sitio web debe tener un origen que puede ser digitación directa en la barra de direcciones del navegador o bien a través de un vínculo o link presente en otra página o documento. El campo en cuestión almacena dicha información como *referrer*, de manera que si existe información en este campo, quiere decir que hubo origen desde un link. Este dato es un buen referente para localización de fuentes de origen de acceso al *website*.

- **Información sobre versión del navegador**: Corresponde a la identificación del agente con que esta haciendo los requerimientos el usuario (navegador, motor de búsqueda, agente). Esta información detallada es una de las principales en un *web log* cuando se desea realizar una sesionización reactiva. Un usuario de un *website* tiene un agente o navegador que puede fácilmente contener actualizaciones o *plugins* que lo mejoran o les presta mayor

utilidad. Con esta data es posible detectar navegadores únicos y, por lo tanto, obtener una sesionización más precisa y cercana a la realidad concatenándolo con la ip de origen y la fecha del requerimiento.

- **Otros campos**: Los servidores de Internet generan un *web log* con campos por defecto. A esta configuración se puede añadir información de otra índole que pueden agregar valor al contenido del registro. El más importante y significativo de los campos adicionales es el *querystring* que es una línea de código que va después de la dirección o path del documento o página web separados por el símbolo "?" y que en su *string* contiene información de variables traspasadas de una página a otra mediante un *submit* (botón de formulario o link que pincha el usuario) y que es traducida por la página de destino para desplegar información en base a esa variable. Este tipo de variables es mayormente utilizado en páginas dinámicas (extensiones asp, php, jsp, cfm) las cuales son un tipo de plantillas que al momento de leer la variable traspasada genera el contenido de la página desplegada. Por ejemplo, una página *contenido.php* se alimenta de una variable de nombre *ID* de tipo entera. Si *ID* es 1 desplegara el contenido bajo el *ID* 1, si es *ID* 2 desplegará el contenido bajo el ID 2 y así sucesivamente. Este tipo de páginas son las llamadas ASP: Active Server Pages.

3.4.2.1.2. Ordenamiento y unión de archivos *web log*

Ante un requerimiento del usuario se obtiene un registro por cada objeto devuelto por el servidor, por lo que el volumen de datos generado por un día de transacción, puede ser enorme. El control de estos casos se realiza mediante la generación de un archivo de *web log* a diario (por ejemplo, archivo log *fecha.log*) o bien para un período de tiempo determinado (por ejemplo, archivo *fechainicio_fechafin.log).* De esta manera es necesario que la data sea consistente en los campos que se generan, es decir, que no exista un descuadre entre campos y que por otra parte se haga una unión de todos los archivos obtenidos y que están dentro del análisis.

El proceso de ordenamiento y unión hace que todos los registros en diferentes archivos sean consolidados en uno sólo lo que permitirá realizar un proceso de reconstrucción de sesiones en un continuo de registros hechos por el servidor.

3.4.2.1.3. Codificación del archivo y la base de datos

Los textos extraídos desde un documento vienen de acuerdo a cierto formato o set de caracteres. Este set es el resultado de una configuración de tipo de idioma del texto utilizado en el *web log*. En algunos casos el tipo de caracteres del archivo de origen y del usuario que utiliza la sesionización no son compatibles y no se puede por lo tanto obtener comparaciones entre registros y por ende la sesionización puede resultar errónea. Se requiere de alinear los sets de caracteres utilizados para que los cruces y comparaciones sean consistentes a través de la transformación a un set en común. El más habitual es UTF-8 o bien Windows (ISO).

3.4.2.1.4. Reconstrucción de sesiones de un sitio web. Sesionización.

El siguiente paso después del ordenamiento y verificación de los archivos de log es la

sesionización [10] de los datos, proceso que se genera siguiendo los siguientes pasos.

- Se hace uso de una base de dato relacional por la cantidad de registros que se pueden llegar a manejar y por el fácil manejo de altos volúmenes de información. En el modelo relacional es necesario crear dos tablas principalmente:
    - *Web Log*, que es la información que almacenan los logs íntegramente cargados en una tabla.
    - *Web Log Clean*: que es la tabla que contendrá los datos depurados y sesionizados.
- La carga de los web data se realiza a medida que se van sesionizando los diferentes archivos. En teoría, en un equipo de alta capacidad, el proceso de consolidación podría hacerse en un solo paso para luego sesionizar, sin embargo las limitantes de capacidad de respuesta y procesamiento de los equipos hacen que sea más efectiva la sesionización por archivo. Es decir, se carga un archivo, se sesioniza y luego se pasa al siguiente. Este proceso se repetirá tantas veces como archivos del *web log* sean detectados.
- El siguiente paso es el detectar todos los registros que pueden generar error [26] al momento de intentar generar una sesionización los cuales pueden ser:
    - Respuestas de Error: al momento de hacer un requerimiento el usuario este puede caer en una página que entregue una respuesta en error, o bien que no se encuentre la página. Estos registros son marcados y luego ignorados del conteo de tiempo de las sesiones.
    - Archivos de Multimedia: Dentro de los objetos devueltos por el servidor se encuentran los archivos de media (imágenes, sonidos, animaciones, videos y documentos). Estos suman registros y tiempos en la navegación pero no son parte dentro de este estudio del tiempo invertido por el usuario al momento de cargar el contenido de un sitio web por lo que al igual que los errores son marcados y posteriormente omitidos en el análisis.
    - Robots o Crawlers: Las direcciones IP y/o agentes detectados en los *web logs* que corresponden a robots son eliminados del análisis dado que es una suma de navegación de un usuario "virtual" que no tiene comportamiento asociado.
- Casos especiales: Direcciones IP no válidas como usuario desde el *querystring*. Cuando se tiene como dato del registro el *querystring* se puede obtener información adicional y que puede ser relevante al momento de sesionizar. Si por ejemplo es sabida la existencia de una variable que indica el origen del requerimiento (ejemplo: origen=sucursal) se puede marcar dicha fila y eliminar al momento de sesionizar. Lo mismo sucede cuando a priori se conocen las direcciones IP que corresponden a equipos que son de la empresa o de algún segmento de direcciones IP que se conocen como no válidas para los análisis las que son eliminadas.
- Como se mencionó con anterioridad, es posible que las direcciones IP no sean del cliente directamente sino que sean de un *firewall* o *web Proxy* que provee la empresa o ISP del usuario que navega. Mediante el "agente" o browser es posible

diferenciar en un mayor grado a los usuarios, por lo que este paso se trata de generar una columna con una "llave" de la tabla de datos compuesta por: Fecha + Agente + IP.

Ya generada las columnas necesarias se da inicio al proceso de sesionización el cual tiene como objetivo reconstruir la mayor cantidad de sesiones desde un archivo de *web log* mediante la utilización de un diferencia de tiempo entre sesiones de 30 minutos como máximo [7].

El resultado de la sesionización es una tabla de web log reconstruida y depurada la cual agrega los campos identificadores de sesión y tiempo de navegación con lo que ya es posible determinar parte del comportamiento del usuario en el sitio web.

3.4.2.2. Preprocesamiento y Selección de *web pages*.

En las páginas web la acción de preprocesamiento y selección se enmarca en la obtención de una "imagen" del sitio web al momento de realizar el estudio y análisis. La idea de esta foto es almacenar el contenido y estructura exacta que los usuarios visitaron y navegaron con el fin de que sea consistente la relación entre registro en el *web log* y *web page* visitada. Anexo a este procesamiento y de acuerdo a este trabajo, se puede omitir la extracción de imágenes, archivos y documentos complementarios como hojas de estilo (CSS), Javascripts (JS), documentos adjuntos o para descarga (PDF, DOC, XLS, PPT, PS) ya que si bien son una componente existente de la página web, no son parte del estudio de contenidos que se realiza.

Como resultado se obtiene finalmente un set de páginas html, htm, asp, php y jsp que contienen una estructura de etiquetas y contenidos de texto libre. Un preprocesamiento posible para cada página, es la extracción de las etiquetas para dejar los textos libres y sin formato. Sin embargo, el proceso de tokenización, a explicarse en la sub sección 2.4.2.3, requiere conocer con anterioridad de la limpieza de los contenidos el tipo de etiqueta utilizada con el fin de determinar las palabras que son solamente normales y aquellas que son especiales.

3.4.2.3. Extracción de palabras de un sitio web. Tokenización.

El trabajo de transformación de un texto en un set de palabras es llamado tokenización. Como parte importante del análisis a las páginas web [6], se debe realizar este proceso para identificar el set de palabras que conforman el contenido del documento.

La tokenización del contenido de una página web se inicia extrayendo el contenido existente dentro de cada página, comenzando por la inicial o *home page* de un *website,* pasando por todos los documentos que se encuentran enlazados y haciendo este proceso de forma recursiva pero siempre dentro del dominio de análisis ya que pueden existir vínculos a páginas externas que no son parte del proceso de investigación y por ende al proceso de tokenización. El proceso de extracción debe contener la totalidad de las *web pages* ya que se debe contemplar las preferencias de todos los usuarios y el comportamiento que los acompaña. La forma de extracción como proceso se explica a continuación.

- **Generación de Link de Acceso**. El primer paso para el proceso de tokenización es establecer cual será la página de inicio o *home page* desde donde se extraerán las palabras. Este hito es importante dentro del proceso ya que es posible iniciar el proceso desde niveles superiores que no necesariamente llegan nuevamente a la

página de inicio. Un ejemplo de estos casos son los inicios de portales. El sitio Web raíz contiene links a sus principales portales de acceso, sin embargo, desde uno de sus portales no es posible acceder nuevamente al menú de selección inicial mencionado.

• **Fijación de Nivel Máximo**: El proceso de extracción de links se desencadena desde cada página extraída dentro del *website* de análisis. La fijación de un nivel permite detener el proceso por profundidad de links ya que por lo general, si el mapa del sitio es redundante, llegará un momento en que todos los links dentro del servidor no tendrán más links o comenzarán a apuntar a las mismas páginas anteriormente procesadas.

• **Localización de links detectados:** Existen dos tipos de vínculos dentro de un sitio web.

  - El vínculo o link duro que comienza con http como inicio del *string*, por lo que el acceso a este es directo.
  - Links que no contienen http por lo que es necesario establecer en estas circunstancias cual es el link de origen de esta página y se debe reconstruir nuevamente concatenando el origen del documento encontrado con el link en análisis.

- **Creación de arreglo de links**: Con lo anterior, dado que lo que importa es el contenido y no el orden de extracción en este análisis, se extrae, carga y ordena el arreglo contenedor que posteriormente será la base para el proceso de extracción de los contenidos.
- **Extracción de contenidos**: El proceso de extracción obtiene la respuesta que entrega el servidor (por ejemplo 200 para el caso que es correcto o 404 cuando no se encuentra el sitio web). De esta manera se restringe la extracción de textos a las respuestas correctas del servidor y no a páginas de error. Luego del encabezado, se encuentra el contenido que es la colección de tags y textos de cada una de las páginas analizadas de forma íntegra. Este contenido se almacena en una variable pues es necesario pasarla posteriormente por una serie de análisis y filtros para limpiarla.
- **Limpieza y filtrado de palabras**: Luego del proceso de extracción de contenidos, es necesario llegar a una palabra "limpia", sin etiquetas que la rodeen, por lo que se genera a continuación el listado de limpiezas y filtros que se realiza a cada contenido.
- **Extracción de scripts (<script></script>) y estilos (<style></style>)**: Un proceso previo a la extracción de la totalidad de etiquetas es la extracción de los scripts y hojas de estilos de un sitio web ya que si se realiza posterior a la extracción de las etiquetas, queda como contenido del sitio lo que se encontraba dentro de estas marcas, por ejemplo:
  - Antes de la extracción de etiquetas se tiene dentro del contenido: <script>function hola(){print('saludo');} </script>.
  - Después de la extracción de etiquetas la parte del contenido afecta queda:

function hola(){print('saludo');}

- Luego, se genera un set de palabras que es almacenada como parte del contenido: function, hola, print, saludo.

Como se observa en el ejemplo, de forma innecesaria y errónea se agrega un set de palabras que no es parte del contenido de interés del usuario, sino una componente de apoyo a la estructura del funcionamiento del *website*.

- **Extracción de etiquetas**: Luego de extraer los scripts y hojas de estilo se puede proceder a quitar las etiquetas existentes dentro del sitio web. En PHP, la función strip_tags se encarga de quitar todas los etiquetas que se encuentran dentro de un documento o página. Hay que considerar que las etiquetas son eliminadas junto con su contenido interno que en algunas ocasiones es información adicional de la etiqueta.

- **Eliminación de caracteres no textuales**: Dado que la separación de las palabras de un contenido se hace en base a los espacios que hay entre las palabras, en algunas ocasiones quedan caracteres adyacentes al texto que vienen de acuerdo al contexto de la página. Por ejemplo, cuando se separa el contenido "(entre paréntesis)" las palabras extraídas son "(entre" y "paréntesis)". Luego es necesario eliminar este tipo de caracteres de las palabras extraídas, para el caso de este ejemplo el inicio del paréntesis de la palabra entre ("(") y el fin de paréntesis de la palabra "paréntesis" (")").

- **Eliminación de caracteres especiales HTML**: Dentro de la generación de las páginas Web se utilizan algunos códigos especiales para los casos de caracteres que no son comunes en un contenido verbal. Por ejemplo, las vocales acentuadas dentro de una palabra quedaban entre un signo "&" y un "acute;" con lo que el browser al momento de leer el contenido interpretaba estas vocales como una vocal acentuada. Al igual que con los caracteres no textuales es necesario identificar todos los códigos que tienen esa nomenclatura para ser extraídos o bien reemplazados de manera que el texto quede directamente escrito para su lectura y no mediante un código html que no será entendido en la investigación. Dentro del análisis se detectaron los siguientes caracteres especiales: "©"(©), "á"(á), "é"(é), "í"(í), "ó"(ó), "ú"(ú), "&"(&), "¿"(¿), """("), "ñ"(ñ).

- **Separación de palabras**: finalmente con los contenidos limpios y sin formatos se inicia la separación de palabras de acuerdo a los espacios existentes entre ellas. Estas palabras son almacenadas en una variable para luego stemizarlas. Dentro de este proceso es posible que se pierdan cierto número de palabras puesto que en los procesos de depuración se eliminan los signos, que en algunas ocasiones separan estas palabras. Por ejemplo, se encuentra dentro de un set de links de la página la siguiente línea: home|banco|contacto|salir. Al eliminar los caracteres no textuales se elimina el carácter "pipe" (|), luego, la palabra cargada es de la forma: homebancocontactosalir, por lo que esas 4 palabras quedan ilegibles dentro del texto. Para resolver este inconveniente se prefirió reemplazar los caracteres no textuales por espacios ya que la utilización de estos es arbitraria generalmente regida por las

líneas de diseño que son aplicadas en el *website.*

- **Stemización [22]:** El proceso de stemización busca la forma más cercana a la raíz de la palabra según diversos criterios en un algoritmo[14] de ordenamiento. Con la stemización las palabras cuando tienen una raíz similar quedan bajo un nombre parecido y en un mismo formato ya que elimina conjugaciones, acentos y palabras cortas. Este paso permite a su vez agrupar familias de palabras dentro de un mismo texto. Por ejemplo la palabra cred es utilizado para *crédito*, *credito* y *crediticio*.
- **Carga final del archivo:** Finalmente se carga dentro de una base de datos el set de palabras extraídas desde el sitio web con su respectiva stemización.

3.4.3. Transformación de Datos.

La transformación dentro del proceso de KDD busca tomar los datos obtenidos por el preprocesamiento y selección para dejarlos en una representación que permita obtener información adicional o bien para generar una entrada a procesos de análisis y algoritmos de web mining.

3.4.3.1. Transformación de datos en web logs.

En los *web logs* la transformación realizada es la generación de vectores desde de comportamiento del usuario desde el archivo sesionizado. Corresponde a la identificación de las sesiones versus sus preferencias de navegación y tiempos de navegación. El proceso de transformación toma los registros de los *web log* del tipo Pagina Requerida, Id de Sesión y Tiempo de Navegación desde el archivo de *web logs* preprocesada y la transforma en el vector $v = [(p_1, t_1)...(p_n, t_n)]$, donde (p_i, t_i) son los parámetros que representan la i-ésima página del visitante y el tiempo gastado en ella en la sesión, respectivamente. En esta expresión, p_i es el identificador o etiqueta de la página.

La ecuación anterior enfatiza que el comportamiento del usuario en el sitio web viene dado por:

- La Secuencia de Páginas: que corresponde a las páginas visitas y registradas en los archivos logs.

---

[14] El Algoritmo de Porter: Porter publicó en 1980 un algoritmo para el método de Stemming que fue tomado como base por muchos investigadores. El algoritmo lee un archivo, toma una serie de caracteres y de esa serie, una palabra; luego la valida verificando que todos los caracteres involucrados sean letras, de ser así, aplica Stemming sobre ella.:
La aplicación de Stemmer consiste en hacer pasar esta palabra a través de varios conjuntos de reglas, cada conjunto de reglas está formada por n reglas y cada regla por::

1. Un identificador de regla.
2. El sufijo a identificar.
3. El texto por el cual debe ser reemplazado al encontrar el sufijo.
4. El tamaño del sufijo.
5. El tamaño del texto de reemplazo.
6. El tamaño mínimo que debe tener la raíz resultante luego de aplicar la regla (esto es a los efectos de no procesar palabras demasiado pequeñas).
7. Una función de validación (una función que verifica si se debe aplicar la regla una vez encontrado el sufijo)

- Contenido de la página: contenido que es una mezcla entre texto libre y archivos multimediales (imágenes, sonidos, videos etc.).
- Tiempo invertido: que corresponde al tiempo utilizado por el usuario en cada página web. Se asume que la cantidad de tiempo invertido en la página es directamente proporcional al interés que el usuario presenta en ella.

La combinación de los aspectos anteriores [34] son los que entregan el Vector de Comportamiento del Usuario o UBV por sus siglas en inglés y son un buen indicador para registrar los intereses del usuario en el *website*. Como parte del estudio, se establece que son las 3 primeras páginas las que más interesan al usuario dentro del *website* con sus respectivos tiempos de navegación y es el contenido de texto libre el que será estudiado como parte del análisis de intereses en contenidos.

3.4.3.2. Transformación de Datos en *web pages*.

El proceso de transformación de las web pages en un vector de características, tiene como paso previo y fundamental la realización de una limpieza de contenidos pues en las web pages aparecen gran cantidad de datos que no son de utilidad en la investigación (imágenes, tabla, objetos) o bien que son repetitivos [24]. El fin de este etapa, es reducir la cantidad de palabras, dado que no todas tienen el mismo peso y se requiere de un proceso eficiente. Para propósitos de realizar una presentación vectorial, se considera como R el número total de palabras diferentes y Q el total de páginas de un sitio web. La representación vectorial de las páginas web en su conjunto se muestra en la matriz de la ecuación 1.

$$WPV = \begin{bmatrix} m_{11} & \dots & \dots & \dots & m_{Q1} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ m_{1R} & \dots & \dots & \dots & m_{QR} \end{bmatrix}$$

Ecuación 1: Ejemplo matricial de Word Page Vector

Donde m_{ij} es el peso de la palabra i en la página j. Basado en *tfxidf-weighting* introducido en [24] los pesos son estimados como:

$$m_{ij} = f_{ij}(1 + sw(i)) * \log(Q/n_i)$$

Ecuación 2: Peso de la i-ésima palabra de una página web.

Siendo f_{ij} el número de veces que la palabra i se encuentra en la página j y n_i es el número total de veces que la palabra i aparece en el sitio web completo. El componente Q es como se mencionó anteriormente la cantidad de veces que aparece la palabra en la totalidad del *website*.

Existen palabras que tienen una mayor importancia que otras por encontrarse subrayadas, destacadas, en un link o por ser parte del título de la página web. Estos términos son los que constituyen las palabras especiales o *special words*, que corresponden a los términos en la página web que son más importantes que otras debido a que están marcadas de forma especial u ocupan un lugar importante en la página web, por ejemplo, el titulo de un texto o el título de la página web. La detección de estas palabras en el contenido de una página es haciendo uso de etiquetas HTML, palabras utilizadas por el usuario en la búsqueda de información y, en general, palabras que implican los deseos y necesidades de los usuarios. La importancia de palabras especiales es almacenada en un arreglo *sw* de dimensión R, donde *sw*(i) representa un peso adicional para la i-ésima palabra.

El arreglo *sw* "Special words" permite enfatizar la importancia de una palabra respecto de las otras. Esta distinción se utiliza cuando una palabra es marcada con alguna etiqueta, pertenece a otros sitios web relacionados o simplemente porque el constructor del sitio la consideró importante de destacar.

Los orígenes de las palabras especiales pueden ser diversos entre los que se encuentran:

1. **E-Mails**: Los correos electrónicos que los clientes o usuarios del sitio web envían a través del correo corporativo o de formularios contienen textos que son una fuente importante de palabras destacadas o de interés para el usuario. Sea $ew_i = w^i_{email} / TE$ el arreglo de las palabras contenidas en los e-mails, que también se encuentran presentes en el sitio web donde w^i_{e-mail} es la frecuencia de la i-ésima palabra y TE es la cantidad total de palabras en el grupo completo del arreglo de palabras de e-mail.

2. **Palabras de consultas**: Los sitios web de las organizaciones cuentan muchas veces con un motor de búsqueda propio que permite al usuario del sitio web asuntos específicos a través de la introducción de palabras clave. El almacenamiento de estas consultas pueden ser una buena fuente de palabras destacadas para el sitio web. Sea $aw_i = w^i_{ask} / TA$ el arreglo de palabras usadas por el usuario en el motor de búsqueda y que esta contenida en el sitio web, donde w^i_{ask} es la frecuencia de la i-ésima palabra y TA es la cantidad total de palabras en el grupo completo.

3. **Palabras destacadas en las páginas del sitio web**: En los textos libres del sitio web en análisis se encuentran palabras que son destacadas a través de tags específicos: letra cursiva, negrita, subrayada o fuente de tamaño grande. Esto con el fin de dar énfasis a algún concepto específico o idea del texto. Sea $mw_i = w^i_{marks} / TM$ el arreglo de palabras destacadas dentro de las páginas web, donde w^i_{marks} es la frecuencia de la i-ésima palabra y TM es la cantidad de palabras destacadas en el sitio web completo.

4. **Sitios web relacionados:** Usualmente un sitio web pertenece a un segmento de mercado, en este caso el mercado de las instituciones bancarias. Luego, es posible recolectar páginas de sitios web que pertenecen a otros sitios en el mismo mercado. Sea $rw_i = w^i_{rws} / RWS$ el arreglo con palabras utilizadas en el mercado de sitios web incluyendo

el sitio web bajo estudio, donde w^i_{rws} es la frecuencia de la i-ésima palabra y RWS es el número total de palabras en todos los sitios web considerados.

La expresión final $sw_i = ew_i + mw_i + aw_i + rw_i$ es la suma simple de los pesos descritos anteriormente.

3.4.4. Aplicación de *web mining* y descubrimiento de patrones.

Para el proceso de aplicación de *web mining* se utilizarán los algoritmos de *Kohonen* y *K-means* con el fin de clasificar el comportamiento de los usuarios y poder de esta forma detectar los contenidos de interés y por ende las *website keywords*.

Para el correcto y eficiente funcionamiento de las redes de aprendizaje se requiere del elemento de "medición" entre los vectores para poder generar las comparaciones y diferencias entre ellos en búsqueda de los patrones de comportamiento de los usuarios. Estas medidas de similitud y distancia se presentan a continuación. Seguido de lo anterior se hará mención de los algoritmos SOFM [15]de *Kohonen* y *K-means*.

3.4.4.1. Medición y/o comparación entre páginas.

- **Distancia**: la distancia entre las páginas en su formato vectorial permite obtener un indicador de cuan parecidas son las páginas web. Para ello se utiliza el coseno del ángulo formado entre las páginas (vectores) y se calcula de la siguiente forma:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} wp_k^i wp_k^j}{\sqrt{\sum_{k=1}^{R} (wp_k^i)^2} \sqrt{\sum_{k=1}^{R} (wp_k^j)^2}}$$

Ecuación 3: Coseno entre dos vectores

La ecuación (3) permite comparar a nivel contextual dos páginas web. Esta comparación entrega como resultado un valor numérico entre 0 y 1. Cuando la distancia calculada entre las páginas en medición es cercana a 1, quiere decir que son muy parecidas en contenido. En el caso contrario, indica que son diferentes en un mayor grado. Cuando 2 páginas son iguales retorna el valor 1, de otra forma, retorna 0. Un aspecto a considerar en la ecuación anterior es que cumple con el requerimiento de ser computacionalmente eficiente, lo cual la hace más apropiada para ser utilizada en algoritmos de *web mining*.

[15] SOFM – Self Organized Feature Map.

- **Medida de Similitud:** La medida de similitud se aplica sobre los vectores de comportamiento del usuario y se constituye según la siguiente fórmula:

$$st(\vartheta_t(\alpha),\vartheta_t(\beta)) = \frac{1}{t}\sum_{k=1}^{t} \min\left\{ \tau_k^{\alpha} \Big/ \tau_k^{\beta}, \tau_k^{\beta} \Big/ \tau_k^{\alpha} \right\} * dp(\rho_k^{\alpha}, \rho_k^{\beta})$$

Ecuación 4: Similarity Measure o Medida de Similitud de dos vectores de comportamiento de usuario

El primer elemento de la formula ($\min\{\cdot,\cdot\}$) corresponde al interés que tiene el usuario en la página visitada. Si para los usuarios α y β en la k-ésima página visitada es cercana a la otra, el valor de la expresión $\min\{\cdot,\cdot\}$ será cercano a 1, de lo contrario, será cercana a 0. Se anexa en la fórmula la distancia entre páginas en representación vectorial. La distinción que realiza esta formula es que páginas con contenidos similares son distinguidas según los intereses (tiempo invertido por página) de cada usuario.

Los cálculos anteriores permitirán para los algoritmos de *web mining* generar las mediciones y comparaciones para avanzar en el proceso de búsqueda de patrones.

3.4.5. Clustering de sesiones.

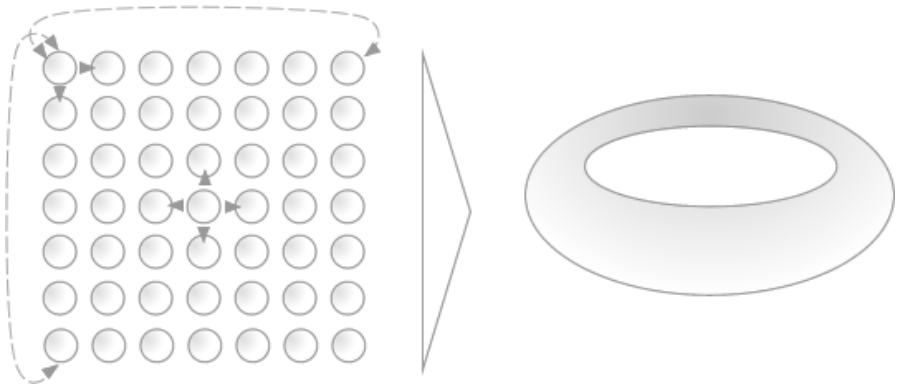
Posterior al proceso de preprocesamiento, selección y transformación de los datos, se inicia el proceso de minado, el cual, en este caso es realizado utilizando algoritmos de clustering, específicamente, mediante una red neuronal del tipo *Kohonen* (*Self Organized Feature Map*) y con el algoritmo de *K-means*. Ambos procesos obtienen su input desde los vectores de comportamiento del usuario (UBV) y la salida se espera que sea similar entre un proceso y otro en cuanto a clasificación final.

3.4.5.1. SOFM – Self Organized Feature Man.

La red neuronal de Kohonen, conocida como *Self Organized Feature Map* o SOFM es una red de neuronas bidimensionales y que se constituye por Vectores de comportamiento del usuario (UBV) obtenidos de forma aleatoria del listado completo generado y que es llamado vector de entrenamiento. La red de Kohonen toma una red cuadrada de dimensión n la cual es inicializada con neuronas aleatorias de la muestra y que posteriormente comienza el proceso de entrenamiento. Este proceso es de tipo no supervisado dado que la evolución de las clasificaciones se da de acuerdo a las modificaciones y operaciones que se realizan en el procesamiento de la información.

Existen diversas topologías asociadas a un mapa SOFM, sin embargo en este estudio se utiliza la topología toroidal [31], que quiere decir que la neurona que se encuentra en la primera posición del mapa es vecina es la ultima de abajo y de la primera y ultima de la derecha del mapa.

Figura 4: Proximidad de contenido importante de una página web en una red toroidal de Kohonen



Dada la significancia del vector de páginas importantes, y la composición por dos partes de este vector (página y tiempo), es necesario modificar ambos componentes de este vector cuando el entrenamiento encuentra una neurona ganadora y la modifica junto con sus vecinos.

Sea N una neurona de la red (mapa) y sea E el ejemplo de pagina importante a comparar en la red. El componente de tiempo del vector es modificado de acuerdo a una transformación numérica, es decir,

$$t_{i+1}^{N} = t_{i}^{N} * f_{\delta}$$

Ecuación 5: Modificación de la componente de tiempo en el procesamiento de una red SOFM

donde f_{δ} es el factor de ajuste del tiempo en el proceso, con i =1,..,n , donde "n", para el caso de este trabajo, es de valor 3.

La componente de página del vector de comportamiento del usuario, requiere de otro procedimiento de modificación para llevar a cabo el aprendizaje. Lo anterior requiere del cálculo de la distancia entre la página de la neurona a modificar y la del vector de entrenamiento en su i-ésima componente, que en este caso puede ser 1, 2 o 3.

Las distancias se obtienen durante el proceso de comparación neurona y vector de entrenamiento y se pueden definir como un arreglo, tal como lo muestra la ecuación (6). Siguiendo con la nomenclatura anterior, E corresponde a la neurona que se encuentra como input en el momento del entrenamiento y N la neurona a ser comparada. Con lo anterior se obtiene la expresión de distancias representadas en un vector con componentes numéricos:

$$D_{NE} = \left[dp\left(\rho_1^N, \rho_1^E\right), \ldots, dp\left(\rho_n^N, \rho_n^E\right) \right] \quad (6)$$

Ecuación 6: Distancias entre neuronas y páginas

Siguiendo la misma lógica de ajuste de tiempo se obtiene un nuevo vector que es ponderado por un factor de ajuste con lo cual se puede derivar una nueva expresión de distancias que corresponde a:

$$D'_{NE} = D_{NE} * f_\varepsilon \quad (7)$$

Ecuación 7: Factor de Modificación de distancia en el proceso de aprendizaje de SOFM.

Siendo f_ε el factor de ajuste de las páginas. Luego, con el nuevo valor de distancias es necesario encontrar el set de páginas cuyas distancias son similares y parecidas a D_{NE}. El ajuste final realizado para la neurona ganadora se obtiene de la expresión:

$$\rho_{i+1}^N = \gamma \in \Gamma / D'_{NE,i} \approx dp\left(\gamma, \rho_i^N\right) \quad (8)$$

Ecuación 8: Factor de ajuste de páginas

Con $\Gamma = \{\gamma_1, \ldots, \gamma_Q\}$ el set completo de páginas del sitio web, $D'_{NE,i}$ la i-ésima componente del vector D'_{NE}. Luego ρ_{i+1}^N es la menor distancia que reemplaza la i-ésima componente ajustado de las distancias de D_{NE}.

Para las vecindades, el cálculo de las distancias sigue la misma lógica de cálculo entre la neurona ganadora y la de entrenamiento originales, pero el ajuste realizado tanto para tiempo como para página se hace con un delta que disminuye de forma exponencial según la lejanía a la neurona ganadora. A la formula de modificación de distancia se agrega un Δ adicional y que generalmente corresponde a una función

$$e^{-\frac{k}{\lambda(t)}} \quad (9)$$

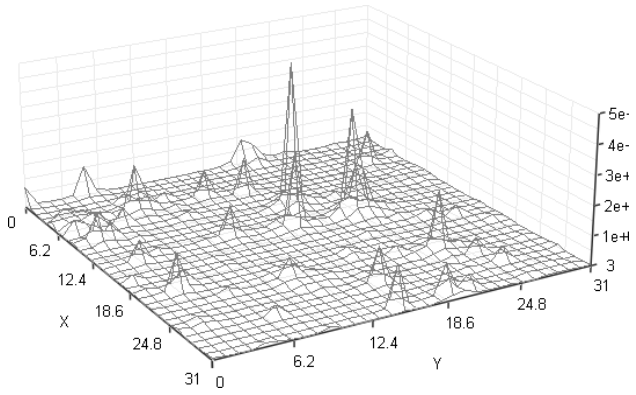
Ecuación 9: factor de corrección aplicado a vecindades.

Que pondera el factor de ajuste f del tiempo y las páginas. En la ecuación, k corresponde al radio de la neurona y $\lambda(t)$ es la tasa de aprendizaje en el *epoch* t del proceso.

El resultado final del proceso de SOFM genera como salida un mapa de neuronas con el proceso de modificaciones anteriormente descrito y que adicionalmente contiene un mapa de

conteo de veces en que una neurona resulta ganadora. En la figura 5 se observa un diagrama con el proceso de conteo, lo cual permite identificar, de forma gráfica, la posición de los clusters y su nivel de importancia según las veces que resultó ganadora.

Figura 5: Ejemplo de resultado de proceso de SOFM

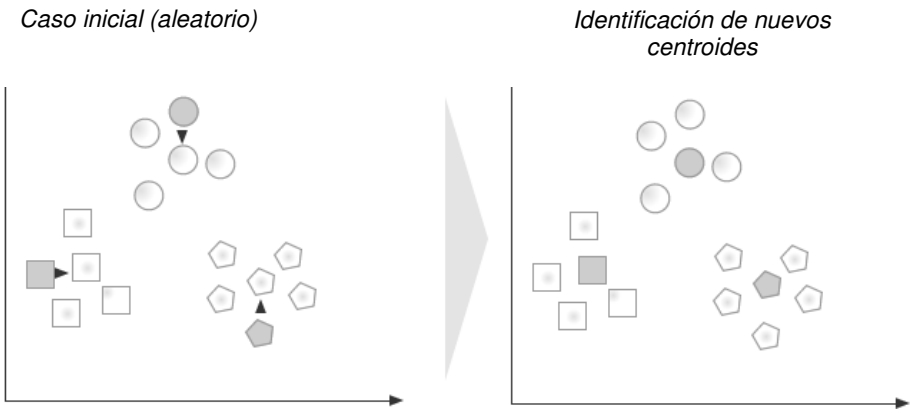


3.4.5.2. K-means.

El entrenamiento de K-means está dentro de los algoritmos supervisados, dado que requiere como input inicial la cantidad de centroides donde serán agrupados los diferentes miembros del archivo de entrenamiento.

El proceso de entrenamiento es simple, con respecto a SOFM en cuanto a pasos, pero tiene un mayor costo de procesamiento, dado que se requiere de hacer en un solo análisis del archivo de entrenamiento, la comparación de dos veces la cantidad de miembros del vector. La primera es para identificar a que centroide pertenecen y el segundo para detectar el centroide más representativo del grupo.

Figura 6: **K-means. Proceso de identificación de centroides.**



Los pasos en el proceso *K-means* son los siguientes:

- **Iniciación de Centroides**: El proceso se inicia indicando el número de centroides K, donde se ubicarán los miembros del archivo de entrenamiento en búsqueda de un mejor representante del grupo completo. Este valor puede ser dado de forma aleatoria o viendo el comportamiento de iteraciones con diferentes valores de K. En el caso de este estudio, se hace uso del número de clusters obtenidos de SOFM con el fin de buscar la semejanza que tienen los dos procesos donde mediante un proceso no supervisado se pueden identificar las clasificaciones que se realizaron en el proceso y por ende el número de clusters identificados.
- **Identificación de Miembros del centroide**: Para cada IPV del vector de entrenamiento se busca a cual de los centroides inicializados pertenece. Esto se hace mediante la medida de similitud indicada en la ecuación (4). Aquel centroide que tenga la mayor similitud será el que contenga el miembro identificado. Este proceso se genera con la totalidad del archivo de entrenamiento hasta que se logren agrupar todos los componentes de entrenamiento en los K centroides.
- **Identificación de mejor representante**: El proceso siguiente se realiza con el subconjunto de IPV del vector de entrenamiento que quedo relacionado al centroide en análisis. Se realiza la comparación de similitud entre cada uno de los miembros. Lo anterior se registra mediante el conteo de las veces que el miembro que esta siendo comparado con el resto, marca una medida de similitud alta comparada con la totalidad de miembros. Con lo anterior se obtiene un arreglo de miembros al centroide y un arreglo de veces en que ese miembro resulto tener una distancia menor en la comparación total del archivo. Con lo anterior se obtiene el nuevo candidato a centroide.
- **Reemplazo de Centroides**: El reemplazo de centroides tiene por objeto mejorar la situación inicial dada la identificación anteriormente descrita. Se obtiene entonces de acuerdo al análisis de los miembros de centroides, K nuevos representantes que son reemplazados por los originalmente inicializados. Con lo anterior se reinicia el proceso de análisis. La detención de *K-means* se da cuando la diferencia entre una iteración y la anterior es cercana a 0 en cuanto a cambios de centroides realizados.
- **El resultado final de K-means** son K centroides que corresponden a *Important Page Vectors* identificados según las iteraciones como los más frecuentes y mejores representantes de la clasificación sugerida por K. Estos vectores posteriormente serán utilizados para los trabajos de evaluación e interpretación que le siguen.

3.4.5.3. Evaluación e interpretación.

La etapa final de la extracción de patrones desde los web data, es interpretar los resultados obtenidos desde los algoritmos de *SOFM* y *K-means*. Para ello utiliza los *outputs* identificados de los respectivos entrenamientos para dar con lo que se llama las palabras claves de un sitio web o

website keywords.

La extracción de las palabras clave se hace utilizando la ecuación.

$$kw[i] = \sqrt{\prod_{p \in \varsigma} m_{ip}} \qquad (10)$$

Ecuación 10: Identificación de Keywords desde vectores resultantes.

Donde i = 1,…,R, *kw* es un arreglo que contiene los pesos para cada palabra relativa a un cluster dado y ς el grupo de páginas representando el cluster. Las palabras claves del sitio web son el resultado del ordenamiento de *kw* y de la detección de palabras con los pesos más altos, por ejemplo, las 10 palabras con mayor peso.

El proceso anterior se hace con los clusters identificados en SOFM para obtener las palabras frecuentes y de forma análoga se realiza el mismo procedimiento con K-means pero se utilizan los centroides para obtener las páginas visitadas más frecuentes.

Sobre el resultado obtenido es necesario aplicar preprocesamientos y limpieza dado que existen palabras identificadas que son frecuentes y no claves del sitio web dado que es principalmente la redundancia utilizada en el sitio lo que genera que algunas palabras sean frecuentes y no importantes. En estos casos aparecen palabras como “home”, “mapa”, “contactenos”, “ayuda” o bien los nombres de la organización dueña del sitio web.

CAPITULO 3

4. IMPLEMENTACIÓN DE HERRAMIENTAS PARA WEB MINING.

Como fue mencionado en la sección 2.2, los *web data*, principalmente los relacionados directamente con el comportamiento del usuario, corresponden a un gran volumen de datos que se almacenan como registros en los web logs del sitio web.

Se hace necesario recurrir a herramientas que permitan facilitar aquellos procesos que al realizarlos de forma "manual" requieren tiempo de trabajo, procesamiento y además comprensión dado que son diferentes los criterios que son aplicados al archivo al momento de realizar limpiezas y filtros. Es muy posible que bajo un proceso no estandarizado para esta etapa, genere errores o resultados deficientes por el gran volumen de información que se maneja.

Para las diferentes etapas de KDD existen diversas herramientas de preprocesamiento y selección que no son gratuitas y las que lo son, son deficientes y limitadas en los resultados que retornan. Las herramientas comerciales, además cuentan con el inconveniente que vienen desarrolladas para un estándar de trabajo en base al *data mining* y el proceso de aprendizaje de utilización de esa herramienta puede ir en contra del tiempo requerido para efectuar un estudio de un *website*.

Dado que el objetivo principal del trabajo es identificar las palabras claves de un sitio web y se cuenta con el obstáculo de que la utilización de herramientas no se acomoda a la investigación ni metodología realizada, es que se diseño un procesamiento de datos ad hoc y se decidió desarrollar las herramientas necesarias para la investigación como apoyo al trabajo realizado.

Dentro del proceso KDD, el proceso de *web mining* y análisis es el que mayormente involucra intervención humana puesto que se requiere de criterio y toma de decisiones para poder avanzar a la siguiente etapa o concluir el análisis del proceso. La implementación desarrollada abarcó principalmente las tres primeras etapas del proceso en cuestión. Según la etapa, se desarrollaron los siguientes módulos de trabajo:

Para la etapa de **Preprocesamiento y Selección**, se generaron módulos de sesionización o reconstrucción de sesiones y del tokenizador o extractor de palabras del contenido de una página web.

En la etapa de **Transformación** se generaron los módulos de generación de vector de comportamiento del usuario (*User Behavior Vector*) y el vector de páginas web (*Web Page Vector*).

En la etapa de **Web Mining** se desarrollaron los algoritmos de SOFM y K-means adaptados según la teoría de obtención de *website keywords* y de medición indicada en el capítulo 2.

Las especificaciones técnicas del equipo y software utilizado en el desarrollo de los módulos son:

Servidor Internet: Apache (http://www.apache.org)

Lenguaje de Programación: PHP (http://www.php.org)

Base de datos: MySQL (http://www.mysql.org)

La plataforma donde fueron desarrollados los módulos fue Windows XP Home Edition, pero dado que el software utilizado es *Open Source*, se pudo realizar el testeo tanto en este sistema operativo como en Linux.

4.1. Implementación para etapa de preprocesamiento.

El preprocesamiento y selección de datos es una etapa que requiere de rigurosidad y de tiempo puesto que el volumen de información es muy grande y la existencia de fuentes que provoquen error en la interpretación de los datos no es menor. La automatización de esta etapa trae como ventaja el desarrollar una herramienta automatizada que minimiza el error humano y genera una salida limpia de forma precisa y de acuerdo a lo estándares establecidos previamente a la generación del archivo de salida. Como entrada en esta etapa se recibe un archivo fuente que contiene la información pura desde el generador de *web data* y como salida se obtiene un archivo de iguales características pero libres de contenidos que induzcan el estudio a errores o malas interpretaciones. La generación de errores por parte de los datos se da principalmente por las siguientes situaciones:

- Convergencia de algoritmos: Muchas veces la precisión de los datos es fundamental para generar *clustering* y clasificación. Si existe un punto en que los datos convergen a nivel de medición, es posible que se trate de un error en el resultado y por ende un posible error en la interpretación final.
- Datos sin incidencia: Es la eliminación de datos que se encuentran dentro del *web data* ya que en la investigación no generan un aporte significativo. En el caso de este trabajo, la importancia del *web log* era aquellos registros con contenido (*web pages*). Luego, todos los registros con imágenes, sonidos, videos no cumplen con el criterio por lo tanto son eliminados.
- Los principales procesos necesarios para el análisis dentro del preprocesamiento y selección en *web mining* son la sesionización, que corresponde a la reconstrucción de sesiones desde el *web log* y la tokenización, que es la extracción y separación de contenidos desde un website.

4.2. Componentes complementarios para los módulos desarrollados.

Se desarrollaron funciones modulares para el caso de reutilización de código. Se generaron archivos separados para estos componentes de manera que si se requería en el proceso, se hacia la inclusión del archivo respectivamente para ser llamado dentro del código. Se detallan todos aquellos módulos de uso común para más de un módulo en la construcción de la herramienta de web mining.

4.2.1. Configuración de variables.

Como todo procesamiento de información existen diferentes formas de parametrizar las aplicaciones para obtener resultados diversos. Por lo anterior, se generó un archivo que almacena todas las variables utilizadas en los diferentes módulos. El módulo contiene información que es

almacenada en variables para luego ser utilizada por los diferentes procesos.

4.2.2. Configuración y conexión a base de datos.

El apoyo de una base de datos en los diferentes procesos da la facilidad en administración y gestión de información. Estos archivos son incluidos en cada módulo y contienen la información necesaria para conectarse con la base de datos generada para el análisis y generación de vectores.

- El archivo config.values.php almacena los valores de configuración y conexión de la base de datos.
- open.php contiene el *string* de iniciación y apertura de la base cuando se desea hacer una consulta.
- close.php contiene la información de cierre de la base de datos.

4.2.3. Análisis de archivos y directorios.

La componente de listado de directorios contiene la fuente que realiza lectura sobre una carpeta y enlista todos los subdirectorios y archivos que ahí se encuentran. Este módulo presta utilidad tanto para el sesionizador, que obtiene los registros desde una serie de archivos en una carpeta, así como también para la tokenización de un sitio web, el cual ha sido almacenado con anterioridad desde la red.

Existen otros archivos complementarios a los módulos principales, pero que serán mencionados dentro del análisis de los mismos ya que sólo son utilizados puntualmente por ellos.

4.3. Módulo de reconstrucción de sesiones. Sesionizador.

Una forma simple de definir una sesión, corresponde a la secuencia de páginas que visita el usuario en un sitio web durante un período de tiempo determinado. Cada requerimiento del usuario, y respuesta del servidor, quedan almacenados en los *web logs* de manera que se puede ver como se van "cargando" las páginas a medida que el usuario las solicita. De la misma forma, si el usuario solicita una página con documentos estos quedan almacenados dentro el *web log* y si solicita una página que no existe o ya no se encuentra el *web log* almacena un código de error en su contenido.

El módulo de reconstrucción de sesiones se llama sesionizador y genera un procesamiento de los *web data* que se encuentra en el *web log* para identificar las sesiones y generar un reconstrucción de ellas. Para realizar estos pasos se deben considerar los siguientes aspectos al momento de iniciar el proceso.

Web logs distribuidos: a cada solicitud de página del usuario se genera un registro en el *web log*. El volumen de estos archivos pueden ser de un tamaño enorme que por lo general es distribuido en varios archivos generalmente de acuerdo a la fecha en que se genera el log, por lo que usualmente se trabaja con tantos archivos de log como días hay en el análisis, luego se necesita unificar el archivo final para obtener el *web log* total del periodo de análisis.

Campos de un web log: Los *web logs* tienen campos estándar según NCSA y W3C para

almacenar los requerimientos, sin embargo, también existe la posibilidad de agregar información a los *web logs*. Un ejemplo es el campo *Server* que contiene la dirección o nombre del servidor que origina el requerimiento o el campo *Querystring* que contiene las variables almacenadas en el *string* de dirección de la página web. Estos campos por lo general, dada la especificidad del estudio, son removidos ya que al ser variables arbitrarias en estructura, requieren de un análisis a parte como para poder ser incluidas dentro de un análisis de *web mining* o estadístico. Como opción de trabajo en este aspecto, se sugiere estandarizar y ordenar los campos según el estándar solicitado de los *web logs*.

Normalización de Path de requerimiento: En ocasiones es posible que ciertos *websites* almacenen dentro del path de requerimiento de la página web algunas variables de información entre páginas, por ejemplo una ruta http://.../pagina.html puede ser también encontrada en el *web log* como http://.../pagina.html?variable=x que es una página probablemente muy similar con un contenido dinámico o que requiere de una variable para entregarlo. En este caso de preferencia se trabaja con páginas "estáticas". Luego, este tipo de condición en la dirección altera el resultado puesto que se pueden interpretar como páginas diferentes, por lo tanto, el path es normalizado eliminando todo contenido posterior al separador "?" que utilizan en el URL para identificar variables.

4.3.1. Modelamiento del sesionizador.

La base de la construcción del módulo del sesionizador se detalla según las tareas realizadas en el proceso.

4.3.1.1. Requerimientos.

Los web logs contienen un gran volumen de datos respecto de las páginas y objetos web en general, solicitados por el browser cuando el usuario navega por el sitio web. Estos logs son generados sobre un archivo no transaccional y quedan almacenados para ser analizados posteriormente. Existen muchos sistemas de análisis estadísticos de estos archivos, pero lo que se requiere es reconstruir las sesiones de los usuarios de forma reactiva, es decir, con los datos existentes identificar aquellas instancias en que el usuario navego en el sitio.

- La duración de la navegación se estima tiene un máximo de duración de 30 minutos.
- Son consideradas en la reconstrucción sólo páginas web. Se omiten imágenes, documentos y archivos anexos.
- Se deben eliminar aquellos registros con resultado de error por parte del servidor.
- Existen Crawlers o Robots que quedan registrados en los web logs. Estos no deben ser considerados como usuarios.
- El resultado debe ser una replicación del web log que debe sumar un campo de identificación de la sesión.

4.3.1.2. Caso de Uso Principal. Sesionización de un web log.

Caso de Uso: Sesionizar Web Log.

Participantes: Usuario, Analista.

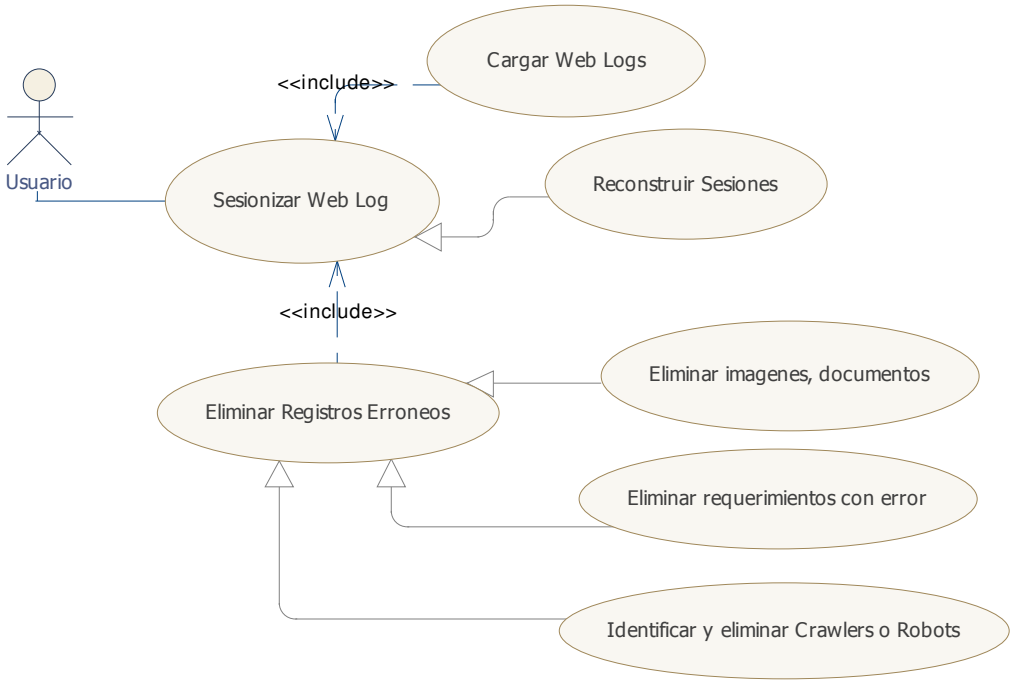
Propósito: Reconstruir sesiones de forma reactiva a partir de un conjunto de archivos web log.

Tipo: Primario y Esencial.

Descripción: Este caso de uso comienza identificando y cargando los archivos web log en la base de datos. Este proceso inicial almacena en una tabla los registros logs para un mejor manejo de los datos. En este mismo proceso el participante identifica las imágenes, sonidos, video, crawlers o robots y requerimientos erróneos devueltos por el servidor los cuales elimina. Desde los registros calcula los tiempos entre páginas web y luego genera el proceso de reconstrucción de sesiones de acuerdo al límite de tiempo indicado (30 minutos por sesión)

4.3.1.2.1. Diagrama de Casos de Uso para Sesionizador

Figura 7: **Diagrama de casos de uso para módulo sesionizador**



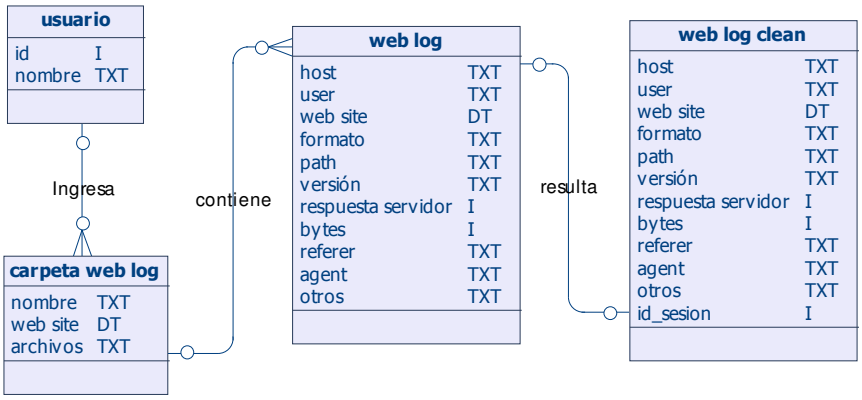
4.3.1.3. Modelo Conceptual.

El modelo conceptual permite identificar las principales entidades y sus atributos para el sesionizador. En este caso los participantes del proceso son el usuario, la carpeta donde se contienen los archivos a sesionizar y los *web logs* en análisis. Estos *web logs* son luego

trasformados a un *web log clean* que es un documento limpio y filtrado y que contiene los identificadores de sesiones resultantes del proceso.

4.3.1.3.1. Diagrama de Modelo Conceptual Sesionizador.

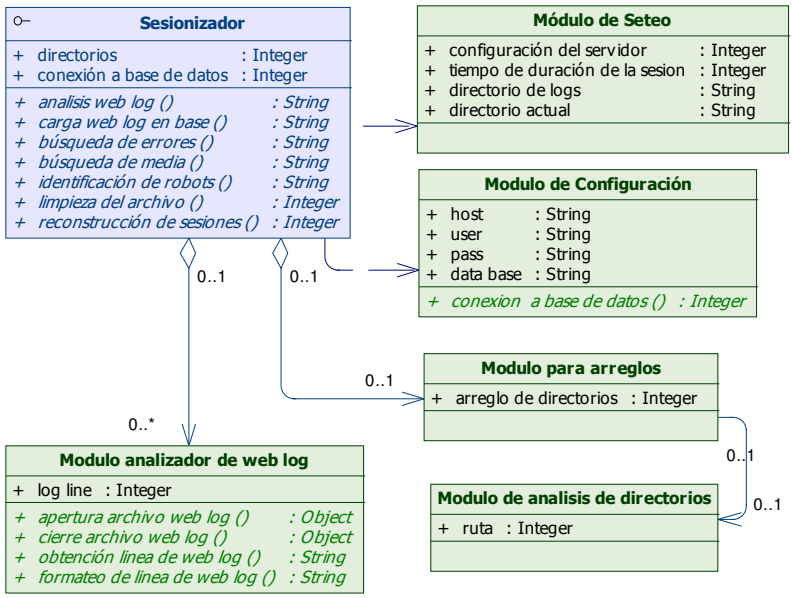
Figura 8: ***Diagrama de modelo conceptual para el sesionizador***



4.3.1.4. Diseño de clases.

La construcción del sesionizador y el resto de los módulos se realizó en php en conexión con una base de datos *MySQL*. El diagrama de clases muestra las componentes a construir o ya construidas y que serán utilizadas en el desarrollo del módulo. Se observan en el diseño de clases los módulos de inclusión de seteo y configuración que son parte común dentro de los módulos desarrollados. La inclusión más importante dentro de los módulos es el analizador de web log que es una clase que lee las líneas de un web log, las analiza y devuelve los campos respectivos.

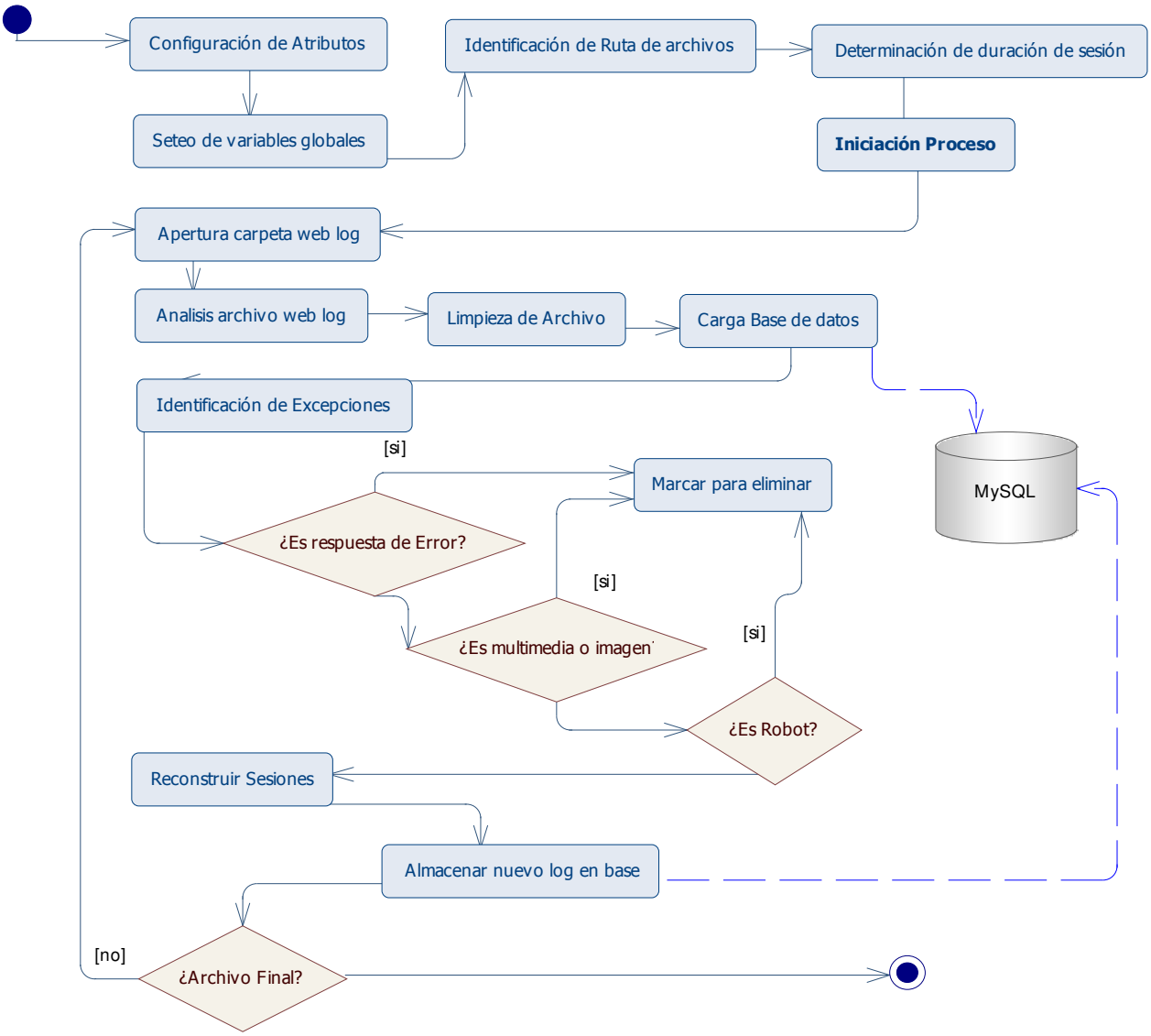
Figura 9: ***Diagrama de modelo de clases para el sesionizador***



4.3.1.5. Diseño del proceso de sesionización.

El proceso de sesionización se realiza de acuerdo al flujo del diagrama 10. Es relevante en el proceso la identificación de archivos con error y el preprocesamiento de la data.

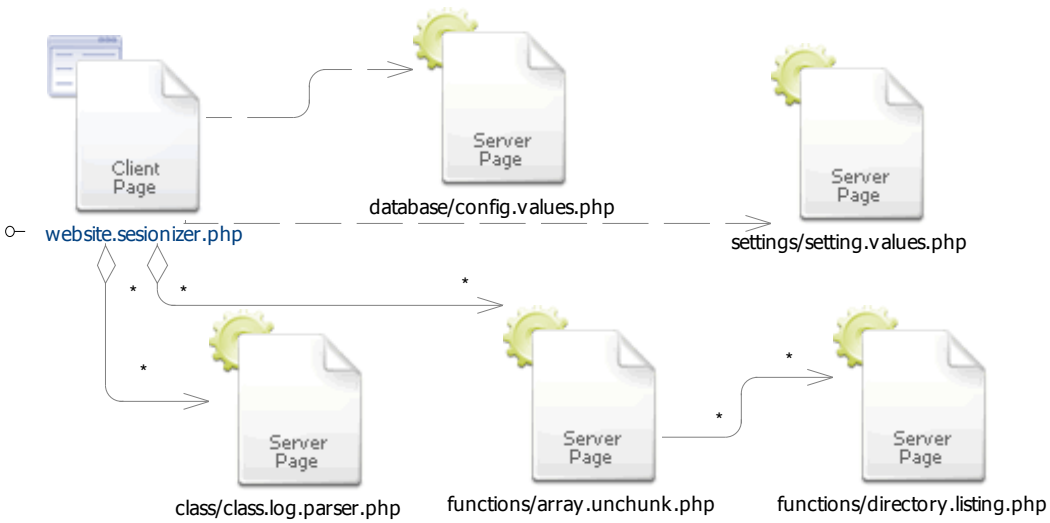
Figura 10: ***Diagrama de flujo párale proceso de sesionización.***



4.3.1.6. Diagrama de Conallen

Seguido de la composición de clases y del flujo a seguir, se construye el diagrama de Conallen para la implementación de las páginas con código para realización de la sesionización, con lo que queda el diseño de la siguiente forma:

Figura 11: ***Diagrama de Conallen para el sesionizador***



4.4. Módulo de extracción de palabras. Tokenizador.

El propósito del tokenizador es extraer el contenido textual de un sitio web, quitarle las etiquetas, obtener las palabras, obtener la raíz de las palabras y almacenar el resultado para análisis y procesamiento.

Como paso inicial y previo al proceso de tokenización se debe almacenar el sitio web en un directorio local, de manera que se tenga certeza que el contenido no cambiará con respecto a los registros de *web data* en el *web log* asociado. Como requerimiento del proceso de análisis se debe establecer como norma la no modificación del sitio web durante el período en estudio con el fin de no alterar el comportamiento del usuario en el *website* y de esta forma generar ambigüedades o información errónea por tratarse de contenidos diferentes al momento de ser visitado.

El proceso de tokenización se compone de los siguientes pasos:

1. Determinación de carpeta de análisis e inicialización: se debe indicar a la aplicación donde se encuentra la carpeta con las *web pages* para iniciar el proceso de extracción de contenidos. Luego, se da inicio al proceso de tokenización que comenzará a leer los archivos contenidos en el directorio mencionado.
2. Estructura de directorios, lectura de archivos: iniciado el proceso, comienza la lectura de *web pages* almacenadas en el directorio indicado en el paso anterior y los subdirectorios correspondientes. El proceso lee el archivo y almacena toda la estructura de contenidos en una variable la cual posteriormente pasará por los procesos de filtraje respectivo.
3. Limpieza de contenidos: seguido de la lectura y almacenamiento en una variable, comienza el proceso de limpieza de contenidos extraídos, lo cual significa:

    3.1. Eliminación de etiquetas HTML: como se indicó con anterioridad, la estructura de una web page es una combinación y orden de contenidos junto con etiquetas que le dan el formato requerido. Esta parte del proceso elimina las etiquetas dejando

prácticamente sin formato las palabras del contenido analizado.

3.2. Eliminación de caracteres especiales: como estándar de una página web se utilizan ciertas "configuraciones" para los casos de palabras especiales, por ejemplo, los acentos son incluidos en una palabra usandoel siguiente texto entre la vocal: &[vocal]acute;. Con lo anterior se genera el llamado acento, pero a nivel de contenido, no es identificable la palabra con ese texto añadido. Procesos posteriores a la limpieza remueven el & (amperson) de la palabra, pero en ese caso quedaría la palabra con el texto "acute" lo cual genera error o una mala interpretación y lectura de ella. Se necesita por lo tanto eliminar todos aquellos caracteres especiales contenidos en el texto. Se detalla en la siguiente tabla algunos caracteres especiales:

Tabla 1: ***Caracteres HTML Especiales***

Formato HTML	Carácter Correspondiente
©	©
á	á
é	é
í	í
ó	ó
ú	ú
&	&
¿	?
"	"
ñ	ñ

3.3. Eliminación de caracteres anexos a las letras o palabras: El español cuenta con una gran cantidad de caracteres que dan énfasis a palabras o frases. Este es el caso de los acentos, comillas, signos de interrogación entre otros. Estos caracteres son representados de la misma forma habitual en español y están siendo utilizado en las palabras. Este tipo de caracteres son extraídos de la palabra y reemplazados por la misma parte pero sin el énfasis del carácter. En este caso se puede perder contexto o significado de la palabra, pero en este caso el trabajo tiene esa limitante que debe ser analizada y estudiada para poder dar significado semántico a algunas palabras a las cuales se les elimina.

3.4. Stemización: Corresponde a la generación o identificación de la raíz de la palabra de la forma más cercana posible. Dado que muchas palabras tiene una misma escritura para los países de habla hispana, el proceso de stemización aproxima una palabra a su raíz. Al igual que en el caso anterior, se puede perder contexto de la palabra lo cual provoca desorientación en el análisis.

4. Almacenamiento de palabras: Finalmente, las palabras extraídas y stemizadas son almacenadas en una base de datos para realizar posteriormente la generación de pesos de las palabras por página para los análisis a realizar.

4.4.1. Modelamiento del tokenizador.

El módulo de tokenización trabaja sobre otra web data y que a diferencia de los web logs corresponde a datos que no se modifican por el usuario y son los archivos que el usuario ve durante la navegación.

4.4.1.1. Requerimientos.

Se requiere extraer desde las páginas web de un sitio web la información textual que es mostrada al usuario. El contenido extraído debe ser separado por palabras y estas a su vez deben ser derivadas o aproximadas a su raíz. Algunos aspectos importantes a ser considerados en el proceso de extracción son:

- Se deben eliminar las etiquetas html del contenido para dejar la palabra limpia.
- No se deben considerar palabras de detención como preposiciones, pronombres, etc.
- Se deben extraer las funciones en las etiquetas de script y los formatos y estilos de las etiquetas style ya que al eliminar las etiquetas el contenido interno queda como contenido textual.
- Se deben identificar en la página el título, links y palabras destacadas ya que estas corresponden a palabras especiales y que se almacenan en un arreglo adicional.
- Los archivos a analizar deben ser aquellos que contienen contenidos textuales como html, htm, jsp, asp y php. Otro tipo de páginas contiene contenidos que no son parte del análisis de textos que ve el usuario.

Esta conjugación de requerimientos debe acompañarse de la carga de las palabras a una base de datos *MySQL* que contenga la información de las palabras y su stem o raíz.

4.4.1.2. Caso de uso principal. Extracción de palabras.

Caso de Uso: Extraer palabras.

Participantes: Usuario

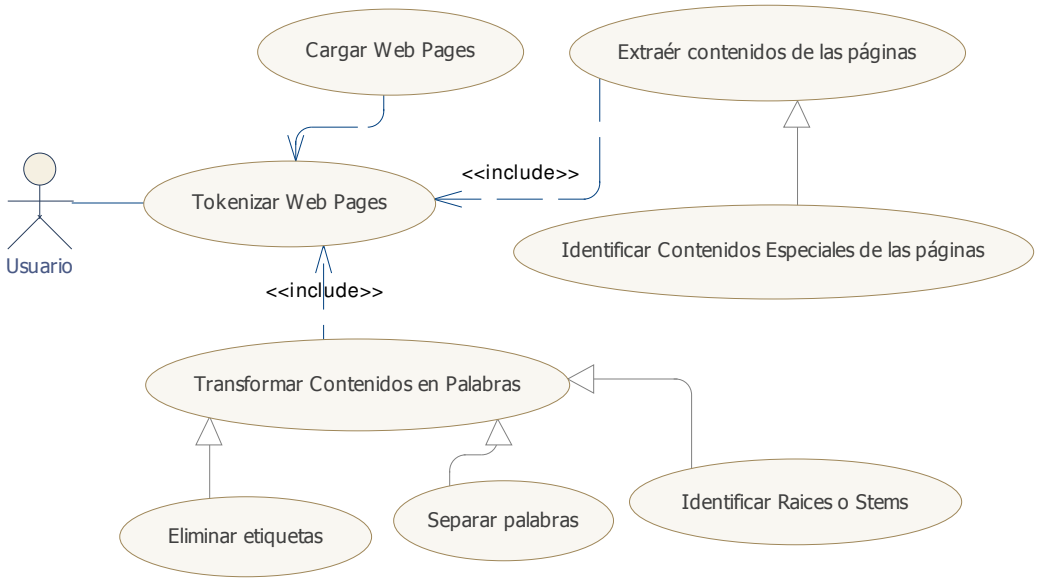
Propósito: Extraer las palabras de las páginas web que constituyen un sitio web.

Tipo: Primario u Esencial.

Descripción: La extracción de palabras consiste en la lectura del set de archivos html, htm, asp, php o jsp que conforman el sitio web. Estas palabras deben ser separadas y almacenadas en una tabla. Adicionalmente se debe identificar la raíz o stem de la palabra extraída.

4.4.1.2.1. Diagrama de Casos de Uso. Extraer Palabras.

Figura 12: ***Diagrama de casos de uso del tokenizador***

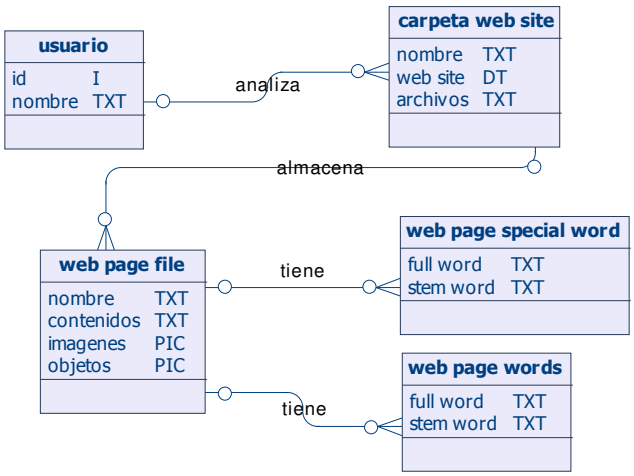


4.4.1.3. Modelo Conceptual.

La tokenización se compone esencialmente de los mismos actores que el modelo de sesionización siendo la única diferencia entre estos modelos el tipo de *web data* que se procesa. Se distingue principalmente la participación del usuario como entidad principal, el directorio de *web pages* y archivos web adjuntos, y la tabla de palabras donde se almacenarán las coincidencias encontradas.

4.4.1.3.1. Diagrama de Modelo Conceptual.

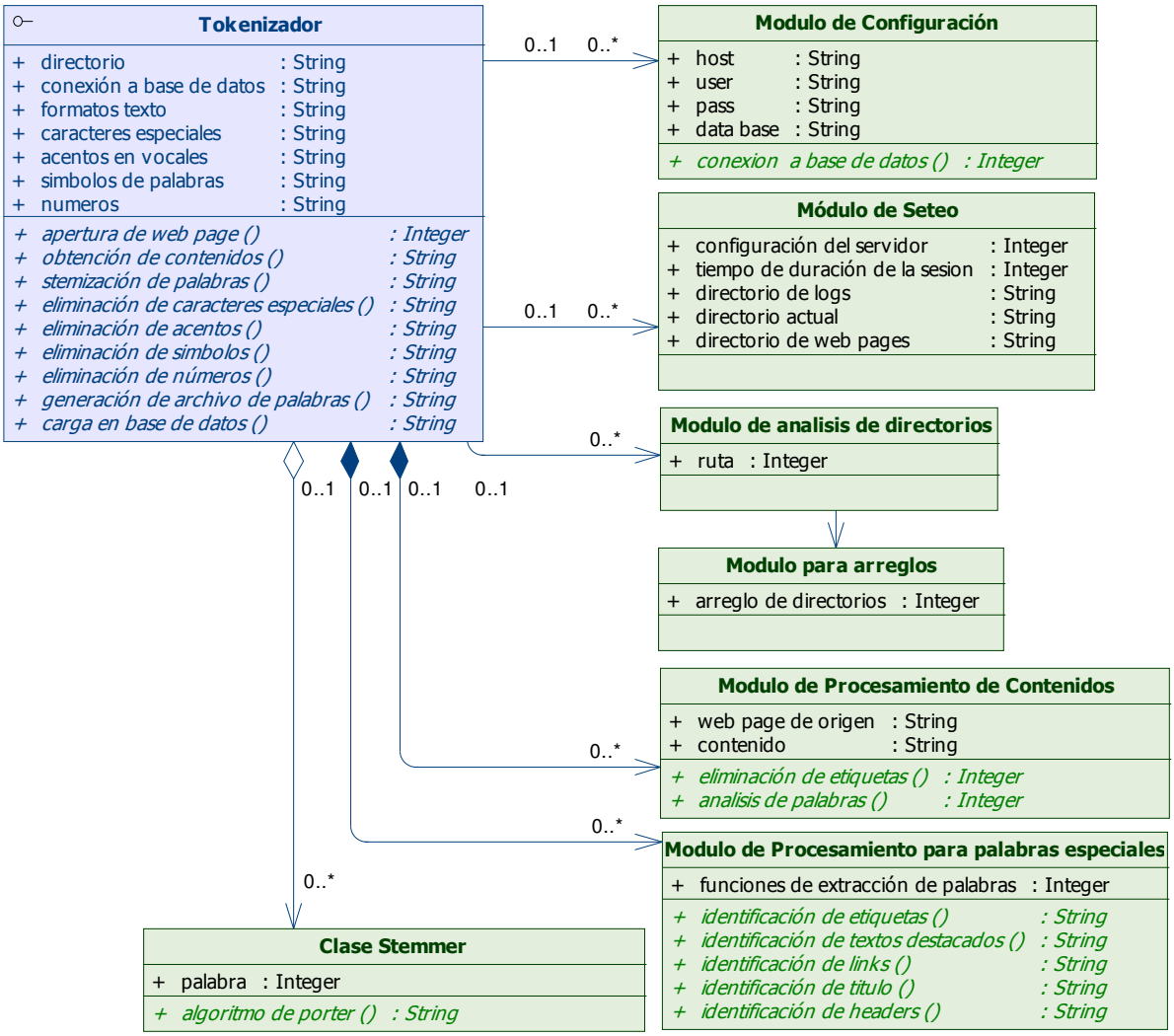
Figura 13: ***Modelo Conceptual de Tokenizador.***



4.4.1.4. Diseño de clases.

Las clases que componen el tokenizador, tienen como principal función limpiar los contenidos y separar las palabras que componen el contenido de la página web. Es relevante en este módulo desarrollado el módulo de procesamiento de contenidos que genera la limpieza del contenido de sus etiquetas y caracteres HTML; el módulo de procesamiento de palabras especiales, que previo a la limpieza identifica, separa y stemiza las palabras especiales del contenido y finalmente el módulo de stemización, que mediante el algoritmo de Porter obtiene la raíz o stem de la palabra.

Figura 14: ***Diagrama de clases de tokenizador.***

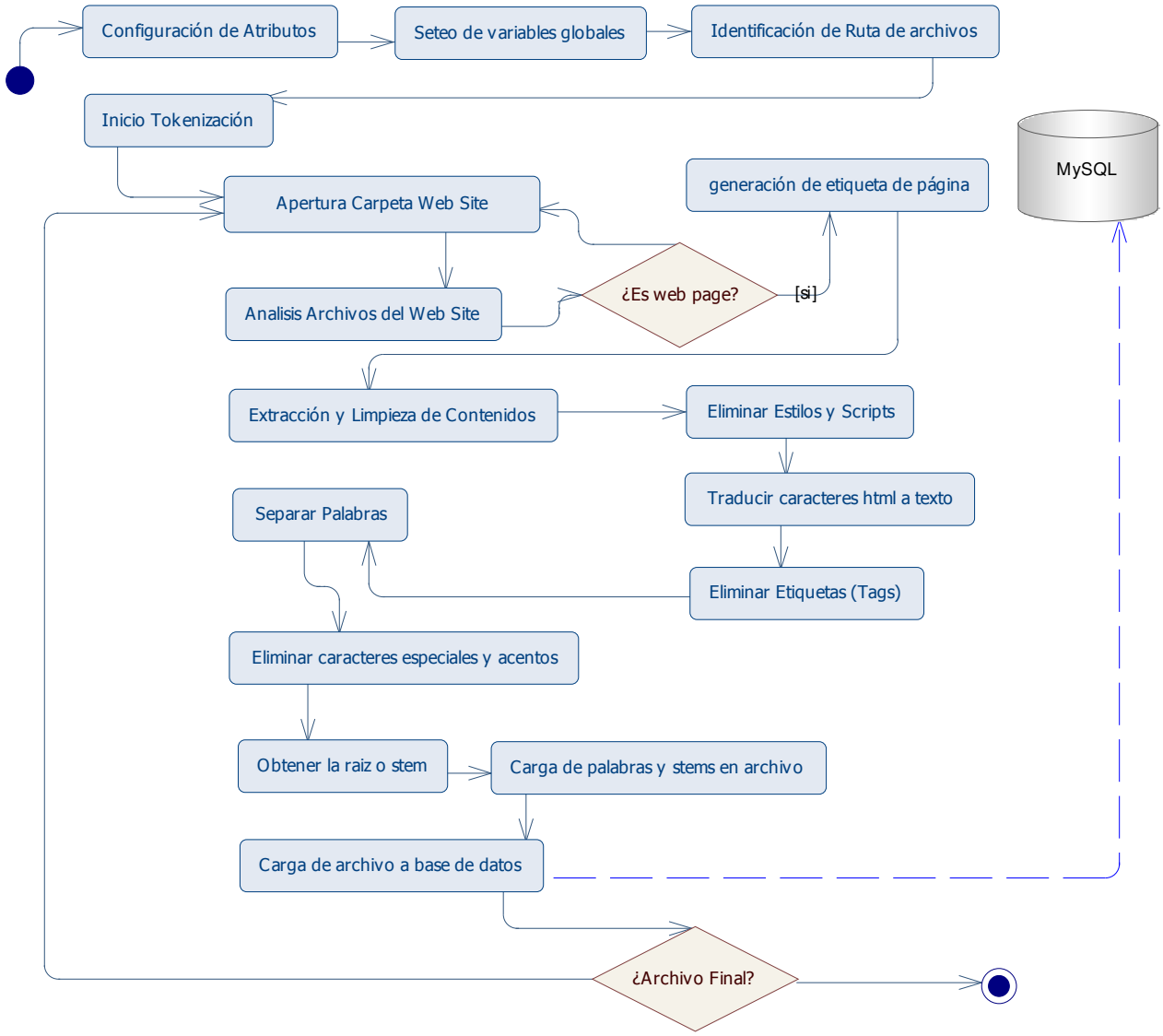


4.4.1.5. Diseño del proceso de tokenización.

El proceso de tokenización sigue una línea similar a la sesionización. La tokenización toma

un documento web y genera el proceso de extracción de contenidos y limpieza. Es relevante en este proceso la limpieza y depuración de las palabras y el archivado de ellas en la base de datos.

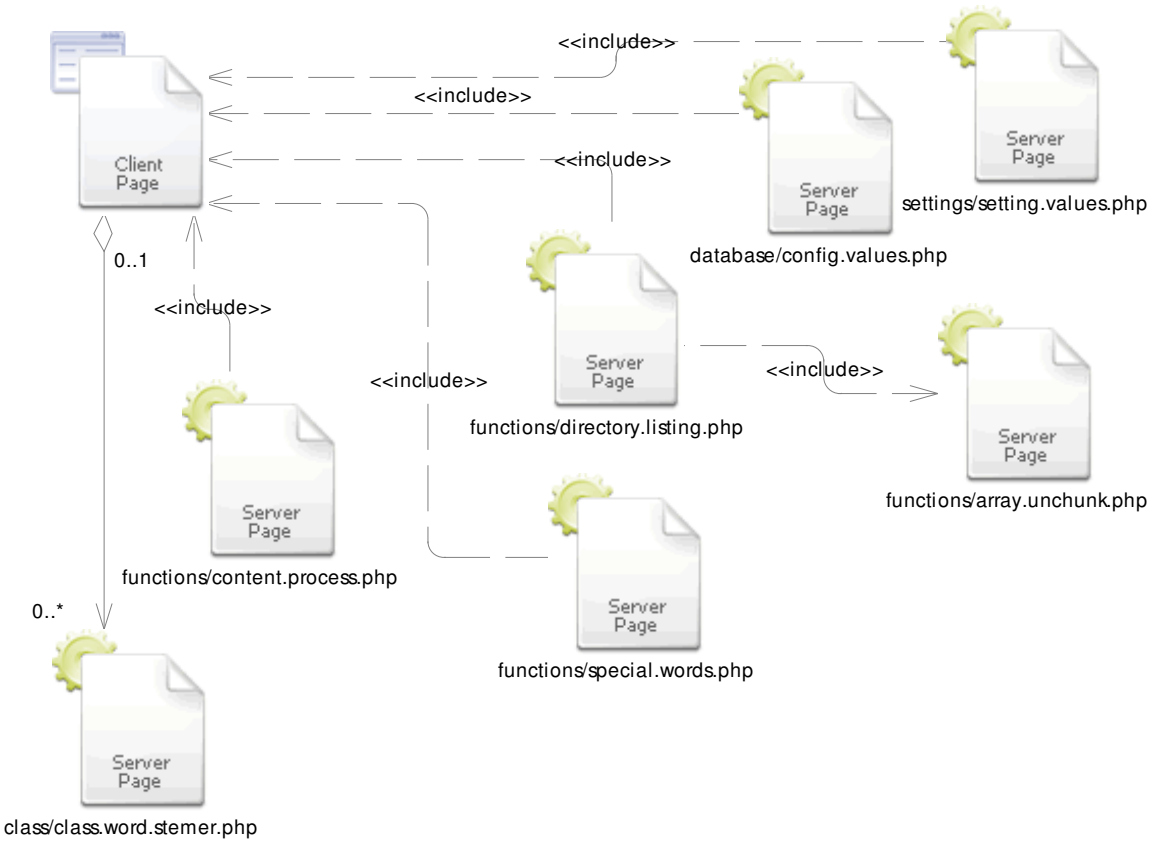
Figura 15: ***Flujo de proceso de Tokenización.***



4.4.1.6. Diagrama de Conallen.

El diagrama de Conallen muestra la interacción de la página con el módulo principal de tokenización con las otras páginas con funciones y procedimientos de apoyo. Se destacan los módulos referentes al procesamiento de las palabras principalmente.

Figura 16: ***Diagrama de Conallen del Proceso de Tokenización***



4.5. Implementación para etapa de transformación

En la etapa de transformación se hace uso de los *web data* preprocesado y seleccionado de la etapa anterior y los transforma en vectores que son utilizados para el paso posterior de *web mining*. Dentro del estudio realizado se utilizan principalmente dos vectores desde el *web data*: El vector de comportamiento del usuario (*User Behavior Vector*) que contiene información de las páginas visitadas y el tiempo invertido en ellas y el vector página web (*Web Page Vector*) que es la transformación de la pagina web en un vector según las palabras y su frecuencia en el website. Se detallan a continuación estos dos módulos de generación de vectores:

Generación del *User Behavior Vector*: El UBV es un vector originado desde el *web log* y contiene la información del comportamiento de navegación del usuario en el sitio web. Los vectores están compuestos principalmente de dos componentes por tupla: La página que visitó el usuario y el tiempo que estuvo en ella. El diagrama muestra un ejemplo de comportamiento del usuario sobre un sitio web de 5 páginas. La secuencia de navegación es la que se muestra con la flecha punteada y enmarcada con números del 1 al 4. Si se asume que esta es la i-ésima sesión analizada en un *web log,* el vector queda:

$$UBV = [(1,30),(2,10),(1,10),(3,50),(4,40)]$$

Que como se observa, considera el orden en que se navega por las páginas. Como en análisis y estudio realizado busca analizar la importancia de los contenidos y palabras claves que hay en un *website*, se hace un paso adicional en la obtención del UBV y se calcula con respecto al total de la navegación sin importar el orden las n páginas más importantes y el tiempo invertido

en ellas. Esto se conoce como *Important Page Vector* y siguiendo el ejemplo anterior tomando las 3 principales páginas queda de la siguiente forma:

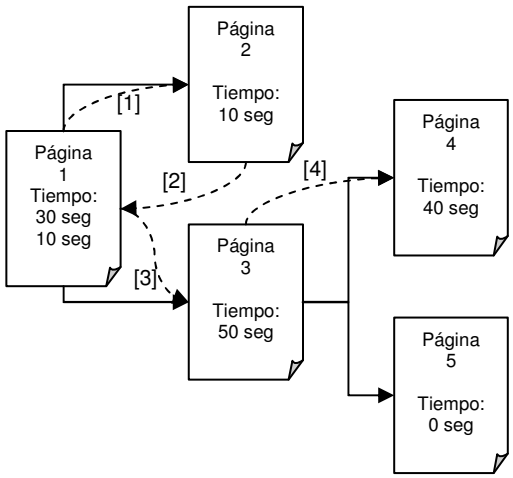
Tiempo total de navegación: 140 segundos.

Páginas navegadas: 1,2,3,4

$$IPV = [(3,36\%),(4,29\%),(1,29\%),]$$

Donde se puede distinguir la primera componente de la tupla como la página (etiquetada) y la segunda el porcentaje invertido de navegación en el sitio web.

Figura 17: ***Distinción de páginas y tiempos de un usuario.***



4.5.1. Modelamiento del *Important Page Vector*.

El vector de páginas importantes se genera desde el archivo de sesionización ya que este contiene las páginas visitadas, el tiempo invertido en las páginas y el identificador de la sesión resultante del proceso de reconstrucción de sesiones. La construcción del módulo toma como inicio el término el proceso de sesionización.

4.5.1.1. Requerimientos.

Construir el vector de páginas importantes del usuario identificado en el proceso de sesionización que contenga:

- Las tres páginas con mayor tiempo de navegación por sesión.
- Etiquetas de páginas para permitir una mejor administración de los vectores.
- Tiempos deben ser representados como porcentajes de tiempo del total de la navegación en la sesión. De esta forma se determina el interés por la página más que el tiempo gastado en ella.

4.5.1.2. Caso de uso importante: Construcción del IPV

Caso de Uso: Construir Vector de Páginas Importantes (IPV)

Participantes: Usuario

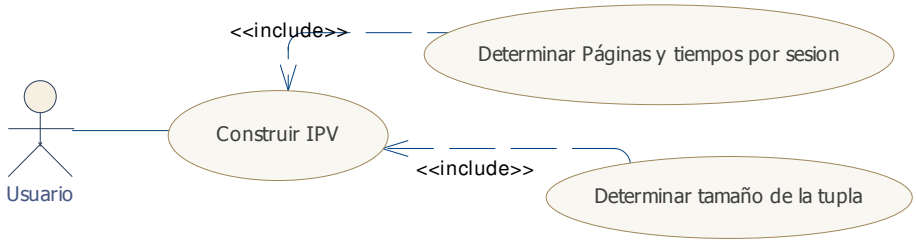
Propósito: Construir la representación vectorial de las páginas más importantes para el usuario en su navegación según el tiempo que invirtió en ellas.

Tipo: Primaria

Descripción: El vector de páginas importantes para el usuario contiene aquellas páginas en que más tiempo invirtió. Luego, el módulo a construir debe tener como input el largo de la tupla, la etiqueta de la página obtenida de la tokenización y las sesiones reconstruidas e identificadas en el proceso de sesionización.

4.5.1.2.1. Construcción de Vector de Páginas Importantes.

Figura 18: ***Diagrama de casos de uso para la construcción del Important Page Vector.***

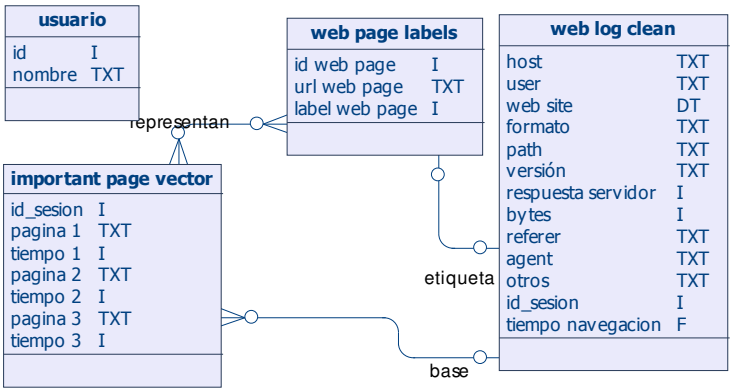


4.5.1.3. Modelo Conceptual.

El componente esencial en el modelo es la existencia previa del archivo de *web log* con la información de la sesionización resultante y los tiempos de navegación respectivos.

4.5.1.3.1. Diagrama de Modelo Conceptual.

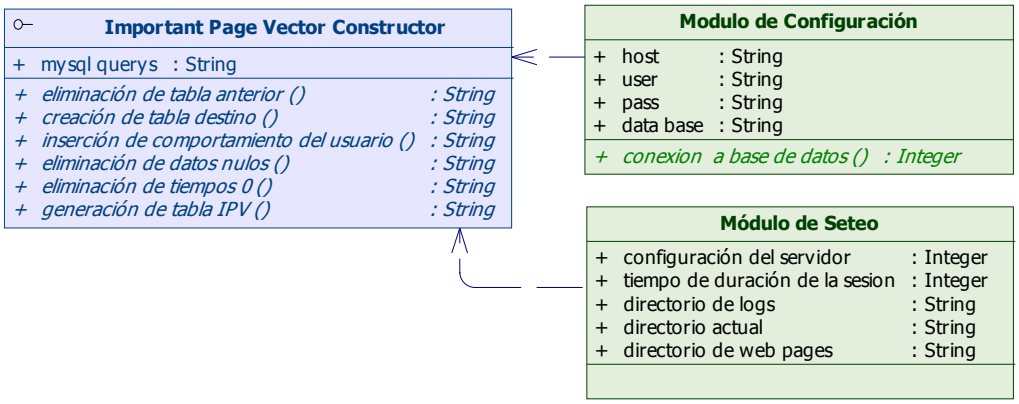
Figura 19: ***Diagrama de modelo conceptual del constructor IPV***



4.5.1.4. Diseño de clases.

A diferencia de los módulos de sesionización y tokenización, el constructor del IPV contiene la totalidad de la construcción del vector dentro de sus líneas de códigos. Los llamados externos que realiza se hacen a través del "call" de procedimientos almacenados de la base de datos que contiene las tablas de páginas importantes.

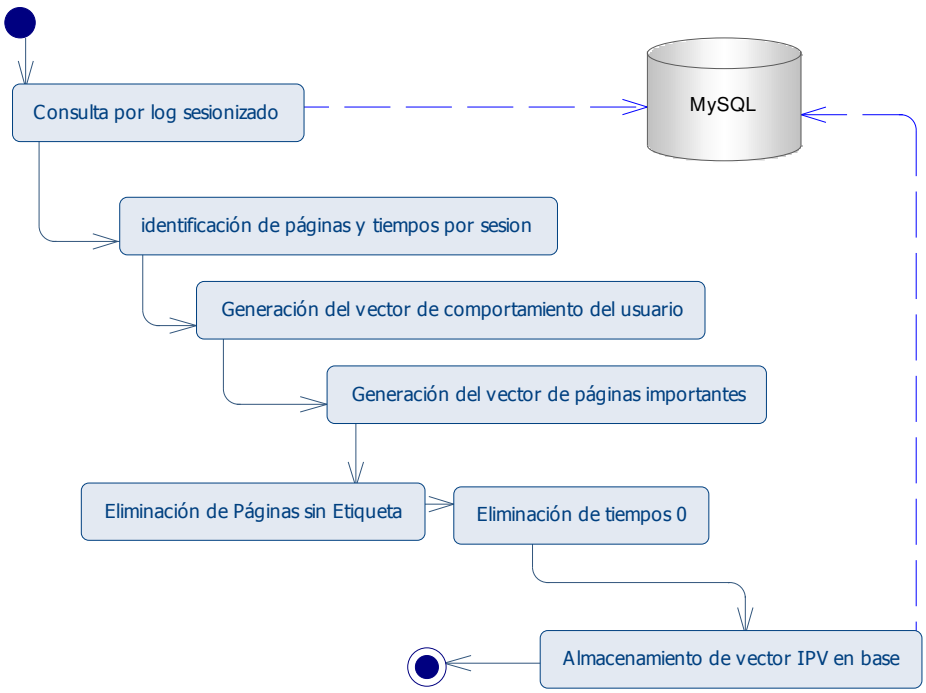
Figura 20: ***Diagrama de clases del constructor de IPV***



4.5.1.5. Diseño del proceso de creación de IPV.

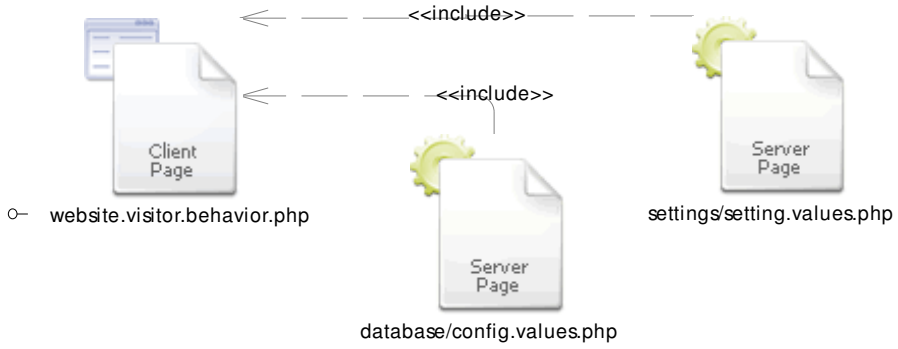
En el proceso de creación del IPV, se hace una interacción desde el inicio con la base de datos pues es en ella que se encuentra la información del resultado de la sesionización.

Figura 21: ***Flujo del proceso de construcción de IPV***



4.5.1.6. Diagrama de Conallen

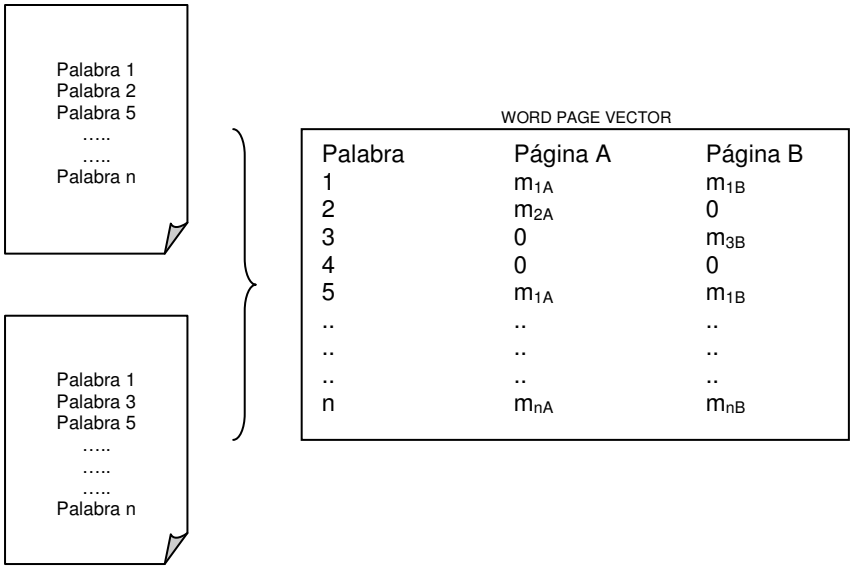
Figura 22: ***Diagrama de Conallen del proceso de construcción del IPV.***



4.5.2. Modelamiento del *Word Page Vector*

Generación del WPV: el *Word Page Vector* es la representación en el espacio vectorial de las páginas web. Este vector es construido según las palabras que contienen las páginas y peso que tienen en el sitio web según la ecuación 2Word del capítulo 2. Si se asume que para la i-ésima palabra el peso para la página A será m_{iA} y para la página B m_{iB} el vector queda:

Figura 23: ***Ejemplo de Word Page Vector (WPV)***



Para la identificación de las preferencias de los usuarios se requiere representar las páginas web en base a sus contenidos textuales (palabras). Para ello se propone generar un vector que contenga para cada página todas sus palabras y el peso que esta palabra tiene en la página y en el sitio web (ecuación (2)). Este modelamiento se debe hacer automatizado y debe tomar como

referencia principal el resultado obtenido del procesamiento de tokenización realizado.

4.5.2.1. Requerimientos.

Se requiere generar una vista vectorial de las páginas web mediante el modelo de *Word Page Vector* (WPV) que es un vector que contiene las palabras identificadas en la página del sitio web y a las cuales se aplica el peso de cada palabra calculado según la ecuación (2).

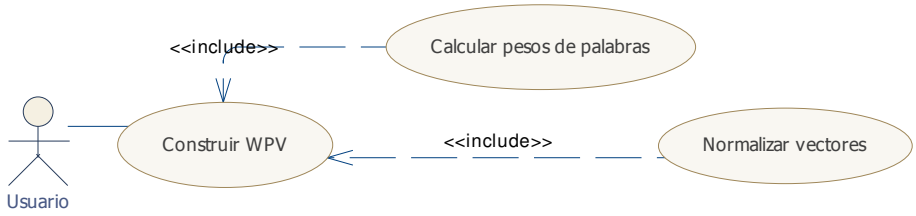
El WPV debe construirse de forma dinámica y como paso siguiente al proceso de tokenización ya que en este proceso se genera todo el material de input necesario para generar esta matriz.

4.5.2.2. Caso de Uso Principal. Construir la WPV.

Caso de Uso: Construir el Word Page Vector (WPV)
Participantes: Usuario
Propósito: Construir la representación vectorial del sitio web a través de las Word Page Vectors.
Tipo: Primaria**.**
Descripción: las páginas web pueden ser representadas por su contenido según el peso de las palabras contenidas en la página.

4.5.2.2.1. Construcción de Vector WPV.

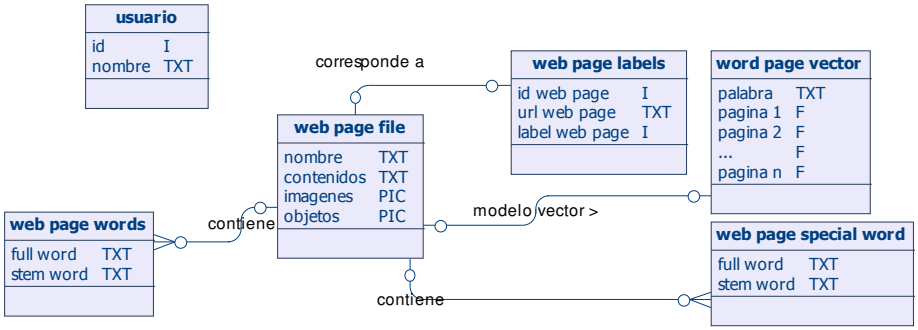
Figura 24: ***Diagrama de casos de uso del constructor WPV.***



4.5.2.3. Modelo Conceptual.

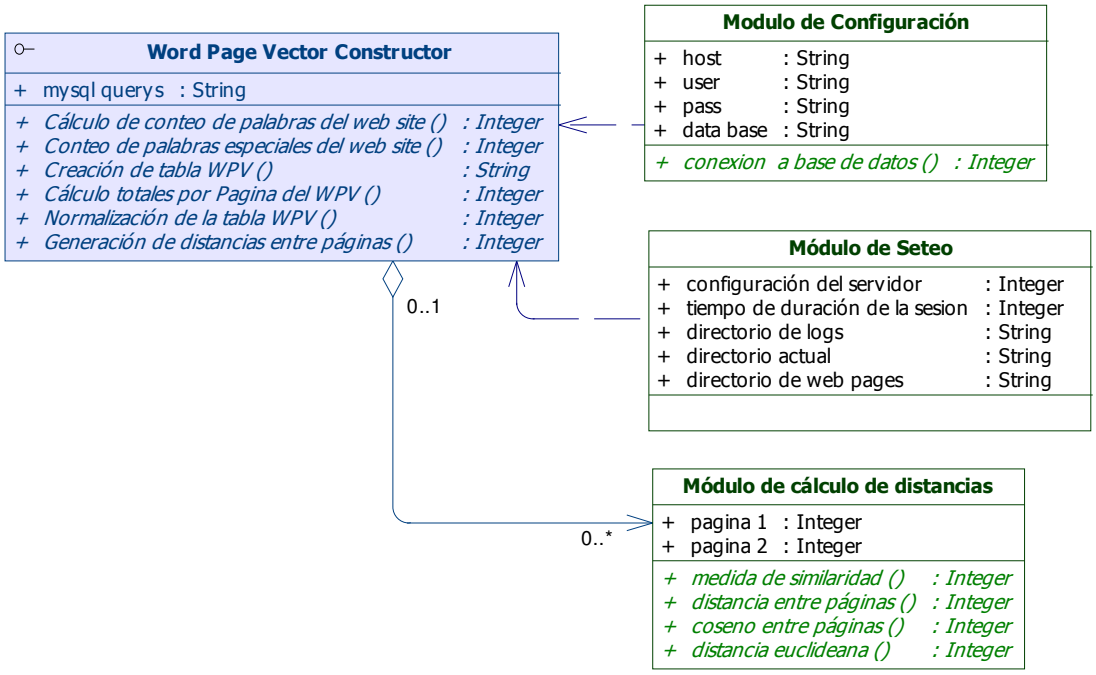
4.5.2.3.1. Diagrama de Modelo Conceptual.

Figura 25: ***Modelo conceptual del constructor WPV.***



4.5.2.4. Diseño de clases.

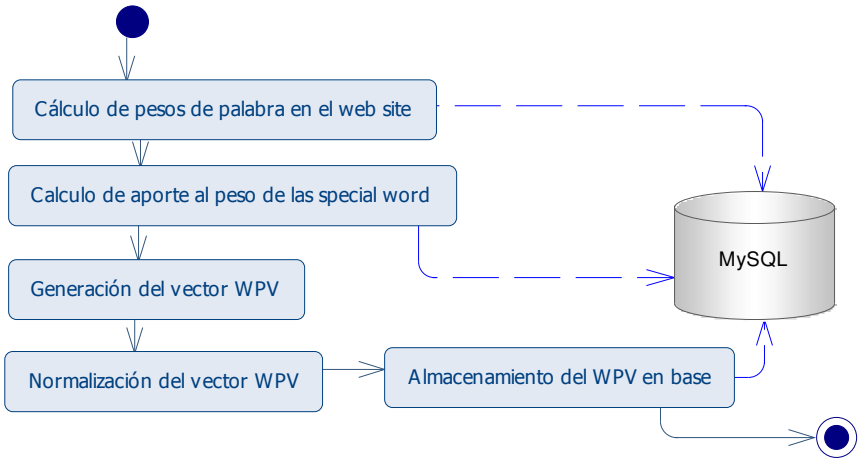
Figura 26: ***Diagrama de Clases del constructor WPV.***



4.5.2.5. Diseño del proceso de creación de WPV.

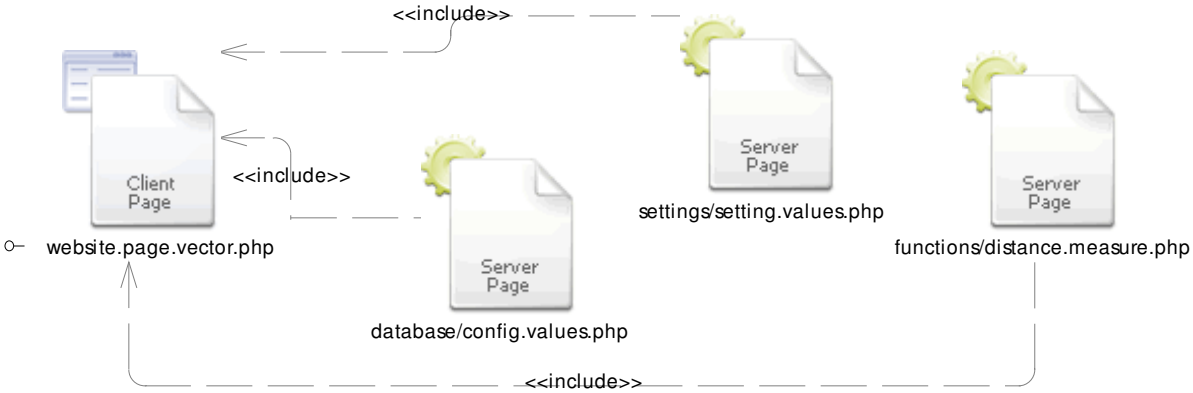
Al igual que en la creación del IPV, el proceso se inicia llamando al resultado de la tokenización desde la base de datos. El proceso que mayor cantidad de subprocesos tiene es el de Generación del WPV.

Figura 27: ***Flujo del proceso de construcción del WPV.***



4.5.2.6. Diagrama de Conallen

Figura 28: ***Diagrama de Conallen del proceso de construcción WPV.***



4.6. Implementación para etapa de web mining

4.6.1. Desarrollo de procedimientos de Comparación y Medición.

Para el procesamiento de los vectores con algoritmos y redes de aprendizaje uno de los pasos fundamentales en el proceso es determinar similitudes y/o distancias entre los vectores. Con esto es posible determinar cuan parecido son dos vectores o cual es la distancia existente entre dos páginas. Para ello se implemento el concepto *Similarity Measure* y *Vector Distance* mencionado en las ecuaciones (4) y (3) respectivamente. Estas distancias son funciones desarrolladas que toman como input los vectores en revisión y generan la medición respectiva.

4.6.1.1. Cálculo de distancia. Coseno entre Vectores.

La distancia entre vectores se puede calcular de forma mediante varias técnicas de cálculo. Para el caso de este trabajo se utilizó el concepto de coseno de un ángulo entre vectores que mediante la ecuación (3) calcula la proyección entre vectores para ver su distancia. El modulo desarrollado es una función que se llama según el requerimiento del algoritmo y entrega como resultado la distancia solicitada.

Cuadro 5: ***Código fuente del cálculo de coseno entre vectores.***

```
function coseno_between_pages($u,$v){
              $imax = count($u);
              $den_x = 0;
              $den_y = 0;
              $num = 0;
              for($i=1;$i<=$imax;$i++){
                     $num = $num + $u[$i]*$v[$i];
                     $den_x = $den_x + pow($u[$i],2);
                     $den_y = $den_y + pow($v[$i],2);
```

```
            }
            if(sqrt($den_x)*sqrt($den_y)==0){
                  $res = 0;
            }
            else{
                  $res = ($num/(sqrt($den_x)*sqrt($den_y)));
            }
            return $res;
      }
```

4.6.1.2. Cálculo de similitud. *Similarity Measure.*

El concepto de similitud es utilizado cuando se desea medir cuán parecidos son dos vectores. En el caso particular de este estudio, la medida de similitud es utilizada para comparar los UBV de los usuarios. Esta medida utiliza en su función de cálculo el coseno entre las páginas existentes en cada tupla del vector y verifica la similitud o semejanza entre los tiempos navegados mediante la tasa existente entre ellos según lo indica la ecuación (4).

Cuadro 6: ***Código Fuente del cálculo de similitud entre vectores.***

```
function distance_between_pages($u,$v){
            $ut = $u[1];
            $vt = $v[1];
            if ($ut == 0 || $vt ==0){
                  $to = 0;
            }
            else{
                  $to = min($ut/$vt,$vt/$ut);
            }
            $den_x = 0;
            $den_y = 0;
            $imax = count($u[0])-1;
            $num = 0;
            for($i=0;$i<=$imax;$i++){
                  $num= $num + $u[0][$i+1]*$v[0][$i+1];
                  $den_x = $den_x + pow($u[0][$i+1],2);
                  $den_y = $den_y + pow($v[0][$i+1],2);
            }
            if(sqrt($den_x)*sqrt($den_y) == 0){
                  $res = 0;
            }
            else{
                  $res = $to*($num/(sqrt($den_x)*sqrt($den_y)));
            }
            return $res;
      }

function similarity_measure($x,$y){
            $vr = count($x);
            $sigma = 0;
            for($j=0;$j<=$vr-1;$j++){
                  $sigma = $sigma + distance_between_pages($x[$j],$y[$j]);
            }
            return $sigma/$vr;
      }
```

4.6.2. Desarrollo de Red neuronal de Kohonen. Self Organized Feature Map.

El proceso de identificación de palabras claves requiere identificar el perfil de los usuarios que han navegado en el sitio web desde los web data existente en los *web logs* y las *web pages.* Para ello se utiliza inicialmente SOFM para identificar dichos clusters. La utilización de mediciones y comparaciones no estándar dificulta la utilización de herramientas existentes en el área de data mining. Por ello es que se desarrollo para SOFM en php para la ejecución de los algoritmos de *web mining*. El detalle del código fuente de este procedimiento desarrollado se encuentra en el anexo.

4.6.3. Desarrollo de Algoritmo de K-means.

El desarrollo consistió en la generación de un procedimiento de cálculo de distancias y similitudes entre vectores para detectar los centroides representativos de cada grupo de miembros (means). El detalle del código realizado se encuentra disponible en los anexos.

CAPITULO 4

5. APLICACIÓN DEL TRABAJO A UN SITIO WEB REAL

La aplicación de toda la teoría desarrollada durante esta memoria, debe hacerse en un sitio web complejo en estructura, volumen de visitas, actualización periódica y con gran cantidad de contenido textual.

5.1. Aplicación de análisis a un sitio web real. Un banco virtual

El estudio realizado fue realizado sobre un miembro de la industria bancaria de Chile, que correspondió a un banco virtual, es decir, no posee sucursales físicas y todas las transacciones son realizadas electrónicamente. Sus páginas están escritas en idioma español, siendo en total 212 páginas estáticas las que componen el sitio web. La cantidad de registros de logs a procesar es de aproximadamente 6,5 millones en total, es decir, contienen tanto páginas como imágenes, flash, javascripts, entre otros. Desde las web pages se obtuvieron 2.035 palabras diferentes de un total de 22.468 palabras encontradas en la totalidad del sitio web.

5.1.1. Proceso de reconstrucción de sesiones.

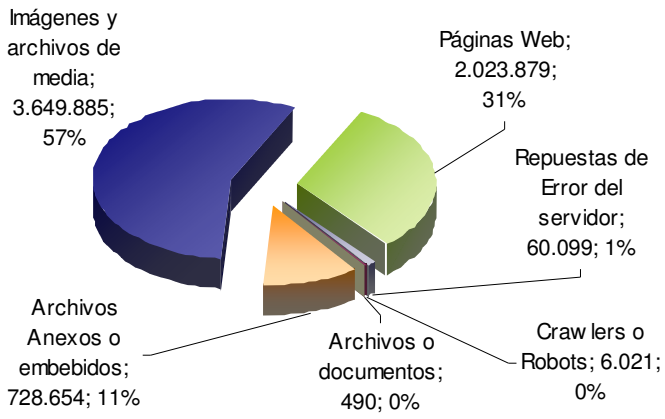
De los 6.469.028 de registros obtenidos del web log del sitio web, sólo 1.939.514 apuntan a páginas web con respuesta de servidor conforme al requerimiento del usuario. El resto de los registros obtenidos correspondió a imágenes, documentos, archivos anexos (javascripts y hojas de estilo) y registros con identificación de error del servidor (error 404, 500, etc.) el detalle de la extracción de los 209 archivos web log es:

Tabla 2: Resultados de aplicación de análisis y herramientas a un sitio web real.

Tipo	Cantidad	Ejemplo
Repuestas de Error del servidor	60.099	Error 404, 500.
Crawlers o Robots	6.021	Google, Scooter, Zibie, etc.
Archivos o documentos	490	PDF, DOC, PPT, XLS
Archivos Anexos o embebidos	728.654	SWF, CSS, JS
Imágenes y archivos de media	3.649.885	GIF, JPG, PNG, BMP
Páginas Web	2.023.879	HTML, HTM, ASP, PHP, JSP
Total	6.469.028	

De los archivos correspondientes a páginas web (2.023.879) fueron consistente en la información 1.939.514 (contenían todos los campos requeridos, no había errores en las columnas o se encontraban el la “imagen” del sitio web en estudio). Finalmente de ellas, se reconstruyeron 1.001.931 sesiones para el período de estudio entre Enero y Marzo del 2003. Finalmente, se generaron 23.027 vectores de comportamiento de usuario

Gráfico 1: ***Distribución de la aplicación del estudio a un sitio web real.***



El proceso de reconstrucción fue exitoso pero no estuvo exento de errores ni dificultades:

- Campos en web logs: un requerimiento fundamental para inicializar el proceso de sesionización, es que se cuente con la totalidad de los campos "mínimos" en el web log, indicados en el capítulo 2. Si bien durante el proceso se encontraron estos campos, aparecieron otros adicionales y que se encontraban posterior a los campos requeridos. Si bien estos campos adicionales fueron ignorados, pudieron haber aportado con mayor información o también pueden haber sido un dato no requerido para el proceso final. En el caso que el campo fuese un aporte, por ejemplo, si se tratase de un querystring, este puede contener información adicional que puede servir para identificar comportamiento, orígenes, procesos entre otros.
- Formatos de campos: los campos son de un formato estándar según el tipo de columna, sin embargo pueden estar afectas a la dependencia de set de caracteres asociados, principalmente la columna correspondiente al path de visita ya que puede estar conformado por letras o caracteres que no serán reconocidas por la configuración de la base de dato al cruzarla con datos similares. Un ejemplo de lo anterior es cuando la columna de path esta en formato UTF-8 y la etiqueta tiene el path en formato LATIN. También dentro de los mismos archivos se encontró el inconveniente de idioma de las fechas. Para el caso del analizador de las fechas, este contenía los meses en formato de 3 letras en español, pero el web log contenía las fechas en ingles. Luego, al proceso de análisis y conversión se sumó la identificación de fechas en ingles y/o español.
- Path con texto adicional (querystring): El path es la ruta de acceso que toma el navegador y que la traduce para ir a buscare la página solicitada. Este path en algunas ocasiones se acompaña de variables para determinar el tipo de contenido según diferentes criterios de filtraje y que utilizan páginas web dinámicas. En el caso del análisis de un sitio web, se requiere que el path este limpio de variables ya que en rigor las páginas no poseen contenido dinámico, luego, los paths no debiesen tener el

querystring pero en el caso de revisión de los logs si lo contenían. El problema se produce al momento de etiquetar le página, que es un proceso que facilita el manejo de datos sobre todo en vectores. Al tener un texto adicional el path, al momento de compararlo con los paths disponibles, estos daban una coincidencia nula, perdiendo gran cantidad de registros en el etiquetado. Lo que se realizó ante este inconveniente es la eliminación de todo contenido posterior al .html final del path ya que el querystring viene adjunto en el siguiente formato: http://www.ejemplo.cl/pagina.html?variable=x

El resultado final del proceso entregó un archivo de log reducido conteniendo las sesiones reconstruidas de acuerdo al criterio reactivo de reconstrucción de sesiones, es decir, considera un tiempo máximo de duración igual a 30 minutos.

5.1.2. Identificación del comportamiento del usuario.

Una sesión reconstruida, permite identificar cuáles fueron los requerimientos realizados por el cliente durante su estadía en el sitio web. Estos requerimientos son los que quedan registrados en el web log en el orden de navegación que tuvo y el momento en que se realizó el requerimiento. Lo anterior permite tener una vista de cuál es el patrón de comportamiento de un usuario específico y cuánto es el tiempo que invirtió el usuario visitando las páginas del sitio.

Si bien la secuencia de navegación puede ser determinante al momento de clasificar el comportamiento del usuario, para este estudio no fue considerado principalmente porque lo que interesa es responder cuáles son los contenidos textuales que le interesan, sin importan en que momento de la sesión fueron revisados. Además, como una de las variables que se toman en cuenta es el tiempo que se invertido por página, este permite no conocer el orden, pero si las prioridades del usuario con respecto a las páginas.

Como resultado de proceso se obtuvieron las matrices de comportamiento de usuario, similares a las indicadas en la matriz siguiente.

Tabla 3. Ejemplo de Vectores de Comportamiento del usuario (WBV)

sesión	P_1	P_2	…	P_Q
1	30	43	…	12
2	54	0	…	33
…	…	…	…	…
n	29	54	…	97

La matriz muestra en la primera columna el identificador de sesión y las columnas siguientes las páginas en que estuvo y el tiempo que se invirtió en ellas. Esta matriz se construye para la totalidad de las sesiones identificadas por lo que no es raro encontrar casos en que hay acceso a todas las páginas del website mientras en otros sólo hay acceso a una.

El siguiente paso del proceso es identificar cuales son las paginas más importantes para consolidar el llamado Vector de Paginas Importantes que identifica las i páginas en que más

tiempo se invirtió. La construcción toma el UBV construido anteriormente y en un primer paso ordena las páginas por sesión en que más tiempo se invirtió, luego, el tiempo es normalizado por cada sesión con el fin de identificar el porcentaje de tiempo en que estuvo el usuario en una página, de esta forma se omite la velocidad de lectura por usuario cuando se desea saber para que usuario es más importante. Si un usuario lee de forma pausada versus alguien que lee rápido, quizás los tiempos en cantidad serán muy diferentes, pero a nivel de interés con respecto al resto del sitio web serán similares. El vector construido arroja las k páginas más importantes para el usuario según el tiempo que invirtió. En el caso del trabajo k = 3, luego el vector quedo construido como se indica en el ejemplo:

Tabla 4: Ejemplo de Vectores de Páginas Importantes.

Id sesión	p1	t1	p2	t2	p3	t3
11	211	0,6105	28	0,3346	85	0,0549
64	28	0,5000	1	0,2500	211	0,2500
107	74	0,4596	31	0,3106	66	0,1429
175	1	0,5196	74	0,1947	114	0,1032
191	211	0,7418	1	0,2022	142	0,0561
388	211	0,6741	28	0,1975	115	0,1170
440	211	0,7473	28	0,2500	1	0,0027
445	211	0,5203	1	0,3684	28	0,1112
454	211	0,6071	15	0,2039	152	0,1429
518	211	0,5181	28	0,4799	1	0,0020
523	28	0,6291	211	0,3646	1	0,0063
534	28	0,8722	211	0,1263	1	0,0015

Con estos pasos se obtuvo el primer vector que forma parte del estudio del uso del sitio web por lo tanto alineado al *Web Usage Mining*. El resultado en volumen fue de 23027 vectores de páginas importantes de los usuarios.

5.1.3. Análisis de contenido de un website.

La segunda parte importante del procesamiento de los web data, es el análisis del contenido existente en las páginas web que componen el sitio bajo estudio. El proceso realizado sobre el sitio web, es la tokenización, mencionada en el capítulo 2. Este proceso extrae los contenidos de un sitio web y las convierte en palabras y también en palabras stemizadas, que es llegar al stem o raíz de ella mediante el algoritmo de Porter.

El proceso abarcó 212 páginas de las 213 que componían el website. Una de las páginas era de contenidos referentes a una aplicación web (software) y que no aportaba valor a la navegación, además de contener caracteres incompatibles a la lectura de la aplicación del tokenizador. Las 212 páginas contenían en total 22.468 palabras sin considerar las palabras de detención como pronombres, preposiciones, entre otras. Las únicas en total eran 2.867 y que como resultado del proceso de stemización quedo en 2.035 palabras. En promedio cada página contenía 74 palabras diferentes y un total de 106 palabras por página.

5.1.3.1. Palabras Especiales.

De las diferentes fuentes de origen de una palabra especial, se tuvo acceso a dos de ellas. Las palabras destacadas del sitio web y las palabras que se encuentran en los diferentes sitios web pertenecientes a la industria bancaria.

Palabras destacadas: del set de palabras que se encontraban destacadas según los diferentes criterios de identificación de palabras especiales en las páginas web, se encontraron 3.824 palabras, quedando 646 diferentes. Al realizar la stemización el valor se reduce solamente a 569 lo que indica que mayoritariamente las palabras especiales están en un formato similar a su raíz. De las palabras especiales se encontraron en promedio 18 de ellas por página.

Palabras de la industria: se realizó un análisis de las páginas pertenecientes a la industria bancaria, entre las que se encuentran las páginas de los bancos Itau, Estado, Bice, Bci, Corpbanca, Santander. En esta fuente de palabras especiales se detectaron 404 de ellas que eran coincidentes con las palabras especiales detectadas en el sitio web del banco en análisis. De estas palabras encontradas, la página que mayor aporte de contenido tenía fue BCI con un total de 119 páginas encontradas (ver gráfico). De la misma forma, la cantidad de palabras que aportan las entidades al arreglo total esta predominada por los bancos que más links tenían.

Gráfico 2: ***Páginas web extraídas desde los sitios web de la industria bancaria.***

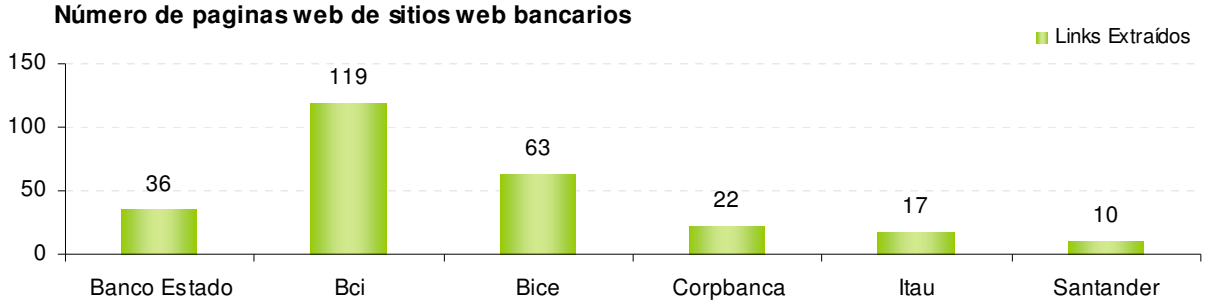
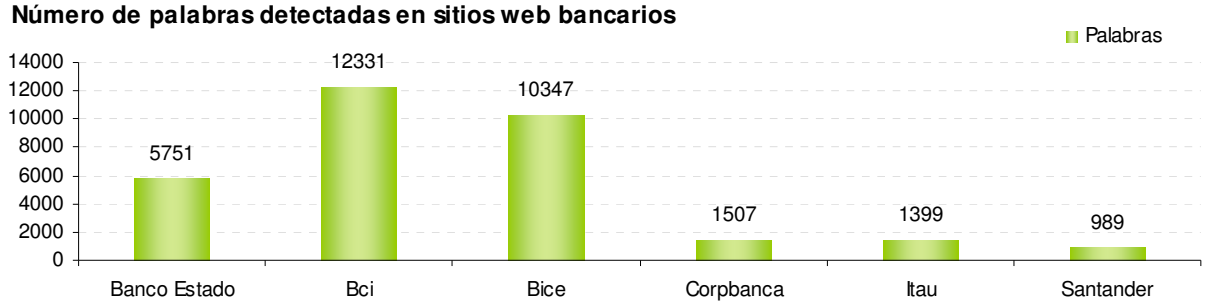
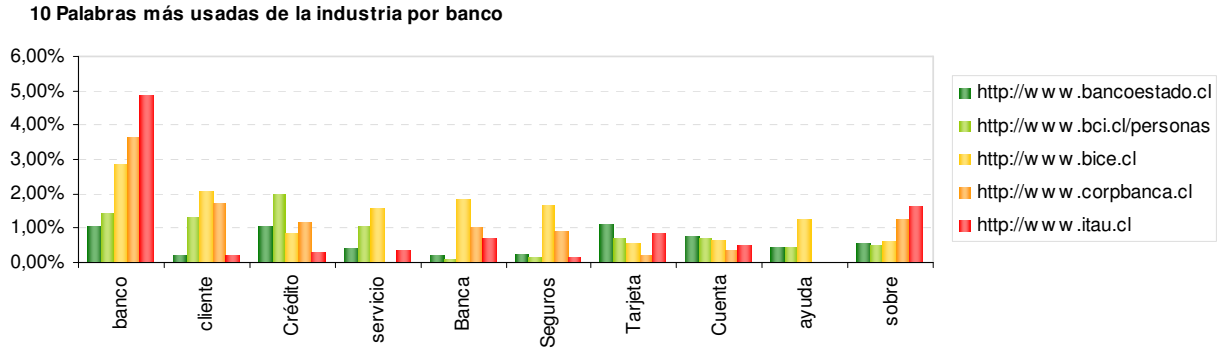


Gráfico 3: ***Palabras extraídas desde los sitios web de la industria bancaria.***



Palabras más utilizadas entre los bancos analizados: Sobre el total de palabras extraídas por banco se calculó el porcentaje de participación sobre el total para cada banco y se omitieron algunos nombres propios para no alterar resultados (ejm: Corpbanca, SBIF, ABIF). De esta forma se extrajo las palabras con mayor presencia de acuerdo al total general.

Gráfico 4: ***Palabras más utilizadas en la industria bancaria.***



Se aprecia en las palabras la significancia que tienen en el contexto de la industria en la que se encuentran inmersos y el porque estas palabras detectadas coinciden con las palabras especiales del sitio web en análisis.

Gráfico 5: ***Agrupación de páginas según cantidad de palabras en su contenido.***

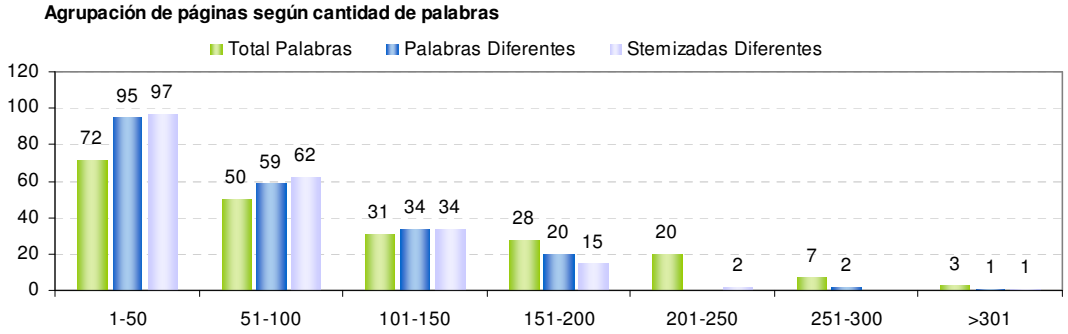
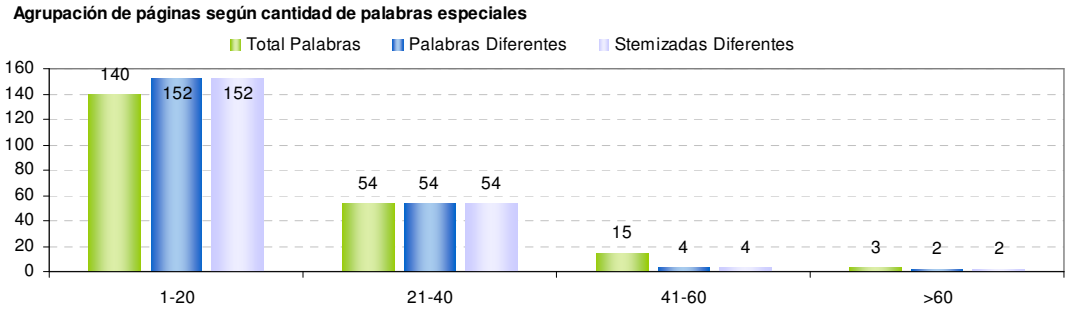


Gráfico 6: ***Agrupación de páginas según cantidad de palabras especiales.***



5.1.3.2. Análisis de palabras encontradas.

Del resultado de la extracción de la palabras es posible realizar un primer análisis de los contenidos (palabras) encontrados en el website según dos criterios. La revisión de frecuencia total de la palabra en el website y de aparición de palabra por página del website.

5.1.3.3. Revisión de palabras por Website.

El resultado del proceso de tokenización dio como resultado el siguiente set de palabras. En este caso son mostradas solamente 20 de las palabras más importantes.

Tabla 5: Ranking de palabras originales más frecuentes.

Ranking	Palabra	Total conteo palabras normales	Total conteo palabras especiales	Total de palabras
1	****	427	190	617
2	fondos	279	57	336
3	inversiones	179	105	284
4	ahorro	212	63	275
5	cuenta	228	32	260
6	linea	151	99	250
7	mutuos	188	48	236
8	banco	139	92	231
9	credito	182	47	229
10	comision	143	59	202
11	dias	168	27	195
12	pago	149	45	194
13	home	95	93	188
14	cliente	177	9	186
15	plazo	141	41	182
16	servicio	167	12	179
17	mapa	87	86	173
18	contactanos	86	85	171
19	monto	152	11	163
20	tarjeta	146	17	163

La palabra con más aparición dentro de todos los contenidos del sitio web es el nombre del banco, lo cual es consistente ya que parte de la gestión del contenido realizada es la imagen de la marca que se desea dar al cliente. Luego, una segunda tanda de palabras muestra una mayor relación con respecto a la industria bancaria; palabras como fondos, industria, ahorro son muy frecuentes y puede ser por el contexto de resguardo o cliente a futuro que desea dar el banco a modo de fidelización. Un último conjunto de palabras muestra palabras utilizadas frecuentemente en la estructura del banco como mapa, contáctanos, banco y home.

Tabla 6: Ranking de palabras stemizadas más frecuente.

Ranking	Palabra	Total conteo palabras normales	Total conteo palabras especiales	Total de palabras
1	****	427	190	617
2	fond	407	70	477
3	pag	288	59	347
4	cuent	310	35	345
5	ahorr	250	64	314
6	inversiones	179	105	284
7	mutu	219	63	282
8	credit	215	62	277
9	servici	235	41	276
10	banc	168	100	268
11	line	152	99	251
12	tarjet	194	37	231
13	plaz	160	44	204
14	comision	143	59	202
15	dias	173	27	200
16	home	95	93	188
17	cliente	177	9	186
18	cuot	147	33	180
19	map	87	86	173
20	contactan	86	85	171

El resultado del set de palabras es similar al anterior, pero se ve más que nada afectado el orden de la frecuencia puesto que pueden haber una agrupación de las palabras productos de la stemización, luego aumenta su frecuencia general. El inconveniente de este proceso es que en algunas ocasiones la palabra puede perder contexto y puede ser asociada a un grupo que no corresponde. Si por ejemplo se toma la palabra stemizada “line”, puede estar asociada a una línea de crédito o bien a una línea de teléfono.

5.1.3.4. Revisión de palabras por páginas del website.

Se agrega al análisis anterior, la búsqueda o conteo de lugares desde donde es referenciada una palabra. La siguiente revisión busca identificar en cuantos documentos están presente las palabras respectivamente:

Tabla 7: Ranking de palabras originales más frecuentes por página web.

Ranking	palabra	normal	special	Total general
1	****	149	132	281
2	inversiones	120	93	213
3	linea	106	94	200
4	banco	100	92	192
5	home	93	92	185

6	mapa	85	85	170
7	contactanos	85	85	170
8	cliente	135	7	142
9	pago	81	42	123
10	cuenta	102	18	120
11	ahorro	74	35	109
12	servicio	101	5	106
13	fondos	69	31	100
14	atencion	96	3	99
15	ejecutivos	96	2	98
16	autoconsulta	95	2	97
17	hazte	94	3	97
18	plazo	61	34	95
19	personas	59	34	93

En una primera revisión, se observa que nuevamente es el nombre del banco el mayoritariamente referenciado en la totalidad del documento. Para el caso de las otras palabras hay un reordenamiento de las palabras y aparición de nuevas de ellas con respecto a la revisión anterior. Se observa un segundo grupo asociado a la estructura del sitio web con las palabras home, mapa, contáctanos. Seguido, aparecen palabras más frecuentemente asociadas a la cuenta corriente y a las utilidades que presta el banco al cliente como pago, cuenta, ahorro, servicio, fondos. Posteriormente un tercer grupo señala la posibilidad de recibir atención de ejecutivos y servicio de auto consulta. Con lo anterior el banco da la posibilidad constante de atención o auto atención.

Tabla 8: Ranking de palabras stemizadas más frecuentes por página web.

Palabra	Normal	Especial	Total general
tbanc	149	132	281
pag	170	52	222
banc	119	99	218
inversiones	120	93	213
line	107	94	201
home	93	92	185
map	85	85	170
contactan	85	85	170
fond	120	44	164
cuent	139	20	159
servici	133	23	156
cliente	135	7	142
ahorr	100	36	136
ejecut	111	4	115
credit	77	34	111
plaz	73	36	109
mutu	68	40	108
voluntari	94	12	106
person	70	35	105

El segundo conjunto en análisis, correspondiente a las palabras stemizadas muestra un comportamiento muy similar en frecuencia de aparición en páginas respecto al análisis de palabras completas. No hay mayor cambio en ellas.

5.1.4. Modelando las páginas web al espacio de vectores.

En el capítulo 2, se indicó que para la realización del estudio de las preferencias de contenido de los usuarios, una etapa importante en el proceso es la de transformación de una página web a un modelo matemático que sea utilizable por procesos computacionales y algoritmos de web mining. Esto último se realiza mediante la transformación de una página web a un modelo de vectores. Luego del proceso de extracción de contenidos, se inicia el proceso de conformación del WPV o web page vector que es una tupla por cada página que contiene según la palabra el peso mij que le corresponde según la ecuación (2). Como resultado de la postulación anterior, se obtiene una página inicial que contiene los pesos según el ejemplo:

Tabla 9: Ejemplo de WPV. Extracto.

Palabra	1	2	3	…	212
aeropuert	0,0000	0,0000	0,0000	…	0,0000
apertur	0,0000	0,0000	0,0000	…	0,0000
asegur	0,0006	0,0006	0,0006	…	0,0006
aspect	0,0000	0,0000	0,0000	…	0,0000
colocac	0,0081	0,0081	0,0081	…	0,0081
comis	0,0031	0,0031	0,0031	…	0,0031
estructur	0,0000	0,0000	0,0000	…	0,0000
financi	0,0000	0,0000	0,0000	…	0,0000
hog	0,0010	0,0010	0,0010	…	0,0010
portafoli	0,0024	0,0024	0,0024	…	0,0024
promoc	2,5273	2,5273	0,0040	…	0,0040
real	0,0000	0,0000	0,0000	…	0,0000
reemplaz	0,0000	0,0000	0,0000	…	0,0000
superior	0,0009	0,0009	0,0009	…	0,0009

La revisión de las palabras, permite de forma adicional al análisis que se pretende hacer al momento de identificar las *website keywords*, obtener una idea de cómo esta constituido el website bajo estudio, revisando si existe consistencia entre las palabras que utiliza y la industria en la que se encuentra inmerso. Permite además conocer las prioridades actuales en el uso de palabras y así poder contrarrestar en el caso de diferencia y corroborar en el caso de coincidencia de los resultados.

5.2. Analizando las preferencias de texto de los usuarios del sitio web.

Correspondiente a la etapa de *web mining*, el análisis de preferencia de los usuarios hace uso de las páginas web transformadas a vectores (WPV) y de las páginas importantes del usuario para identificar mediante un mix el contenido que le importa y por ende las palabras más importantes en él. Estos vectores son el input necesario para que los algoritmos desarrollados, SOFM y K-means, generen el proceso de *clustering* para luego identificar aquellos grupos de usuarios y sus preferencias.

Se muestra a continuación los resultados de la aplicación por algoritmo realizado. Según lo anterior, se hará análisis de los contenidos de cada cluster junto con su vecindad para luego identificar aquellos contenidos y palabras más importantes.

5.2.1. Aplicación de SOFM.

El mapa donde se correría el proceso para identificar los clusters mediante esta red no supervisada tendría como entrada 3 neuronas y como salida 32 (mapa de 1024 neuronas) y partiría con una vecindad de 5 neuronas cuyo radio se reduce a medida que avanza el algoritmo.

El resultado del procesamiento de los datos proporcionados en los procesos anteriores dio como resultado el mapa mostrado en la figura:

El mapa muestra 10 *clusters* que contienen las páginas más importantes del sitio web según los usuarios que la visitan. Para la validación de que estos clusters se toman los temas centrales por página y si hay relación o coincidencia en ellos se validan, en caso contrario no es un cluster válido o consistente quedando nulo el centroide y sus vecindades para el análisis.

Para que el proceso quede libre de errores o malas interpretaciones se consideran aquellas palabras que tienen significado con la industria en la que esta inmersa la organización. Aquellas palabras que sean principalmente parte de la estructura del sitio web, como palabras “home”, “mapa”, “contáctenos” o nombres propios (en los que se incluye el nombre del banco), serán omitidos con el fin de obtener una mejor interpretación y restarle importancia a palabras que se repiten pero que no agregar semántica al grupo de palabras obtenidos

Desde la información anteriormente obtenida se obtuvo además los vectores que componían dichos clusters junto con su vecindad.

5.2.1.1. Análisis de Clusters Resultantes del proceso

Se realizaron dos análisis para la aplicación de SOFM. El primero de ellos correspondió a un mapa de 16 x 16 neuronas y el segundo análisis realizado se hizo con un mapa de 32 x 32 neuronas. Para cada revisión de clusters se hace principal análisis de su centroide y luego de su vecindad. Para efectos de mostrar los resultados obtenidos se mostrará en detalle el análisis de cada cluster y se mencionará el resultado para cada vecindad del respectivo cluster. El detalle de esta información será incluida en los anexos.

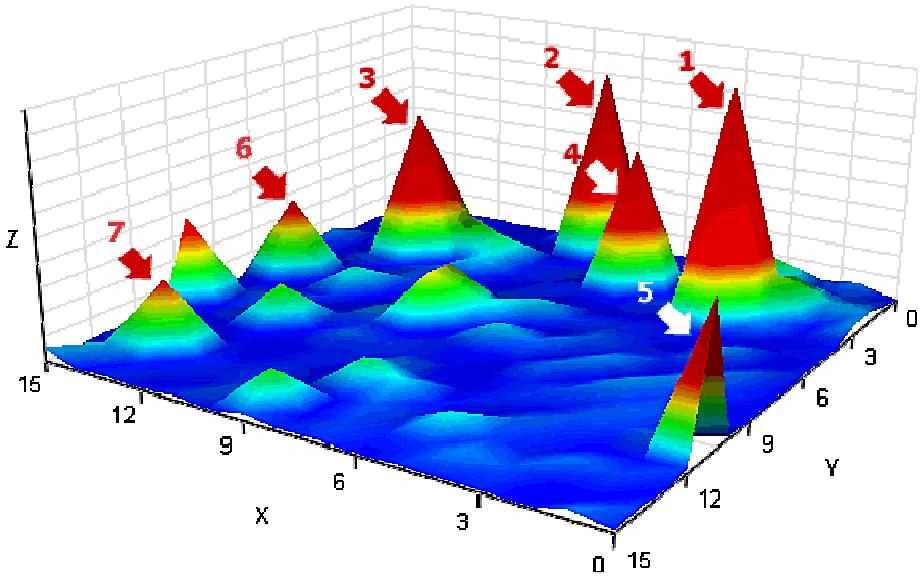
5.2.1.2. Análisis de Clusters para mapa SOFM de 16 x 16.

El resultado del proceso de análisis mediante SOFM con un mapa de 256 neuronas de

salida entregó como resultado el mapa indicado en la figura 29. En esta se observan 7 clusters de los cuales 5 de ellos son notorios de mayor forma respecto a los otros dos. Dada la topología desarrollada en el proceso, el cluster que se encuentra en las neuronas 9 a 12 del eje "***y***" continúa en el eje siguiente y no corresponde, por lo tanto, a un cluster adicional.

El criterio de aceptación o rechazo de un cluster es: si las páginas dentro de cada cluster están relacionadas con un tema principal similar, entonces el cluster es aceptado, en caso contrario, se rechaza. Aplicando este criterio, 5 clusters son aceptados y el patrón contenido en cada una de ellas fue utilizado para extraer las palabras claves del sitio web.

Figura 29: ***Mapa de Clusters resultante del proceso de SOFM con 16 Neuronas.***



De los clusters detectados, se extrajeron las palabras respectivas mediante la utilización de la ecuación (10). Luego, el análisis e interpretación de los clusters identificados son:

5.2.1.2.1. Análisis de Cluster 1.

9,57% de hits en la neurona respecto al total de análisis del mapa.

En el Cluster 1 (etiquetado de la misma forma en la figura anterior) se encontraron los siguientes vectores:

Tabla 10: Resultado SOFM de 16x16. Cluster 1.

Tipo	Hits	Vector
Centroide	85386	(117, 192, 19)
Vecino	3924	(21, 10, 179)
Vecino	1113	(205, 128, 210)
Vecino	2270	(55, 18, 41)
Vecino	832	(24, 104, 95)

Los contenidos detectados en el centroide de acuerdo a las páginas que lo componen corresponden son referentes a consulta de orientación de utilización puntos entregados por compras con la tarjeta de crédito. La segunda página corresponde a la página de despliegue del estado de la tarjeta de crédito en compras internacionales y finalmente la tercera página contiene la información del servicio Premium. Estos tres temas son concordantes entre si por lo que no se descarta el centroide.

El análisis de las palabras contenidas en el cluster según la aplicación de la ecuación 10 al resultado es:

Tabla 11: Resultado SOFM de 16x16. Palabras encontradas.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
actual	0,380958024	4,32%	11	2	actual, actualidad
canje	0,374119236	4,25%	54	6	canje, canjeados, canjeando, canjear, canjearlos, canjearse
pag	0,311101	3,53%	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
total	0,281381752	3,19%	52	2	total, totalidad
compr	0,262167618	2,98%	98	5	compra, comprando, comprar, compras, compro
calcul	0,239716065	2,72%	22	8	calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, cálculos
cost	0,23844382	2,71%	19	3	costado, costo, costos
sald	0,230532417	2,62%	80	2	saldo, saldos
anterior	0,227575573	2,58%	21	1	anterior
pes	0,211510761	2,40%	136	2	peso, pesos

Las palabras detectadas coinciden en la revisión del estado de cuenta de la tarjeta de crédito y consulta por pesos disponibles. Las palabras canje, pago, pesos, compra, saldo y costo tienen relación en el uso y estado de este instrumento financiero. Por lo tanto, se puede inferir que este cluster corresponde a Heavy Users de tarjetas de crédito dado que generan consultas, canjes y compras con los puntos que adquieren en sus compras con este documento.

La vecindad de este cluster hace referencia a consultas de créditos y deudas en general, lo cual se puede asemejar a lo anterior dado que la referencia principal es el uso de una tarjeta de crédito.

5.2.1.2.2. Análisis de Cluster 2.

7,85% de hits en la neurona respecto al total de análisis del mapa.

El cluster 2 dio como resultado los siguientes vectores según centroides y vecindades.

Tabla 12: Resultado SOFM de 16x16. Cluster 2.

Tipo	Hits	Valor
Centroide	72297	(130, 49, 10)
Vecino	8853	(125, 24, 145)
Vecino	794	(104, 15, 43)
Vecino	2838	(10, 104, 104)
Vecino	339	(91, 208, 172)

El contenido de este centroide menciona una simulación de retiro por parte del usuario, tributación de aportes y consulta a minicartola. Las vecindades a su vez contienen información de APV (Ahorro Provisional Voluntario), Fondos Mutuos y estado de Créditos, por lo que se puede deducir de esta información inicial que el centroide puede estar relacionado con la participación de un usuario cercano a la jubilación o bien provisorio.

Tabla 13: Resultado SOFM de 16x16. Palabras encontradas en Cluster 2

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
pag	0,809567137	0,127676	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
impuest	0,350775187	0,05532	93	2	impuesto, impuestos
total	0,322054777	0,050791	52	2	total, totalidad
calcul	0,301996889	0,047628	22	8	calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, calculos
rent	0,276141464	0,04355	65	2	renta, rentas
ingres	0,20415	0,032196	47	8	ingresa, ingresada, ingresan, ingresando, ingresar, ingresas, ingreso, ingresos
mism	0,134217786	0,021167	27	4	misma, mismas, mismo, mismos
hast	0,083356752	0,013146	124	1	hasta
fond	0,077559978	0,012232	477	2	fondo, fondos
retenciones	0,063983096	0,010091	32	1	retenciones

Las palabras contenidas en las páginas del centroide, hacen una referencia principal a pagos, impuestos, rentas, fondos y retenciones. Se puede, por lo tanto, hacer la deducción a que el interés de este grupo de usuarios esta centrado en el pago de APV (según lo indicado por el contenido de las páginas). Este grupo de usuarios es cercano a la tercera edad o interesado en mantener y hacer sus fondos de jubilación. Las principales características de estas palabras son la relación hacia lo referente a renta, ingresos y fondos que son los que se pueden vincular a la asociación que da el usuario al porcentaje a derivar para su pensión.

5.2.1.2.3. Análisis de Cluster 3.

6,21% de hits en la neurona respecto al total de análisis del mapa.

El 3° cluster identificado esta compuesto por el siguiente centroide y vecindades:

Tabla 14: Resultado SOFM de 16x16. Cluster 3

Tipo	Hits	Valor
Centroide	57224	(126, 59, 76)
Vecino	340	(132, 30, 14)
Vecino	1458	(158, 203, 143)
Vecino	947	(29, 27, 19)
Vecino	969	(127, 210, 11)

Este centroide de acuerdo a las páginas que hace referencia se relaciona a conceptos de pago de tarjeta, consulta de saldos, consulta de cheques pagados. Luego, el interés por este cluster detectado se encuentra en la resolución de pago de las deudas y el estado y control de su cuenta corriente. Las palabras encontradas en el cluster corresponden a las siguientes:

Tabla 15: Resultado SOFM de 16x16. Palabras encontradas en Cluster 3.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
nombre	2,05573	0,189969	91	1	nombre
pag	1,918767353	0,177312	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
cuent	1,035216491	0,095664	345	3	cuenta, cuentan, cuentas
corriente	0,775453	0,071659	117	1	corriente
total	0,464213897	0,042898	52	2	total, totalidad
cheques	0,352820975	0,032604	87	1	cheques
estad	0,267148004	0,024687	76	2	estado, estados
informacion	0,234333433	0,021655	60	1	informacion
contable	0,228003492	0,02107	7	1	contable
fech	0,198306553	0,018325	101	2	fecha, fechas

Las palabras encontradas producto de la ecuación (10) hacer referencia a lo que es el pago, cuenta, cuenta corriente, estado de cheques e información. Luego, el centroide indicado se puede relacional a un grupo que hace uso del sitio web bancario para el desarrollo de sus actividades y gestiones en sus cuentas. La relación de las palabras deriva en lo que es la utilización del servicio de Internet para resolución e información de su cuenta. Se puede resumir entonces el cluster como grupo de personas que utilizan el sitio web para transacciones bancarias con su cuenta corriente. Puede tratarse de empresarios ya que la referencia de estado de cheques puede estar relacionada a pagos a fecha, como lo efectúan las empresas a sus proveedores.

5.2.1.2.4. Análisis de Cluster 4.

5,74% de hits en la neurona respecto al total de análisis del mapa.

El cuarto cluster se compone de los siguientes vectores de páginas importantes:

Tabla 16: Resultado SOFM de 16x16. Cluster 4

Tipo	Hits	Valor
Centroide	52844	(58, 64, 10)
Vecino	961	(125, 70, 143)
Vecino	1461	(8, 145, 99)
Vecino	1075	(26, 84, 13)
Vecino	388	(171, 24, 145)

El concepto contenido en el centroide se refiere a pago de deuda internacional, solicitud de crédito de consumo y consulta por tributación. Luego se puede definir una primera instancia de interpretación del cluster como grupo de usuarios que tienen una deuda en el banco y que necesitan acceder a un crédito para efectuar el pago. El análisis del contenido de estas páginas es el que sigue:

Tabla 17: Resultado SOFM de 16x16. Palabras encontradas en Cluster 4

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
mont	0,508359779	0,071973	165	2	monto, montos
deud	0,295860212	0,041888	96	2	deuda, deudas
inversion	0,222793203	0,031543	79	2	inversion, inversionistas
estad	0,212136952	0,030034	76	2	estado, estados
fech	0,198306553	0,028076	101	2	fecha, fechas
compr	0,193247521	0,02736	98	5	compra, comprando, comprar, compras, compro
solicitud	0,17448241	0,024703	35	1	solicitud
cuent	0,100602285	0,014243	345	3	cuenta, cuentan, cuentas
fond	0,090673327	0,012837	477	2	fondo, fondos
futur	0,083665677	0,011845	95	3	futura, futuro, futuros

El resultado de las palabras obtenidas muestra a un cliente interesado en inversión pero al mismo tiempo en solicitar crédito por aparentes deudas de las mismas. Puede tratarse de usuarios que invierten principalmente, pero que necesitan de apoyo financiero por no contar posiblemente con la liquidez necesaria.

5.2.1.2.5. Análisis de Cluster 5.

Tabla 18: Resultado SOFM de 16x16. Cluster 5.

Tipo	Hits	Valor
Centroide	50294	(192, 182, 198)
Vecino	471	(11, 104, 153)
Vecino	3556	(133, 29, 127)
Vecino	9551	(111, 111, 25)
Vecino	9873	(152, 95, 117)

Correspondiente al 5,46% de hits en el mapa, este cluster tiene páginas asociadas a inversión, acciones y crédito hipotecario. Luego, en una primera aproximación se encuentra en este centroide un cluster de usuarios interesado en su futuro al igual que el anterior.

Tabla 19: Resultado SOFM de 16x16. Palabras encontradas en Cluster 5

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
nombre	2,05573	0,204896	91	1	nombre
Pag	1,918767353	0,191245	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
Total	0,464213897	0,046269	52	2	total, totalidad
Deud	0,295860212	0,029489	96	2	deuda, deudas
informacion	0,234333433	0,023356	60	1	informacion
Estad	0,212136952	0,021144	76	2	estado, estados
Fech	0,198306553	0,019765	101	2	fecha, fechas
Compr	0,193247521	0,019261	98	5	compra, comprando, comprar, compras, compro
Cuent	0,100602285	0,010027	345	3	cuenta, cuentan, cuentas
Fond	0,090673327	0,009037	477	2	fondo, fondos

Las palabras obtenidas del procesamiento del vector se asemejan o asocian a la administración de las cuentas del usuario y en particular para el caso de cuentas que se encuentran con deudas. Puede tratarse de usuarios interesados de su estado financiero a futuro.

5.2.1.2.6. Análisis de Cluster 6.

Tabla 20: Resultado SOFM de 16x16. Cluster 6.

Tipo	Hits	Valor
Centroide	29537	(171, 159, 199)
Vecino	1121	(182, 164, 21)
Vecino	631	(115, 117, 150)
Vecino	247	(52, 177, 4)
Vecino	436	(150, 85, 145)

Las páginas que componen el 6° cluster detectado corresponden a “servicios remotos” al que puede acceder el cliente, “simulador de crédito automotriz” correspondiente a un formulario y Fondos Mutuos. Luego, la idea centrales de las 3 componentes del cluster no son consistentes, por lo que son ignoradas en el estudio puesto que no guarda consistencia con algún perfil de clientes.

5.2.1.2.7. Análisis de Cluster 7.

Tabla 21: Resultado SOFM de 16x16. Cluster 7.

Tipo	Hits	Valor
Centroide	26166	(192, 153, 58)
Vecino	1188	(13, 201, 173)
Vecino	383	(140, 20, 161)
Vecino	349	(35, 201, 196)
Vecino	898	(24, 158, 109)

Los contenidos asociados al centroide pueden ser interpretados como una apertura o mirada hacia el mercado extranjero. Dos componentes del centroide corresponden al estado de la cuenta internacional y el pago de la deuda de esta. El tercer componente menciona las conveniencias de los productos de ahorro, principalmente los fondos mutuos de inversión en capitales nacionales y extranjeros. Para los 3 casos se menciona la palabra extranjero y la conveniencia de la inversión en un instrumento de este tipo, por lo que se pude perfilar este cluster como el grupo de personas que ha generado apertura hacia el mercado extranjero.

Tabla 22: Resultado SOFM de 16x16. Palabras encontradas en Cluster 7 .

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
deud	2,874131768	0,254166	96	2	deuda, deudas
person	0,768635	0,067972	135	2	persona, personas
pag	0,665825675	0,058881	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
cuent	0,482402924	0,04266	345	3	cuenta, cuentan, cuentas
internacional	0,308665768	0,027296	23	1	internacional
total	0,281381752	0,024883	52	2	total, totalidad
nombre	0,257156026	0,022741	91	1	nombre
descripcion	0,254636441	0,022518	11	1	descripcion
nacional	0,243087636	0,021497	28	1	nacional
fech	0,198306553	0,017537	101	2	fecha, fechas

Las palabras detectadas en el cluster quedan relacionadas principalmente en inversión nacional e internacional y en deudas y cuentas del mismo tipo. Esto sugiere que el grupo de usuario utiliza el banco para tener acceso al mercado extranjero principalmente para inversión.

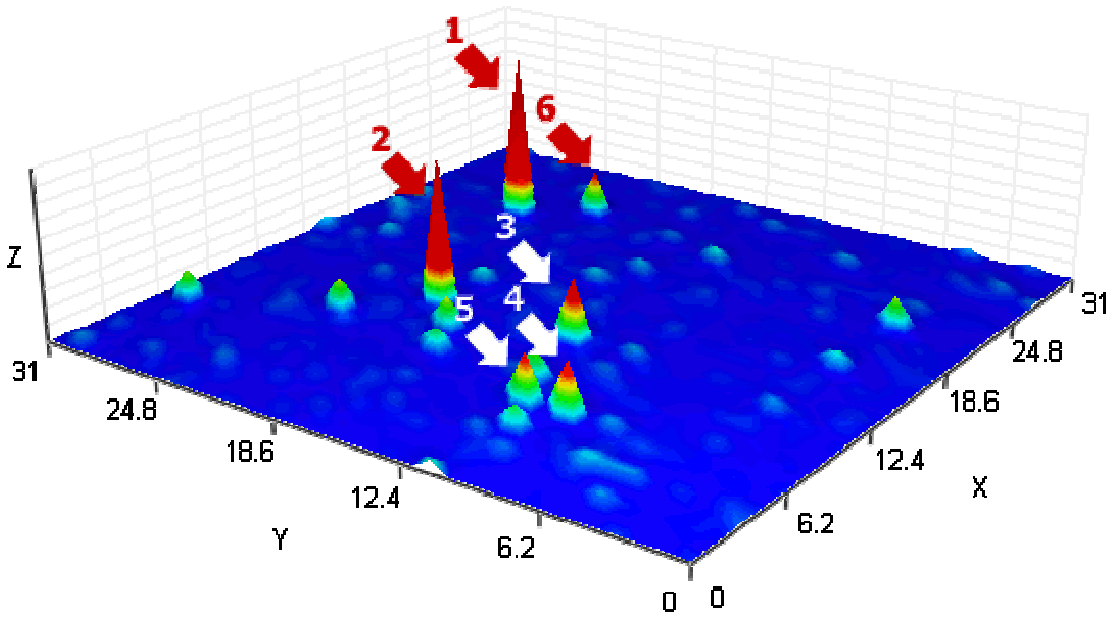
Como resultado final del análisis de clusters en el mapa de 16 x 16 se puede mencionar la detección de 5 grupos principales de perfiles de usuarios a los cuales se les puede asociar un set de palabras clave al momento de su búsqueda. La aplicación de ellas se sugiere en el siguiente capitulo.

Como parte de la investigación y análisis del trabajo realizado se realizará web mining con un mapa de 32 neuronas para revisar y comparar los resultados obtenidos del proceso anterior.

5.2.1.3. Análisis de Clusters para mapa SOFM de 32 x 32.

En relación a la investigación anterior, se resolvió reiterar el trabajo realizado pero sobre un mapa de mayor cantidad de neuronas con el fin de validar e identificar cuantos clusters se detectan sobre un cambio de estructura de salida en el análisis. La modificación fue el aumento de neuronas de salida a 32 con lo que el mapa queda como se indica en la figura:

Figura 30: ***Mapa de Clusters resultante del proceso de SOFM con 32 Neuronas.***



5.2.1.3.1. Análisis Cluster 1.

Las páginas pertenecientes al centroide corresponden a simuladores de crédito, simulador de retiro y revisión de servicios disponibles en Internet. Estas tres páginas muestran el perfil de usuarios que utiliza el sitio web para ver resultados de simulaciones y los servicios

que se prestan en una plataforma tecnológica.

Tabla 23: Resultado SOFM de 32x32. Palabras encontradas en Cluster 1.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
pagin	0,36497069	0,09279	17	2	pagina, paginas
ingres	0,20415	0,051903	47	8	ingresa, ingresada, ingresan, ingresando, ingresar, ingresas, ingreso, ingresos
remot	0,075496055	0,019194	66	3	remota, remotamente, remotos
contactan	0,071851897	0,018268	171	1	contactanos
map	0,071492451	0,018176	173	1	mapa
home	0,067922577	0,017269	188	1	home
retir	0,062995468	0,016016	53	2	retiro, retiros
benefici	0,048668207	0,012373	92	3	beneficiaras, beneficio, beneficios
simul	0,047498693	0,012076	39	3	simula, simulador, simuladores
total	0,044331502	0,011271	52	2	total, totalidad

Las palabras aludidas del centroide corroboran la interpretación según las páginas y muestran un perfil de usuarios interesado en simular, navegar y ver los beneficios disponibles de pertenecer a una banca con sucursal únicamente virtual.

5.2.1.3.2. Análisis Cluster 2.

Las páginas del centroide del segundo cluster detectado muestran un perfil poco claro ya que se mezclan en este centroide consulta por cheques pagados, pago de deuda internacional de la tarjeta de crédito y simulador de retiro. Estos tres conceptos no se observa que se relacionen entre sí en algún perfil de usuarios existente.

Las palabras detectadas en el centroide corroboran la inconsistencia de la información en su conjunto. Se puede por lo tanto obviar el cluster ya que no se identifica claramente el segmento de usuarios y las palabras que se pueden declarar como keywords son divergentes en contenido.

Tabla 24: Resultado SOFM de 32x32. Palabras encontradas en Cluster 2

Términos
total, totalidad
paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
nombre
cuenta, cuentan, cuentas
retiro, retiros
cheques
monto, montos
deuda, deudas
serie
talonario, talonarios

5.2.1.3.3. Análisis Cluster 3.

El centroide del cluster en análisis se relaciona con la utilización de la tarjeta de crédito ya que menciona el estado de movimientos nacionales de la tarjeta, ultimas compras realizadas y además muestra una componente relacionada con tributación de aportes que se aleja de las otras dos pero que con la revisión de las palabras se puede detectar con mayor precisión.

Tabla 25: Resultado SOFM de 32x32. Palabras encontradas en Cluster 3.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
hast	0,821444416	0,115053	124	1	hasta
mont	0,665926778	0,093271	165	2	monto, montos
facturacion	0,328952204	0,046074	9	1	facturacion
descuent	0,262373213	0,036749	18	4	descuenta, descuentan, descuento, descuentos
compr	0,262167618	0,03672	98	5	compra, comprando, comprar, compras, compro
nombre	0,257156026	0,036018	91	1	nombre
descripcion	0,254636441	0,035665	11	1	descripcion
fech	0,173271907	0,024269	101	2	fecha, fechas
decredit	0,16939826	0,023726	16	1	decredito
impuest	0,168739543	0,023634	93	2	impuesto, impuestos

Las palabras claves del centroide guardan relación con lo que son las compras, facturaciones, descuentos, montos e impuestos asociados a la utilización de instrumentos financieros como tarjetas de crédito. La relación de las palabras es consistente y se puede asociar a un perfil de usuarios interesados en hacer uso activo de su tarjeta de crédito.

5.2.1.3.4. Análisis Cluster 4

El centroide del cluster en análisis muestra nuevamente componentes relacionadas a un ámbito provisorio del cliente. Se detecta el análisis de últimas compras con tarjeta de crédito, beneficios de ahorrar en banco y una simulación de retiro. Una inferencia de lo que puede ser el cluster es la relación existente en personas que generan algún gasto u están interesados en ahorrar y ver como resulta el futuro

Tabla 26: Resultado SOFM de 32x32. Palabras encontradas en Cluster 4.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
comercial	0,253937024	0,065736	5	1	comercial
contactan	0,071851897	0,0186	171	1	contactanos
map	0,071492451	0,018507	173	1	mapa
home	0,067922577	0,017583	188	1	home
mont	0,065645597	0,016993	165	2	monto, montos
retir	0,062995468	0,016307	53	2	retiro, retiros
benefici	0,061238405	0,015853	92	3	beneficiaras, beneficio, beneficios
desde	0,044939405	0,011633	106	1	desde
total	0,044331502	0,011476	52	2	total, totalidad
compr	0,042186684	0,010921	98	5	compra, comprando, comprar, compras, compro

Las palabras al igual que el cluster 3 se puede asociar a un grupo de clientes provisorios analizan su costo de oportunidad de realizar un gasto o consultar por un ahorro.

5.2.1.3.5. Análisis Cluster 5.

Las páginas contenidas en este cluster son idénticas a las páginas detectadas en el centroide del cluster 1. Luego, el análisis de palabras es análogo al realizado a ese cluster.

5.2.1.3.6. Análisis Cluster 6.

Contiene páginas de simuladores de retiro y simulador de crédito de consumo, luego, se puede indicar un perfil de cliente interesado en simular acciones en la página del banco.

Tabla 27: Resultado SOFM de 32x32. Palabras encontradas en Cluster 6.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
retir	0,951198561	0,114961	53	2	retiro, retiros
impuest	0,505826884	0,061134	93	2	impuesto, impuestos
total	0,464213897	0,056104	52	2	total, totalidad
tribut	0,420274069	0,050794	10	1	tributable
rent	0,347878794	0,042044	65	2	renta, rentas
resumen	0,342603064	0,041407	2	1	resumen
calcul	0,301996889	0,036499	22	8	calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, calculos
explicacion	0,298270044	0,036049	3	1	explicacion
chilen	0,281815082	0,03406	13	2	chileno, chilenos
situacion	0,272337002	0,032914	4	1	situacion

Las palabras del centroide se asocian al proceso de simulación y cálculo de rentabilidades, créditos y retiros. Luego, el perfil de clientes es de interés en utilizar la página para aprender más de retiros y tener acceso a créditos.

5.2.1.4. Análisis de Clusters por K-means.

La aplicación de K-means permite generar una nueva vista del análisis de cluster, centroides y palabras y adicionalmente puede validar el resultado de los mapas construidos anteriormente. El valor del número de clusters según el procesos SOFM anteriormente realizado fue 7. Luego, se asigno con el mismo valor la iniciación de los centroides en el algoritmo.

5.2.1.4.1. Análisis Cluster 1.

Las páginas contenidas en el centroide corresponden a un grupo asociado a consulta de estado de productos y servicios adquiridos en el banco. Principalmente estado de la cuenta, consulta de crédito hipotecario y la página inicial del banco.

Tabla 28: Resultado K-means. Palabras encontradas en Cluster 1.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
garanti	**0,33404224**	**6%**	**18**	**2**	garantia, garantias
cliente	0,286874545	5%	**186**	**1**	cliente
ahor	**0,218144128**	**4%**	**8**	**1**	ahora
sald	0,201420775	4%	**80**	**2**	saldo, saldos
pag	0,184834191	3%	**347**	**13**	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
segur	0,184460321	3%	**112**	**5**	segura, seguras, seguridad, seguro, seguros
fech	0,137609642	3%	**101**	**2**	fecha, fechas
product	0,118362688	2%	**170**	**2**	producto, productos
plaz	0,101236217	2%	**204**	**3**	plaza, plazo, plazos
fond	0,086735385	2%	**477**	**2**	fondo, fondos

Las palabras clave detectadas son asociadas al perfil de clientes descrito anteriormente y se asocian finalmente a su interés en la consulta de estado de productos, pago, garantías y adquisición de seguros.

5.2.1.4.2. Análisis Cluster 2.

El centroide muestra páginas consultadas frecuentemente en una navegación hacia consultas de estado de créditos ya que se encuentran principalmente las páginas cartola y créditos.

Tabla 29: Resultado K-means. Palabras encontradas en Cluster 2

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
line	**0,61583959**	**17%**	**251**	**2**	linea, lineas
internet	0,257601853	7%	**78**	**1**	internet
envi	**0,198727569**	**5%**	**19**	**7**	envia, envian, enviando, enviar, enviarla, envio, envios
cliente	0,180861	5%	**186**	**1**	cliente
fond	0,086735385	2%	**477**	**2**	fondo, fondos
plaz	0,080504109	2%	**204**	**3**	plaza, plazo, plazos
hazte	0,074071816	2%	**98**	**1**	hazte
retenciones	0,063983096	2%	**32**	**1**	retenciones
cartol	0,058195125	2%	**26**	**1**	cartola
inversiones	0,051975741	1%	**284**	**1**	inversiones

Las palabras del centroide son asociadas a un cliente interesado o en búsqueda de información de clientes interesados en sus fondos e inversiones.

5.2.1.4.3. Análisis Cluster 3.

El centroide muestra páginas de un cliente en consulta de obtención de un plan o servicios del banco, una tercera página de entrega como resultado el error del servidor, por lo que se puede tratar de un cliente en consulta de obtención de servicios del banco pero que no encontró lo que buscaba.

Tabla 30: Resultado K-means. Palabras encontradas en Cluster 3.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
integral	**0,408238227**	**11%**	**40**	**1**	integral
plan	0,31532719	8%	**61**	**2**	plan, plana
segur	**0,128016276**	**3%**	**112**	**5**	segura, seguras, seguridad, seguro, seguros
product	0,118362688	3%	**170**	**2**	producto, productos
clave	0,089884602	2%	**53**	**1**	clave
remot	0,075496055	2%	**66**	**3**	remota, remotamente, remotos
hazte	0,074071816	2%	**98**	**1**	hazte
atencion	0,065719428	2%	**107**	**1**	atencion
inversiones	0,065192404	2%	**284**	**1**	inversiones
password	0,064631368	2%	**28**	**1**	password

Las palabras claves detectadas dan cuenta de que el usuario estaba en busca de información de un plan integral, productos del banco y utilidades de este en el formato virtual. Las palabras plan, integral, hazte y productos dan cuenta del perfil de interés del cliente.

5.2.1.4.4. Análisis Cluster 4.

El grupo de páginas en este caso perfila el interés sobre la adquisición de un crédito hipotecario. Las páginas asociadas mencionan lo que es la simulación del crédito, consulta de crédito y tributaciones.

Tabla 31: Resultado K-means. Palabras encontradas en Cluster 4.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
impuest	**0,168739543**	**6%**	**93**	**2**	impuesto, impuestos
hast	0,12019333	4%	**124**	**1**	hasta
entre	**0,081134715**	**3%**	**91**	**1**	entre
pag	0,077054223	3%	**347**	**13**	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
minim	0,071870506	3%	**89**	**2**	minima, minimo
hipotecari	0,066052839	2%	**29**	**2**	hipotecario, hipotecarios
tributacion	0,046289833	2%	**10**	**1**	tributacion
mont	0,045545166	2%	**165**	**2**	monto, montos
maxim	0,044817233	2%	**84**	**2**	maxima, maximo
convenientre	0,044583298	2%	**16**	**1**	conveniencia

Las palabras en el centroide dan cuenta de este interés y muestran las consultas sobre la simulación de un crédito hipotecario.

5.2.1.4.5. Análisis Cluster 5.

Las páginas del cluster 5 indican clientes con interés en consultas a sistemas provisionales, principalmente la forma de funcionamiento del sistema provisional chileno.

Tabla 32: Resultado K-means. Palabras encontradas en Cluster 5.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
deposit	**0,089252961**	**3%**	**147**	**7**	depositado, depositados, depositantes, depositar, deposito, depositos, depósitos
remot	0,075496055	2%	**66**	**3**	remota, remotamente, remotos
hazte	**0,074071816**	**2%**	**98**	**1**	hazte
ahorr	0,067411685	2%	**314**	**7**	ahorra, ahorrado, ahorran, ahorrantes, ahorrar, ahorro, ahorros
fond	0,065469628	2%	**477**	**2**	fondo, fondos
inversiones	0,065192404	2%	**284**	**1**	inversiones
person	0,058090205	2%	**135**	**2**	persona, personas
sistem	0,051757162	2%	**31**	**2**	sistema, sistemas
integral	0,049908625	2%	**40**	**1**	integral
minim	0,045355182	1%	**89**	**2**	minima, minimo

Las palabras en este centroide muestra información de la información requerida por el cliente interesado en previsión. Palabras como inversión, ahorro, depósitos muestran que pretende adquirir o esta interesado en tener su dinero en alguna cuenta de ahorro o inversión para su pensión.

5.2.1.4.6. Análisis Cluster 6.

Las páginas del 6º centroide detectado hacen mención a créditos de consumo e hipotecarios, que son las paginas que visitó el cliente en su estadía.

Tabla 33: Resultado K-means. Palabras encontradas en Cluster 6.

Palabra	Peso	%	# repeticiones	# términos asociados	Términos asociados
segur	**1,14632269**	**11%**	**112**	**5**	segura, seguras, seguridad, seguro, seguros
autoconsult	0,934595989	9%	**103**	**1**	autoconsulta
atencion	**0,712754**	**7%**	**107**	**1**	atencion
ejecut	0,707327203	7%	**126**	**3**	ejecutan, ejecutivo, ejecutivos
minim	0,540733158	5%	**89**	**2**	minima, minimo
graci	0,461534389	4%	**7**	**1**	gracia
disponibil	0,37860301	4%	**15**	**1**	disponibilidad
maxim	0,345934054	3%	**84**	**2**	maxima, maximo
garanti	0,33404224	3%	**18**	**2**	garantia, garantias
libre	0,333823416	3%	**15**	**1**	libre

Las palabras muestran interés del usuario para la adquisición de seguros, meses de gracia o garantías, lo que hace pensar que son casos de clientes interesados en la adquisición de un crédito de forma segura y con respaldo y garantías.

CAPITULO 5

6. RECOMENDACIONES PARA MEJORAS DE UN SITIO WEB.

Como producto del trabajo realizado, se lograron identificar palabras asociadas a ciertos perfiles de clientes. Estas palabras pueden ser identificadas como palabras clave del sitio web para estos grupos de usuarios que ingresan y utilizan el sistema.

Las recomendaciones que se pueden asociar principalmente a dos ítems:

- Recomendaciones a nivel Estructural: Como parte del trabajo, fueron detectados varios inconvenientes en la estructura del sitio web como el resultado de ciertas prácticas de construcción y diseño. Las recomendaciones en este nivel se asocian a que es lo que se puede hacer en el website para mejorar la experiencia de navegación del cliente.
- Recomendaciones a nivel Textual: Que es la aplicación de las palabras identificadas durante el proceso a diferentes contextos del sitio web según la dependencia existente entre ellas, por ejemplo las palabras asociadas a crédito pueden ser mostradas al usuario durante la consulta de sus deudas o pago.

6.1. Recomendaciones para posibles mejoras estructurales del sitio web.

Se toma como recomendación inicial el ámbito estructural para mejora ya que esta es parte importante del soporte que posteriormente contendrá el contenido de la página ya sea por ser parte del contexto o por ser producto de la recomendación de este trabajo al perfil del cliente.

Las recomendaciones estructurales se asocian a la forma de la página web y de cómo mejorar las prácticas esencialmente de diseño para obtener una página que sea óptima para almacenar los contenidos y para que principalmente sea el usuario el que encuentre lo que esta buscando.

Como se mencionó anteriormente, este tipo de recomendaciones se extrae de los inconvenientes encontrados en el trabajo y se asocian particularmente a alguna acción del cliente para hacerlo.

El listado de prácticas que a continuación se mencionan es útil como referencia principalmente a diseñadores para desarrollar el sitio web.

6.1.1. Recomendaciones configuración de servidor.

En el ámbito de configuración del servidor, que es base para el levantamiento del website, la recomendación es principalmente por el lado de los registros que se almacenan en el web log. Los datos o campos que se almacenen en este registro deben proveer de la información necesaria para aplicar posteriores investigaciones de patrones de comportamiento de los clientes. Los campos mínimos necesarios y en el orden que se necesitan son:

Tabla 34: Tabla de recomendaciones de configuración del servidor.

Campo	Descripción
host	Debe contener la IP o Nombre del servidor desde donde se ingresa. Este campo es esencial para la reconstrucción de sesiones de forma reactiva.
user	En caso que el sitio requiera autenticación este campo puede ser util para identificar efectivamente al usuario que hace el ingreso
fecha	Este campo es obligatorio en un web log ya que a parte de permitir ordenar y obtener estadísticas generales de navegación en el sitio web, permite calcular los tiempos de navegación entre páginas
formato	Este campo no es obligatorio, pero permite saber que tipo de requerimiento es (si es POST o GET)
path	Importante en la identificación de los requerimientos del usuario. Es obligatorio en el formato carpeta/carpeta/…/archivo.algo. Se debe evitar agregar el querystring a este campo ya que genera error en la identificación de páginas.
versión	La versión del requerimiento entregado no es obligatoria
respuesta servidor	La respuesta del servidor es importante para conocer el resultado del requerimiento y al mismo tiempo identificar requerimientos nulos o con error.
bytes	Los bytes transferidos no son un campo obligatorio pero entrega información del volumen de los archivos transferidos.
referer	El referer es un campo que actualmente no fue utilizado en el trabajo, pero puede ser importante en la investigación de las secuencias de navegación e identificación de patrones de navegación para la identificación de orígenes de acceso a las páginas.
agent	El componente Agent es esencial en la identificación del tipo de usuario o cliente que accesa e incluso permite obtener información de si es un cliente “humano” o un robot.
otros	Se sugiere en general la evasión de inclusión de campos adicionales a no ser que agreguen información a ser analizada o incluida en estudios posteriores. Un campo que puede ser de importante utilidad es el querystring, que contiene información de traspaso de variables. Este campo puede ser muy útil al momento de agregar variables en la identificación y perfilación de los clientes como nombre, origen, etc.

Como se menciono y como regla general de este aspecto, se sugiere configurar la actividad de registro de requerimientos con los campos justos y necesarios e intervenir el web log solo si el caso es estrictamente necesario.

6.1.2. Recomendaciones estructurales del sitio web y sus páginas.

El sitio web, compuesto por páginas web tiene diversos ámbitos de configuración los cuales

están afectos a dificultar la navegación del usuario. Se propone mejorar en algunos casos o evitar ciertas prácticas que dan complejidad al sitio pero al usuario final le impide realizar acciones o llegar a la información de forma oportuna y/o eficiente.

6.1.2.1. **Redirección**: Algunos sitios web inician el proceso de navegación desde una url hacia otra que generalmente es una subcarpeta del mismo origen. Este tipo de acción se realiza mediante códigos en javascript o módulos de flash los cuales generan problemas al intentar detectar la importancia de la página. Si se considera que uno de los principales orígenes de ingreso al sitio web es el resultado en buscadores, una redirección en el proceso inicial puede dificultar la obtención de los contenidos del interior del sitio web. La sugerencia se encuentra en que se debe iniciar el sitio web en un portal que en el caso de que deba redireccionar, se apoye con un link la posibilidad de redirección pero mediante etiquetas html.

6.1.2.2. **Frames**: Los frames o marcos en un sitio web dan la posibilidad de trabajar en un sitio web que se mueve en ventanas independientes. Generalmente los marcos son configurados como Top o encabezado, Menú y Cuerpo. Esto puede dar una apariencia de panel de control para el usuario que utiliza la página web, pero el inconveniente que se genera es al momento de obtener los contenidos ya que la página que contiene los marcos que es generalmente la de inicio no tiene contenidos sino fuentes de los marcos. Luego, robots de búsqueda de contenido pueden dar con el sitio web que en su estructura no ofrece contenidos y por lo tanto puede ser mal rankeada o indexada al motor.

6.1.2.3. **Hojas de Estilos** (Palabras especiales): La utilización de hojas de estilo permite dar un diseño almacenado en un archivo en el sitio web. La arbitrariedad de elección de nombres de clases o identificadores para otorgar ciertos diseños a partes del contenido del sitio web hacen muy difícil la identificación de cuando una palabra es destacada respecto a las otras a nivel de etiquetado html. Se recomienda utilizar etiquetas html para palabras destacadas ya que la configuración de ellas es universal y permite identificar las palabras destacadas de un sitio web.

6.1.2.4. **Animaciones y funciones precompiladas:** El impacto gráfico que presenta un banco se apoya de herramientas como flash o applets de java en las cuales también se hacen ofertas de productos y servicios, pero, dado que es "precompilado", no permite obtener la información desde su modulo a no ser que se "decompile". La utilización de este tipo de apoyos para contención de contenidos no se recomienda ya que los motores de búsqueda pueden no encontrar el contenido en el archivo precompilado por lo que se pierde este índice al momento de coincidir con una consulta del cliente. Una recomendación paralela puede ser que se haga uso de estos módulos en el sitio web pero se haga uso de la etiqueta META para indicar las keywords asociadas a estos objetos.

6.1.2.5. **Archivos y Documentos**: Google tiene la posibilidad de hacer búsqueda en la red y entregar resultados de las páginas web hasta documentos como pdf,

ppt, doc y xls. Pero no siempre estos contenidos son consultados por el cliente por la demora de la obtención de la data porque un archivo PDF puede pesar incluso MB lo cual lo hace poco atractivo. Se sugiere la utilización de la etiqueta meta en su valor keyword para indicar los principales conceptos asociados al documento para mostrar a priori el contenido que se encuentra antes de ser desplegado al usuario.

6.1.2.6. **Meta Tags (Keywords):** De los sitios web analizados prácticamente no se utilizan los meta tags, por lo que se desaprovecha una etiqueta que puede contener palabras importantes de la página o incluso metadatos que hagan referencia a objetos que se encuentran en la página en revisión.

6.1.2.7. **Dominio**: Los sitios web en algunas ocasiones son accesados a través del dominio del sitio web (por ejemplo www.ejemplo.org), pero al ingreso son redireccionados a rutas diferentes, por ejemplo http://servidor.hosting.com/carpeta/.../pagina.algo. Esta redirección hace que el contenido finalmente pueda no ser debidamente asociado al dominio del sitio y por ende, el usuario encuentre contenido pero no asociado al dominio que espera.

6.1.2.8. **Uso de caracteres especiales:** Una página web soporta muchos diseños en su estructura, pero en algunas ocasiones los diseñadores utilizan caracteres especiales que son incluidos en su diseño, por ejemplo, la utilización del amperson (&) para separar contenidos o del porcentaje (%) para separar líneas o palabras. Esto hace que los procesos de identificación de palabras sean complejos o con mucha dificultad y por esto un motor de búsqueda puede no lograr identificar el contenido o palabra importante en el sitio web.

6.2. Recomendaciones de uso de palabras claves en el sitio web.

Como resultado del trabajo realizado fue posible identificar segmentos de clientes del sitio web y las palabras claves a las cuales normalmente ingresa en su navegación. Estas palabras llamadas palabras claves o *website keywords* son recomendables para el grupo de clientes asociado por lo que la recomendación va por el lado de inicialmente identificar el perfil del cliente y luego entregar el set de palabras especiales recomendadas.

Las palabras identificadas como clave no deben ser colocadas solamente de forma individual en las páginas del sitio donde se desee realizar la acción de incremento de interés en la página del usuario, sino que deben ir contextualizadas en una frase o texto adicional. Se puede utilizar el meta tag para agregar estas palabras a la página individualmente.

El set de recomendaciones se indica según los siguientes campos:

Tipo de cliente: Corresponde al perfil aproximado del cliente identificado y a quien es óptimo aplicar las palabras claves encontradas.

Palabras: Es el set de palabras clave identificada para el segmento.

Contexto: Aproximación de uso de las palabras en un contexto de interés para el tipo de cliente identificado.

Cuadro 7: ***Recomendaciones de palabras claves o website keywords según el perfil del cliente.***

Cliente	Palabras	Contexto
Heavy Users de Tarjetas de Crédito	actual, actualidad canje, canjeados, canjeando, canjear, canjearlos, canjearse paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues total, totalidad compra, comprando, comprar, compras, compro calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, cálculos costado, costo, costos saldo, saldos anterior peso, pesos	Compras, adquisición de puntos, canje de puntos, catálogo.
Clientes ad portas de jubilación o interesados en ella	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues impuesto, impuestos total, totalidad calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, calculos renta, rentas ingresa, ingresada, ingresan, ingresando, ingresar, ingresas, ingreso, ingresos misma, mismas, mismo, mismos hasta fondo, fondos retenciones	Jubilación, planes de inversión, APV
Clientes Heavy User de instrumentos del banco como tarjetas y cheques. Empresarios	nombre paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues cuenta, cuentan, cuentas corriente total, totalidad cheques estado, estados informacion contable fecha, fechas	Pago de cuentas, pago de deudas, cobertura de cheques, información de cuenta y transacciones.

Inversionistas	monto, montos deuda, deudas inversion, inversionistas estado, estados fecha, fechas compra, comprando, comprar, compras, compro solicitud cuenta, cuentan, cuentas fondo, fondos futura, futuro, futuros	Acceso a liquidez, continuidad de la inversión, respaldo monetario.
Clientes con interés en inversión al extranjero.	deuda, deudas persona, personas paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues cuenta, cuentan, cuentas internacional total, totalidad nombre descripcion nacional fecha, fechas	Inversion, mercado de capitales.

Las recomendaciones anteriormente presentadas muestran que sobre ciertos perfiles de clientes es posible establecer acciones proactivas para mejorar la calidad del servicio que el cliente ve, por lo que dará mayor uso a la página y estará al tanto de las acciones proactivas que la organización tiene para él.

Para una mejora práctica y aplicación de estas recomendaciones se pueden utilizar herramientas de data mining para identificar el perfil que de mejor manera se ajusta al cliente para luego asociarle el set de palabras en las que estaría interesado, como lo muestra la tabla anterior.

Este set de recomendaciones es con respecto a la página del banco y sus registros en web logs. Luego, esas recomendaciones deben ser específicas en la organización y no para toda la industria ya que se realizan dado los perfiles y comportamientos de los clientes de la organización en estudio.

6.3. Testeo de la efectividad de las recomendaciones de texto.

Como se mencionó con anterioridad, la detección y aplicación de no garantizan el éxito de aplicación de estas palabras en un contenido textual. Incluso, el riesgo de utilizarlas puede generar disgusto en un usuario habitual del website y por lo tanto abandonar o dejar de usar con frecuencia el website.

Como medida precautoria, se realizaron test de efectividad de las website keywords. Sobre el contenido del sitio web se extrajeron 10 párrafos que contenían para el caso de 5 de ellos website keywords y otros no las contenían. El resultado se realizó sobre un universo de 10 personas con

el fin de conocer la recepción que ellos tenían respecto a párrafos que contenían las palabras detectadas según el contexto de si entregaban información relevante en un sitio bancario. El resultado de este test es el que se muestra en la tabla 36.

Tabla 35: párrafos testeados para análisis de keywords.

#	incluye website keywords	Párrafo
1	Si	Orientado a empresas que deseen manejar excedentes de caja, así como a Personas que quieran mantener parte de sus recursos **invertidos** en un **fondo mutuo**, cuya cartera esté compuesta exclusivamente por instrumentos de **deuda** nacional, obteniendo rentabilidad y liquidez a corto plazo.
2	Si	Solicitándolos con un día hábil bancario de antelación, se pagarán mediante **cheques** nominativos, **vales** vista o **depósitos** en **cuentas** corrientes, de acuerdo a sus instrucciones.
3	Si	Este plan busca otorgar a tus **Ahorros Previsionales** Voluntarios acumulados a esta fecha y los futuros, una atractiva y segura **rentabilidad** que te permitirá poder mejorar considerablemente tus **ahorros** para una mejor **pensión**.
4	Si	Para obtener información de tu **Cuenta** Corriente y de tu **Línea** de **Sobregiro** debes seguir los siguientes pasos
5	Si	El Servicio de Mensajería es un servicio de entregas y **retiros** de dinero, especies valoradas y documentos que podrás utilizar siendo **cliente**
6	No	Para solicitar tu Plan Integral debes completar la siguiente información y un Ejecutivo se contactará contigo.
7	No	En un sólo lugar tienes la posibilidad de invertir en una amplia gama de productos de acuerdo a tus necesidades y el futuro que quieres proyectar.
8	No	En este caso, la diferencia de $1000 está directamente relacionada con la tasa a la que está afecto. Mientras mayor la tasa mayor es el beneficio.
9	No	Nunca más tendrás que ir al banco ya que para cada una de las transacciones bancarias hemos diseñado servicios de atención remotos disponibles las 24 horas
10	No	Usted ahora puede ver su minicartola en una planilla de cálculo como MS-Excel, para ello debe seguir los siguientes pasos.

Como se puede apreciar, aquellas palabras que contenían website keywords eran para el usuario mucho más interesantes e importantes en el contexto de navegación en que estaban inmersos, versus aquellos párrafos en que no había presencia de dichos website keywords.

Tabla 36: Testeo de efectividad de las palabras claves de un sitio web o website keywords.

#	incluye website keyword	Opinión de aceptabilidad				
		Irrelevante	Moderadamente irrelevante	Algo relevante	Moderadamente relevante	Relevante.
1	Si				8	2
2	Si			4	4	2
3	Si			4	2	4
4	Si				7	3
5	Si			1	2	7
6	No	1	3	5	1	
7	No	3	2	5		
8	No	6	4			
9	No	5	2	1	2	
10	No	7	2	1		

Las *website keywords* atraen la atención del usuario y pueden ser una muy buena guía en el diseño de contenidos específicos de un sitio web. Esta combinación de elementos que se alinean a los que el usuario busca pueden otorgar un mejor resultado en la satisfacción de los clientes.

6.4. Importancia de la aplicación de mejoras en el sitio web para las organizaciones.

Las mejoras de los textos de las páginas de un sitio web no tan sólo se traducen en un sitio web óptimo a las necesidades del usuario, sino también en una optimización de un canal de venta o auto atención.

El utilizar las palabras clave según las preferencias del usuario del sitio web generarán un incremento en las visitas no tan solo por tratarse de lo que el usuario busca y por ende una mayor tasa de visitas, sino también por ser coincidente con las búsquedas que realice el usuario en motores de búsqueda como Google o Yahoo!, los que serán nuevos clientes o clientes fidelizados.

El incremento de las visitas de un sitio web variará según la industria que se estudie. El caso de los sitios web B2C[16] pueden verse mayormente afectados con modificaciones textuales por el tráfico que en ellos incurre, en cambio un sitio web B2B[17] puede ser de un incremento menor dado que las aplicaciones de esta naturaleza estan orientadas a relaciones o alianzas con clientes preestablecidas. Una referencia de las mejoras textuales traducidas en mejoras en los ingresos y disminución de costos se hace a continuación:

Aumento de visitas, nuevas ventas, nuevos clientes: Según la industria que se trate, existe una tasa de visitas al sitio web que se transforman en una venta o en un nuevo cliente por medio de una tasa de conversión (convertion ratio). Luego, al tratarse de una mejora textual que incrementa las visitas de un sitio web, mediante la tasa de conversión se puede observar un posible incremento de ventas o un aumento de nuevos clientes.

Mayor uso del sitio, menos infraestructura: Un aumento de visitas y uso del sitio web hace que servicios como pago de cuentas, servicios adicionales, o transacciones se canalicen por

[16] B2C: Business to Consumer

[17] B2B: Business to Business

Internet. Esta canalización otorga la posibilidad de disminuir costos de mantención e infraestructura puesto que porcentajes de clientes que anteriormente eran atendidos presencialmente pueden optar por comodidad y facilidad a auto atenderse en una sucursal virtual. En el caso de los bancos, por ejemplo, las transferencias de fondos al poder ser realizadas por Internet, evitan visitas al banco puesto que este proceso es automático. Una disminución del flujo en el banco genera una proyección de transacciones presenciales menores y por ende una menor cantidad de turnos de cajeros o de cajas disponibles.

La traducción de estas mejoras a valores monetarios dependerán de la industria en que se encuentre inmerso el sitio web en análisis pues dependerá del uso y demanda del sitio web, los servicios disponibles que pueden ser "trasladados" a la red y las proyecciones de costo unitario por cliente que tenga la empresa para lograr identificar numéricamente la mejora del sitio web.

CAPITULO 6

7. CONCLUSIONES

La evolución de las Tecnologías de Información y Comunicaciones han generado cambios sociales y económicos en los 40 años de evolución que ha tenido. Por esto las empresas han ajustado sus modelos de operación de acuerdo a estas nuevas tendencias. Internet, una de las tecnologías que ha realizado gran parte de estos cambios, es la más utilizada al momento de publicar información corporativa o de los productos y servicios que se ofrecen por la posibilidad de interacción directa con el cliente.

El cliente llega a un sitio web a través de diversos medios, pero el principal, con un 33% de preferencias, es la utilización de buscadores como Google o Yahoo! que a través de un formulario despliegan los resultados según un término buscado. Sobre el 58% de los usuarios consultan sólo la primera página de resultados, por lo que la importancia de aparecer en esta es la prioridad de los encargados de los sitios web de la empresas y corporaciones. Para esto se utilizan diversas técnicas de optimización en buscadores o SEO (Search Engine Optimization) que consideran aspectos como arquitectura, contenido, interacciones y multimedia.

Las palabras que se encuentran en el contenido textual serán las que podrían coincidir con la búsqueda del usuario. Estas palabras coincidentes se denominan *website keywords* o *palabras clave*. Con la identificación de estas palabras clave generando un conjunto de sugerencias de mejoras textuales de las páginas del sitio web y a través de la aplicación del proceso KDD se logró estructurar la información de la web o *web data* para la aplicación de algoritmos de identificación de preferencias del usuario y desde estas, las páginas relevantes del grupo de usuarios asociado. El análisis de las palabras en las páginas de estos grupos o clusters y la detección de aquellas palabras que mayor peso tienen corresponde a la identificación de las palabras claves del sitio web.

Un test de efectividad de las palabras clave permitió corroborar la importancia que estas palabras tienen en los textos del sitio web y, por lo tanto, se generó con ellas un set de palabras clave que pueden ser base de las mejoras textuales del sitio web. La conversión de estas mejoras a un ámbito lucrativo dependerá del tipo de negocio que se prestará a través del sitio web.

Se obtuvo además en el trabajo realizado una metodología y las herramientas facilitadoras para la búsqueda y análisis de las palabras clave de un sitio web o *website keywords*.

En la identificación de las palabras se trabajó sobre los registros de un banco virtual, desde donde se logró apreciar el comportamiento del usuario a través de los *web logs* y el contenido que soportaba ese comportamiento a través del texto libre de las *web pages*. Con la utilización de los módulos desarrollados se obtuvieron varias agrupaciones o clusters de usuarios que navegaron por el sitio web. En estos clusters fue posible identificar las palabras de mayor importancia y, con esto, las palabras claves de los grupos de usuarios del banco y su relevancia en el website. Palabras como crédito, pago, facturación son algunas de las *keywords* que fueron identificadas y que son aplicables a los perfiles de usuarios correspondientes según el interés que demuestren durante la navegación.

La detección de las palabras claves de un sitio web permitirán a un diseñador o administrador de contenidos a mejorar la calidad de textos libres dentro del sitio y lograr con ello un mayor acercamiento a los requerimientos del usuario con lo cual los buscadores como Google o Yahoo! entregarán dentro de los primeros resultados que consulta el usuario la página web con las mejoras y aciertos en texto. Estas recomendaciones no garantizan un éxito total después de su implementación por tratarse de un medio dinámico. El trabajo para una optimización en buscadores debe ser constante pues las empresas competidoras, asociadas al término trabajado se encontrarán también con la meta de aparecer dentro de los primeros lugares. El cliente será el beneficiado pues verá satisfecha de forma más puntual las necesidades que busca en los primeros resultados de los buscadores por coincidir con los intereses que tiene.

Se recomienda por lo tanto la utilización de esta técnica de identificación de palabras importantes para el usuario a aquellas organizaciones que estén interesadas en lograr que el sitio web de la organización cumpla con las expectativas de su mercado objetivo.

Se espera que futuros trabajos dentro del mismo ámbito también comiencen por una parte a incluir los objetos no considerados en este trabajo como imágenes o sonidos que pueden ser parte de la atracción del usuario, o bien conocer otra forma de interés en páginas web que no sea por contenidos sino por los colores que se presentan en la pagina.

Este trabajo, si bien es un interesante y gran avance en materia de investigación en el campo

del web mining, es recién una primera etapa de la investigación que puede ser posible a través de esta técnica.

Quedan muchas más posibilidades de investigación y este trabajo, sin duda, resulta ser un buen comienzo para mejorar de forma concreta un sitio web en su contenido textual.

El web mining actualmente no es mencionado ni utilizado con frecuencia, pero con el crecimiento de Internet y su evolución a la web 2.0 ya no será solamente importante estar en la World Wide Web, sino que también será importante el qué se muestra en el sitio web y cómo se sabe lo que le interesa al cliente o usuario cuando visita un sitio.

El trabajo desarrollado pretende además ser la base de investigaciones futuras del tema y proveer de herramientas las investigaciones de web mining.

BIBLIOGRAFÍA Y REFERENCIAS.

[1] E. Amitay and C. Paris. Automatically summarizing websites: Is there any way around it? In *Procs. of the 9th Int.*

Conf. on Information and Knowledge Management, pages 173–179, McLean, Virginia, USA, 2000.

[2] R. Baeza-Yates. *Web usage mining in search engines*, chapter Web Mining: Applications and Techniques, pages 307–321. Idea Group, 2004.

[3] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In *Proc. in First Int. Semantic Web Conference*, pages 264–278, 2002.

[4] B. Berendt and M. Spiliopoulou. Analysis of navigation behavior in websites integrating multiple information systems. *The VLDB Journal*, 9:56–75, 2001.

[5] D. Buttler. A short survey of document structure similarity algorithms. In *Procs. Int. Conf. on Internet Computing*, pages 3–9, 2004.

[6] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Focused web searching with pdas. *Computer Networks*, 33(1- 6):213–230, June 2000.

[7] L. D. Catledge and J. E. Pitkow. Characterizing browsing behaviors on the world wide web. *Computers Networks and ISDN System*, 27:1065–1073, 1995.

[8] G. Chang, M. Healey, J. McHugh, and J. Wang. *Mining the World Wide Web*. Kluwer Academic Publishers, 2003.

[9] W. Chuang and J. Yang. Extracting sentence segment for text summarization? a machine learning approach. In *Procs. Int. Conf. ACM SIGIR*, pages 152–159, Athens, Greece, 2000.

[10] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowlegde and Information Systems*, 1:5–32, 1999.

[11] U. Hahn and I. Mani. The challenges of automatic summarization. *IEEE Computer*, 33(11):29–36, 2000.

[12] A. Joshi and R. Krishnapuram. On mining web access logs. In *Proc. of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 63–69, 2000.

[13] A. P. Jr and N. Ziviani. Retrieving similar documents from the web. *Journal of Web Engineering*, 2(4):247–261, 2004.

[14] D. Lawrie, B. W. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *Proc. 24th Int*

SIGIR Conf. on Research and Development in Information Retrieval, pages 349–357, New Orleans, Louisiana, USA,

2001. ACM Press.

[15] E. Liddy, K. McVearry, W. Paik, E. Yu, and M. McKenna. Development, implementation and testing of a discourse model for newspaper texts. In *Procs. Int. Conf. on ARPA Workshop on Human Language Technology*, pages 159–164, Princeton, NJ, USA, 1993.

[16] G. Linoff and M. Berry. *Mining the Web*. Jon Wiley & Sons, New York, 2001.

[17] S. Loh, L. Wives, and J. P. M. de Oliveira. Concept based knowledge discovery in texts extracted from the web.

SIGKDD Explorations, 2(1):29–39, 2000.

[18] I. Mani and M. Maybury. *Advances in automatic text summarization*. MIT Press, Cambridge, Mass., 1999.

[19] S. Mitra, S. K. Pal, and P. Mitra. Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 13(1):3–14, 2002.

[20] B.Mobasher, R. Cooley, and J. Srivastava. Creating adaptive websites through usage-based clustering of urls. In *Procs.Int Conf IEEE Knowledge and Data Engineering Exchange*, November 1999.

[21] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[22] M. F. Porter. An algorithm for suffix stripping. *Program; automated library and information systems*, 14(3):130–137, 1980.

[23] T. A. Runkler and J. Bezdek. Web mining with relational clustering. *International Journal of Approximate Reasoning*, 32(2-3):217–236, Feb 2003.

[24] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*

archive, 18(11):613–620, November 1975.

[25] M. Spiliopoulou. Data mining for the web. In *Principles of Data Mining and Knowledge Discovery*, pages 588–589, 1999.

[26] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, 15:171–190, 2003.

[27] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[28] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. Recovering traceability links in multilingual websites. In *Procs. Int Conf. Website Evolution*, pages 14–21. IEEE Press, 2001.

[29] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. Restructuring multilingual websites. In *Procs. Int. Conf. Software Maintenance*, pages 290–299. IEEE Press, 2002.

[30] J. D. Velázquez and V. Palade. A knowledge base for the maintenance of knowledge extracted from web data. *Journal of Knowledge Based Systems (Elsevier)*, page to appear, 2007.

[31] J. D. Velásquez, S. R´ıos, A. Bassi, H. Yasuda, and T. Aoki. Towards the identification of keywords in the website text content: A methodological approach. *International Journal of Web Information Systems*, 1(1):11–15, March 2005.

[32] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. A methodology to find website keywords. In *Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, pages 285–

292, Taipei, Taiwan, March 2004.

[33] J. D. Velásquez, H. Yasuda, and T. Aoki. Combining the web content and usage mining to understand the visitor behavior in a website. In *Procs.* 3th *IEEE Int. Conf. on Data Mining*, pages 669–672, Melbourne, Florida, USA, November 2003.

[34] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. Using the kdd process to support the website reconfiguration. In *Procs. IEEE/WIC Int. Conf. on Web Intelligence*, pages 511–515, Halifax, Canada, October 2003.

[35] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a website. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396, February 2004.

[36] J. Xiao, Y. Zhang, X. Jia, and T. Li. Measuring similarity of interests for clustering web-users. In *ADC '01: Proceedings of the 12th Australasian conference on Database technologies*, pages 107–114, Washington, DC, USA, 2001. IEEE Computer Society.

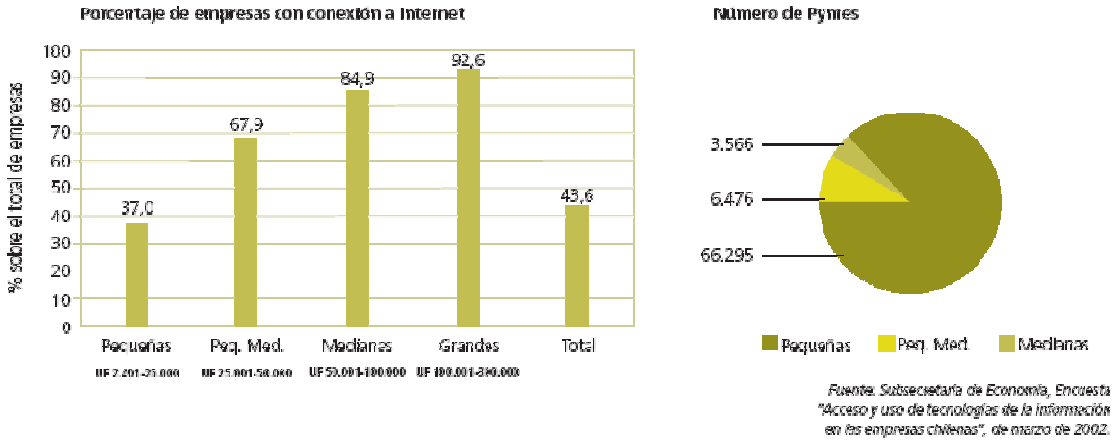
[37] K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Procs. Int. Conf. on Computational Linguistics*, pages 986–989, 1996.

8. ANEXOS

8.1. **Realidad de la Web en Chile.**

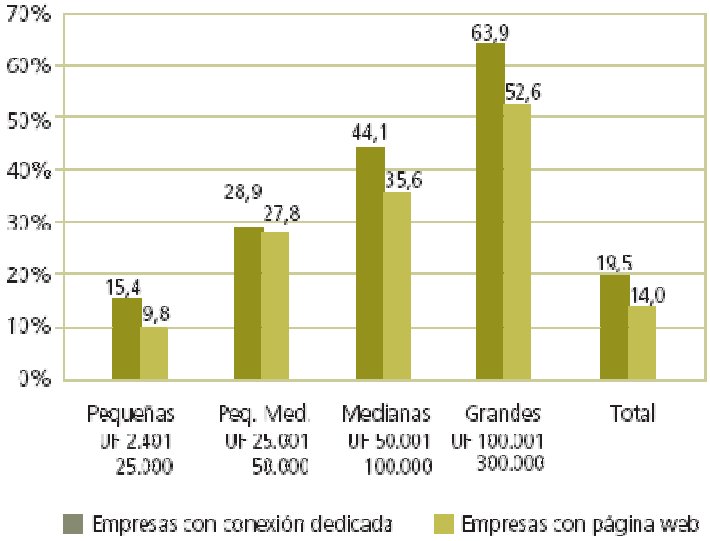
A nivel país, los indicadores de que cada vez es más importante Internet son elocuentes. Se presentan a continuación algunas estadísticas a nivel nacional del uso e importancia de Internet en el país.

Gráfico 1: Aumento de la penetración de Internet en las Empresas.
La presencia de conexión a Internet ha ido en incremento alcanzando niveles importantes en las pequeñas empresas.



Fuente: Subsecretaría de Economía, Encuesta "Acceso y uso de tecnologías de la información en las empresas chilenas", de marzo de 2002.

Gráfico 2: Presencia de página web y conexión dedicada por segmento de empresa.



Fuente: Subsecretaría de Economía, Encuesta "Acceso y uso de tecnologías de la información en las empresas chilenas", de marzo de 2002.

Gráfico 3: Comparativa Internacional: Empresas conectadas y con sitio web.

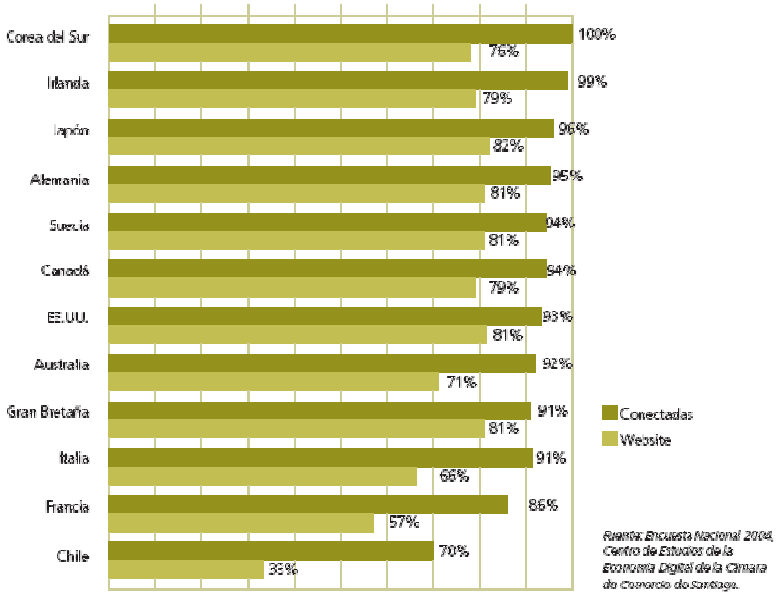


Gráfico 4: Uso de Internet en las empresas.

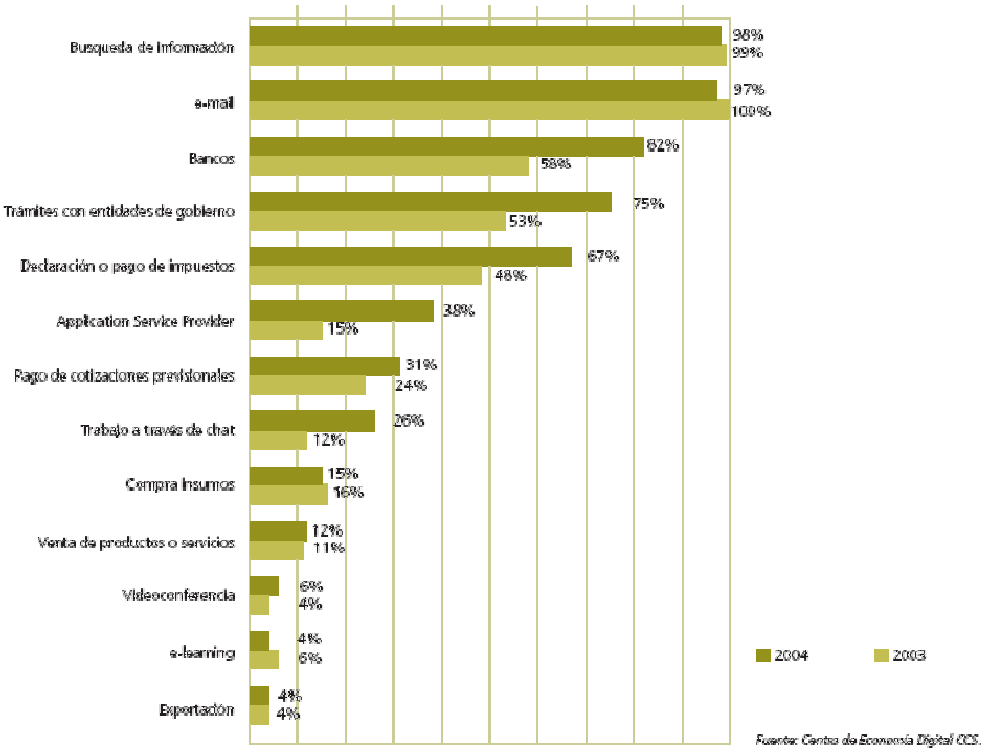
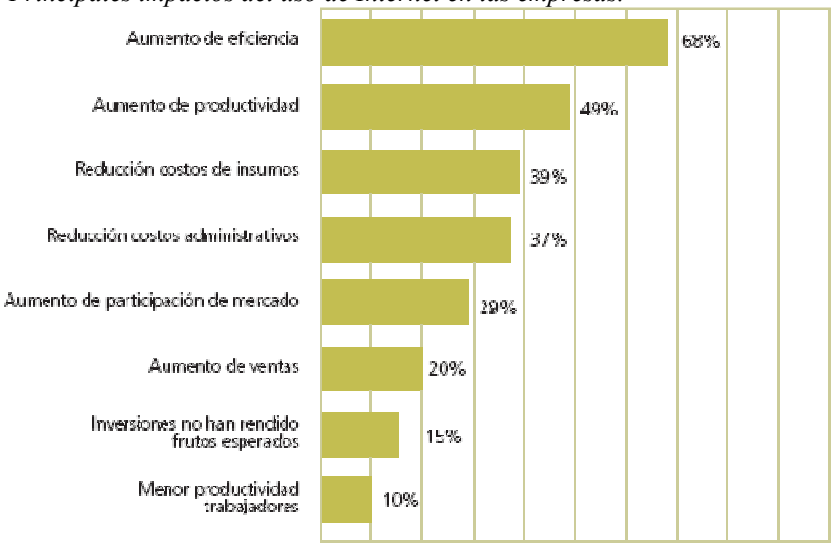


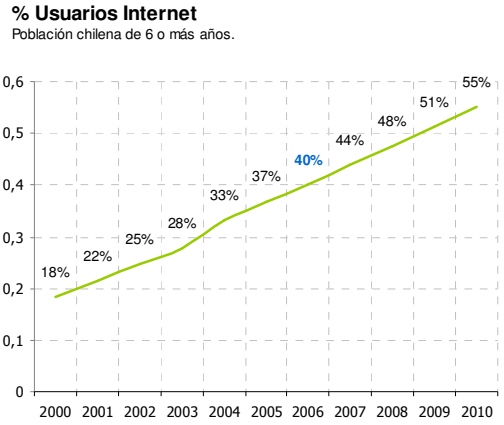
Gráfico 5: Principales impactos del uso de Internet en las empresas.

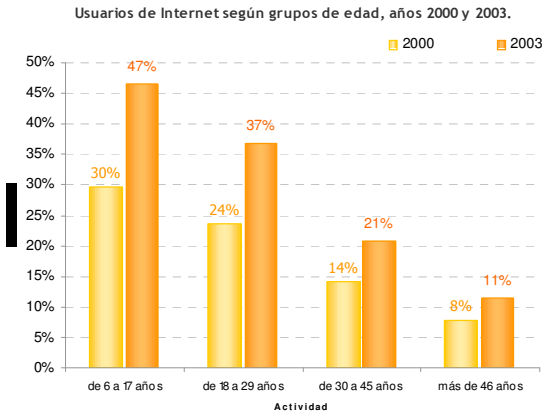


Fuente: Economía digital 2004 - Centro de Estudios de la Economía Digital de la CCS.

8.2. **Realidad de la Web en Chile.**

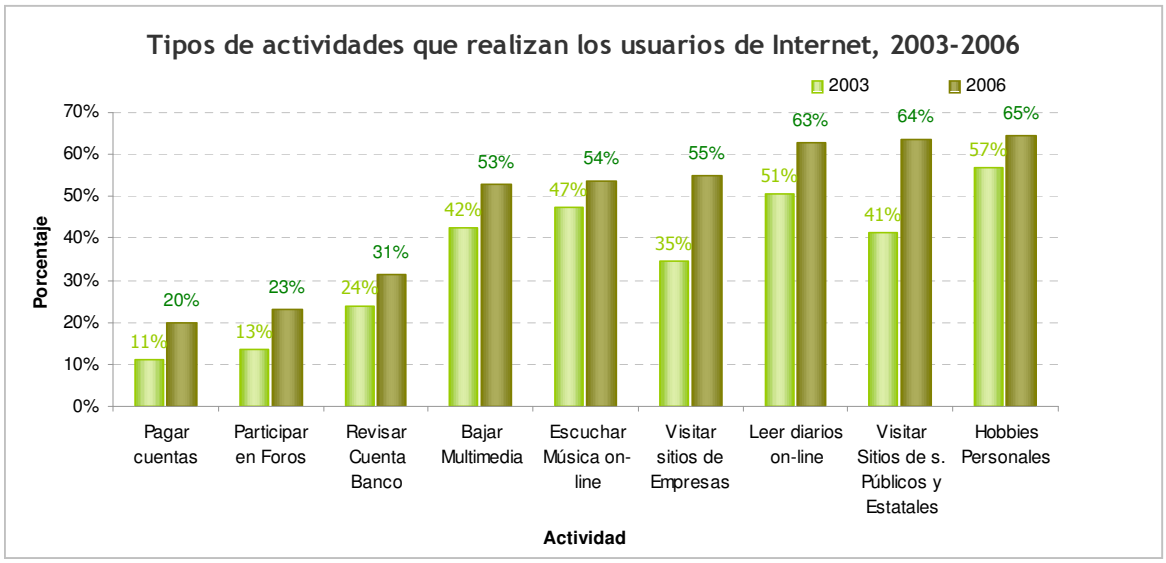
El año 2006 se estima que alrededor de un 40% de la población chilena estaba conectada a la red. La tasa de crecimiento de usuarios de Internet se ha mantenido relativamente constante a partir del año de 2000, aumentando entre 3 y 4 puntos porcentuales por año. Dicho incremento se produce principalmente por la incorporación de población joven a la red, mientras que los usuarios de más edad se han mantenido relativamente más constantes





Actividades que se realizan en Internet

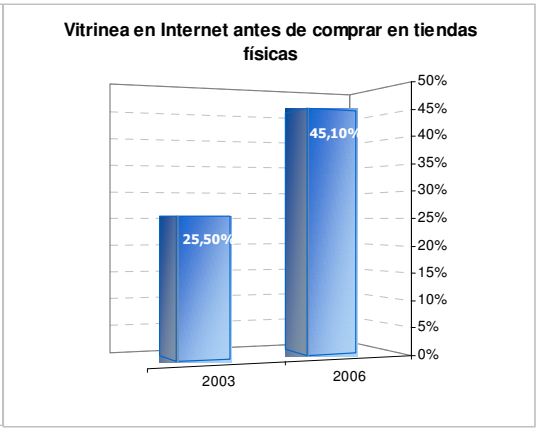
El tipo de actividades que realizan los usuarios no se diferencia mucho por sexo, ni por edad ni por GSE.



Casi todos los tipos de actividades se ven afectados por el lugar de conexión, en particular, los usuarios en sus propios hogares particulares realizan más frecuentemente todos los tipos de actividades

Comportamiento E-Commerce

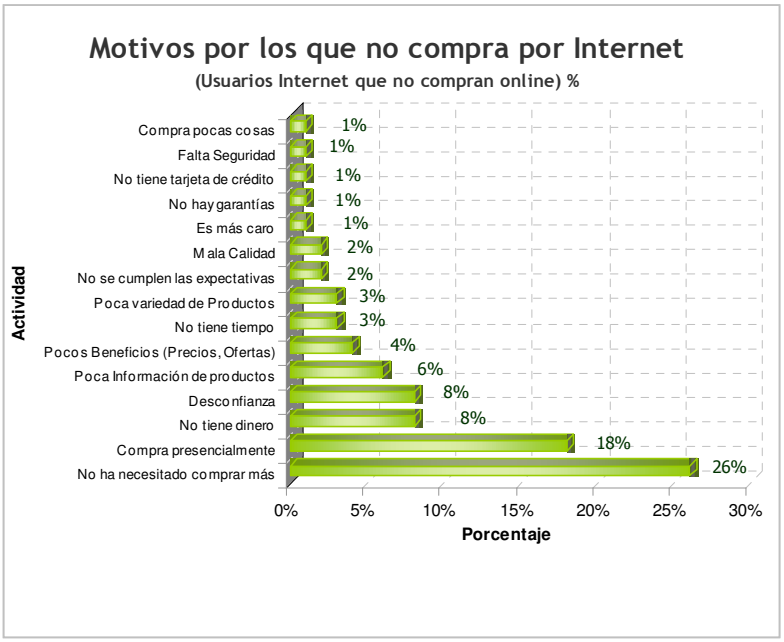
La tasa de compras de los consumidores chilenos a través de Internet ha mostrado una mejoría respecto del período 2003. Un 26.7% de los usuarios mayores a 18 años compra online, 2% más que en 2003, pero varios puntos porcentuales por debajo si nos comparamos con países desarrollados.



Uso de Internet para Servicios

El consumidor valora cada vez más el acceso a precios competitivos en la red, mayores grados de simplificación en los procesos de cotización, de información del producto, de pago y facturación, en general, en toda la cadena que relaciona a la empresa con sus clientes vía web.

Por su parte la principal razón para no comprar online, existe un grupo importante de usuarios que manifiesta cierto nivel de satisfacción respecto de la forma actual como realizan sus compras en los canales tradicionales, por lo tanto desechan por ahora realizar compras por Internet.



Códigos fuente de aplicaciones y procedimientos almacenados.

Los códigos fuente de la aplicación y la estructura de directorios, junto con las procedimientos de creación de tablas y procedimientos almacenados en MySQL se encuentran adjuntos en el disco dentro de la memoria.

8.4. Palabras detectadas de vecindades de centroides según análisis de trabajo.

8.4.1. Palabras de vecindades Aplicación SOFM de 16x16.

8.4.1.1.Cluster 1

Vecino Superior

Palabra	Peso	%	# repeticiones	# términos	Términos asociados
cheques	0,64014418	0,093883	87	1	cheques
caj	0,338761934	0,049682	27	2	caja, cajas
cobr	0,243790465	0,035754	82	8	cobra, cobrado, cobrados, cobrandose, cobrar, cobrara, cobro, cobros
dispon	0,150579838	0,022084	69	3	disponer, disponible, disponibles
line	0,143576613	0,021057	251	2	linea, lineas
banc	0,11079309	0,016249	268	3	banca, banco, bancos
remot	0,075496055	0,011072	66	3	remota, remotamente, remotos
hazte	0,074071816	0,010863	98	1	hazte
ejempl	0,068526202	0,01005	37	2	ejemplo, ejemplos
atencion	0,065719428	0,009638	107	1	atencion

Vecino Inferior

Palabra	Peso	%	# repeticiones	# términos	Términos asociados
minim	0,375509014	0,041113	89	2	minima, minimo
maxim	0,240204401	0,026299	84	2	maxima, maximo
mensual	0,235481982	0,025782	34	1	mensual
caracteristic	0,217009637	0,023759	54	1	caracteristicas
corriente	0,166303473	0,018208	117	1	corriente
sald	0,159944299	0,017512	80	2	saldo, saldos
fech	0,137609642	0,015066	101	2	fecha, fechas
hast	0,105004488	0,011496	124	1	hasta
plaz	0,101236217	0,011084	204	3	plaza, plazo, plazos
cuent	0,100602285	0,011014	345	3	cuenta, cuentan, cuentas

Vecino Izquierdo

Palabra	Peso	%	# repeticiones	# términos	Términos asociados
contactan	0,922294	0,051646	171	1	contactanos
map	0,910709	0,050997	173	1	mapa
home	0,822231	0,046043	188	1	home
hazte	0,803339	0,044985	98	1	hazte
autoconsult	0,741854	0,041542	103	1	autoconsulta
atencion	0,712754	0,039912	107	1	atencion
ejecut	0,561529	0,031444	126	3	ejecutan, ejecutivo, ejecutivos
line	0,538370677	0,030147	251	2	linea, lineas
inversiones	0,392969595	0,022005	284	1	inversiones
fond	0,372805975	0,020876	477	2	fondo, fondos

Vecino Derecho

Palabra	Peso	%	repeticiones	# términos	Términos asociados
ejecut	1,12232	0,056041	126	3	ejecutan, ejecutivo, ejecutivos
mont	0,581833644	0,029053	165	2	monto, montos
line	0,489490927	0,024442	251	2	linea, lineas
fond	0,406820572	0,020314	477	2	fondo, fondos
automat	0,396665904	0,019807	53	4	automatica, automaticamente, automatico, automaticos
cliente	0,361368403	0,018044	186	1	cliente
sobregir	0,361022618	0,018027	38	1	sobregiro
segur	0,358230757	0,017888	112	5	segura, seguras, seguridad, seguro, seguros
oficin	0,346312069	0,017293	9	2	oficina, oficinas
cajer	0,322751171	0,016116	18	2	cajero, cajeros

8.4.1.2.Cluster 2

Vecino Superior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
cuent	1,000397715	0,069764	345	3	cuenta, cuentan, cuentas
person	0,966911464	0,067429	135	2	persona, personas
fond	0,735532123	0,051294	477	2	fondo, fondos
pag	0,491958572	0,034308	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
recomend	0,416436465	0,029041	18	3	recomendable, recomendaciones, recomendamos
dat	0,307883671	0,021471	18	1	datos
numer	0,284098558	0,019812	32	2	numero, numeros
cambi	0,275889404	0,01924	15	4	cambia, cambiar, cambiarla, cambio
estas	0,250664209	0,01748	15	1	estas
automat	0,250014584	0,017435	53	4	automatica, automaticamente, automatico, automaticos

Vecino Inferior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
hazte	0,803339	0,09867	98	1	hazte
line	0,341319	0,041922	251	2	linea, lineas
conveniente	0,287316449	0,03529	8	1	conveniente
banc	0,238706	0,029319	268	3	banca, banco, bancos
cliente	0,227781512	0,027977	186	1	cliente
inversiones	0,173846	0,021353	284	1	inversiones
tas	0,16761434	0,020587	83	2	tasa, tasas
meses	0,13901007	0,017074	50	1	meses
ejecut	0,077312295	0,009496	126	3	ejecutan, ejecutivo, ejecutivos
rescates	0,071556572	0,008789	55	1	rescates

Vecino Izquierdo

Palabra	Peso	%	# repeticiones	# términos	Términos asociados
unic	2,5867	0,18897	20	4	unica, unicamente, unico, unicos
mont	0,508359779	0,037138	165	2	monto, montos
consum	0,455151146	0,033251	24	2	consumas, consumo
igual	0,300577833	0,021959	22	1	igual
conveniente	0,287316449	0,02099	8	1	conveniente
oficin	0,274891971	0,020082	9	2	oficina, oficinas
contactan	0,257426831	0,018806	171	1	contactanos
map	0,255164297	0,018641	173	1	mapa
home	0,236321917	0,017264	188	1	home
equip	0,228003492	0,016657	7	3	equipar, equipo, equipos

Vecino Derecho

Palabra	Peso	%	repeticiones	# términos	Términos asociados
cobr	1,623180457	0,105784	82	8	cobra, cobrado, cobrados, cobrandose, cobrar, cobrara, cobro, cobros
cuent	0,823894207	0,053694	345	3	cuenta, cuentan, cuentas
par	0,407571268	0,026562	277	1	para
fond	0,359964508	0,023459	477	2	fondo, fondos
cheques	0,352820975	0,022994	87	1	cheques
mensual	0,339505376	0,022126	34	1	mensual
contactan	0,257426831	0,016777	171	1	contactanos
map	0,255164297	0,016629	173	1	mapa
pag	0,252561703	0,01646	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
home	0,236321917	0,015401	188	1	home

8.4.1.3.Cluster 3

Vecino Superior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
deposit	2,585989971	0,145307	147	7	depositado, depositados, depositantes, depositar, deposito, depositos, depósitos
hazte	0,803339	0,04514	98	1	hazte
ahorr	0,757744956	0,042578	314	7	ahorra, ahorrado, ahorran, ahorrantes, ahorrar, ahorro, ahorros
autoconsult	0,741854	0,041685	103	1	autoconsulta
atencion	0,712754	0,04005	107	1	atencion
ejecut	0,561529	0,031552	126	3	ejecutan, ejecutivo, ejecutivos
banc	0,453278602	0,02547	268	3	banca, banco, bancos
fond	0,434791911	0,024431	477	2	fondo, fondos
line	0,341319	0,019179	251	2	linea, lineas
institu	0,332898853	0,018706	12	1	instituciones

Vecino Inferior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
fond	4,321437302	0,125507	477	2	fondo, fondos
cuand	1,820836872	0,052882	61	1	cuando
cual	1,820033652	0,052859	50	1	cual
garantiz	1,52783117	0,044373	83	5	garantiza, garantizada, garantizados, garantizan, garantizar
altern	1,50644	0,043751	47	2	alternativa, alternativas
previsional	1,502297705	0,043631	87	1	previsional
acuerd	1,42476	0,041379	51	1	acuerdo
aporte	1,335339458	0,038782	106	1	aporte
hast	1,30350391	0,037857	124	1	hasta
com	1,056734965	0,030691	113	4	coma, comas, comida, como

Vecino Izquierdo

Palabra	Peso	%	repeticiones	# términos	Términos asociados
dispon	1,954011144	0,110386	69	3	disponer, disponible, disponibles
cuent	1,911934264	0,108009	345	3	cuenta, cuentan, cuentas
pag	1,180088012	0,066666	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
line	0,61583959	0,03479	251	2	linea, lineas
fond	0,415908899	0,023496	477	2	fondo, fondos
par	0,343842616	0,019424	277	1	para
sald	0,290313679	0,0164	80	2	saldo, saldos
adicional	0,274211369	0,015491	30	2	adicional, adicionalmente
promedi	0,268770457	0,015183	24	2	promedio, promedios
calcul	0,239716065	0,013542	22	8	calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, calculos

Vecino Derecho

Palabra	Peso	%	repeticiones	# términos	Términos asociados
meses	0,318135976	0,032552	50	1	meses
valor	0,248060919	0,025382	45	2	valor, valoradas
sald	0,201420775	0,02061	80	2	saldo, saldos
necesit	0,173810944	0,017784	15	6	necesita, necesitadas, necesitan, necesitaras, necesitas, necesito
impuest	0,147424579	0,015085	93	2	impuesto, impuestos
fech	0,137609642	0,01408	101	2	fecha, fechas
incl	0,120937066	0,012374	33	3	incluye, incluyen, incluyendo
pes	0,119896759	0,012268	136	2	peso, pesos
product	0,118362688	0,012111	170	2	producto, productos
solicit	0,116476072	0,011918	63	10	solicita, solicitadas, solicitado, solicitandolas, solicitandolo, solicitandolos, solicitante, solicitar, solicitarle, solicitarlo

8.4.1.4.Cluster 4

Vecino Superior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
line	0,341319	0,051867	251	2	linea, lineas
comisiones	0,187598964	0,028508	57	1	comisiones
product	0,1488446	0,022619	170	2	producto, productos
solicit	0,116476072	0,0177	63	10	solicita, solicitadas, solicitado, solicitandolas, solicitandolo, solicitandolos, solicitante, solicitar, solicitarle, solicitarlo
plaz	0,101236217	0,015384	204	3	plaza, plazo, plazos
fond	0,090673327	0,013779	477	2	fondo, fondos
cuent	0,079969156	0,012152	345	3	cuenta, cuentan, cuentas
ejecut	0,077312295	0,011748	126	3	ejecutan, ejecutivo, ejecutivos
inicial	0,076980018	0,011698	58	1	inicial
alfa	0,076022403	0,011552	54	1	alfa

Vecino Inferior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
mont	1,378181814	0,079846	165	2	monto, montos
solicit	1,26323	0,073187	63	10	solicita, solicitadas, solicitado, solicitandolas, solicitandolo, solicitandolos, solicitante, solicitar, solicitarle, solicitarlo
obten	1,24571	0,072172	61	5	obten, obtener, obtenerlas, obtenidas, obteniendo
dispon	1,2317	0,07136	69	3	disponer, disponible, disponibles
esta	0,629198	0,036453	113	1	esta
pag	0,618255312	0,035819	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
simulacion	0,494976145	0,028677	6	1	simulacion
multicredit	0,434305648	0,025162	33	2	multicredito, multicreditos
dispin	0,375861715	0,021776	4	1	dispinible
aqui	0,344163185	0,019939	17	1	aqui

Vecino Izquierdo

Palabra	Peso	%	repeticiones	# términos	Términos asociados
epes	1,221189078	0,053481	18	1	epesos
multitiend	0,958657904	0,041984	9	1	multitienda
contactan	0,922294	0,040391	171	1	contactanos
map	0,910709	0,039884	173	1	mapa
home	0,822231	0,036009	188	1	home
hazte	0,803339	0,035181	98	1	hazte
autoconsult	0,741854	0,032489	103	1	autoconsulta
paris	0,739062127	0,032367	23	1	paris
transform	0,719531887	0,031511	12	3	transforma, transformar, transformaste
atencion	0,712754	0,031214	107	1	atencion

Vecino Derecho

Palabra	Peso	%	repeticiones	# términos	Términos asociados
vis	1,83282139	0,105228	79	1	visa
total	0,446374262	0,025628	52	2	total, totalidad
cartol	0,397818633	0,02284	26	1	cartola
automat	0,396665904	0,022774	53	4	automatica, automaticamente, automatico, automaticos
product	0,324664141	0,01864	170	2	producto, productos
cajer	0,322751171	0,01853	18	2	cajero, cajeros
corriente	0,301736432	0,017324	117	1	corriente
extranjer	0,289950192	0,016647	95	4	extranjera, extranjero, extranjeros, extranjeros;
cuent	0,276552246	0,015878	345	3	cuenta, cuentan, cuentas
carg	0,273802323	0,01572	64	7	carga, cargadas, cargado, cargados, cargar, cargo, cargos

8.4.1.5.Cluster 5

Vecino Superior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
line	0,341319	0,051867	251	2	linea, lineas
comisiones	0,187598964	0,028508	57	1	comisiones
product	0,1488446	0,022619	170	2	producto, productos
solicit	0,116476072	0,0177	63	10	solicita, solicitadas, solicitado, solicitandolas, solicitandolo, solicitandolos, solicitante, solicitar, solicitarle, solicitarlo
plaz	0,101236217	0,015384	204	3	plaza, plazo, plazos
fond	0,090673327	0,013779	477	2	fondo, fondos
cuent	0,079969156	0,012152	345	3	cuenta, cuentan, cuentas
ejecut	0,077312295	0,011748	126	3	ejecutan, ejecutivo, ejecutivos
inicial	0,076980018	0,011698	58	1	inicial
alfa	0,076022403	0,011552	54	1	alfa

Vecino Inferior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
mont	1,378181814	0,079846	165	2	monto, montos
solicit	1,26323	0,073187	63	10	solicita, solicitadas, solicitado, solicitandolas, solicitandolo, solicitandolos, solicitante, solicitar, solicitarle, solicitarlo
obten	1,24571	0,072172	61	5	obten, obtener, obtenerlas, obtenidas, obteniendo
dispon	1,2317	0,07136	69	3	disponer, disponible, disponibles
esta	0,629198	0,036453	113	1	esta
pag	0,618255312	0,035819	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
simulacion	0,494976145	0,028677	6	1	simulacion
multicredit	0,434305648	0,025162	33	2	multicredito, multicreditos
dispin	0,375861715	0,021776	4	1	dispinible
aqui	0,344163185	0,019939	17	1	aqui

Vecino Izquierdo

Palabra	Peso	%	repeticiones	# términos	Términos asociados
epes	1,221189078	0,053481	18	1	epesos
multitiend	0,958657904	0,041984	9	1	multitienda
contactan	0,922294	0,040391	171	1	contactanos
map	0,910709	0,039884	173	1	mapa
home	0,822231	0,036009	188	1	home
hazte	0,803339	0,035181	98	1	hazte
autoconsult	0,741854	0,032489	103	1	autoconsulta
paris	0,739062127	0,032367	23	1	paris
transform	0,719531887	0,031511	12	3	transforma, transformar, transformaste
atencion	0,712754	0,031214	107	1	atencion

Vecino Derecho

Palabra	Peso	%	repeticiones	# términos	Términos asociados
vis	1,83282139	0,105228	79	1	visa
total	0,446374262	0,025628	52	2	total, totalidad
cartol	0,397818633	0,02284	26	1	cartola
automat	0,396665904	0,022774	53	4	automatica, automaticamente, automatico, automaticos
product	0,324664141	0,01864	170	2	producto, productos
cajer	0,322751171	0,01853	18	2	cajero, cajeros
corriente	0,301736432	0,017324	117	1	corriente
extranjer	0,289950192	0,016647	95	4	extranjera, extranjero, extranjeros, extranjeros;
cuent	0,276552246	0,015878	345	3	cuenta, cuentan, cuentas
carg	0,273802323	0,01572	64	7	carga, cargadas, cargado, cargados, cargar, cargo, cargos

8.4.1.6.Cluster 6

Vecino Superior

Palabra	Peso	%	repeticiones	# términos	Términos asociados
banc	0,299467688	0,036468	268	3	banca, banco, bancos
cobr	0,243790465	0,029688	82	8	cobra, cobrado, cobrados, cobrandose, cobrar, cobrara, cobro, cobros
estad	0,24279744	0,029567	76	2	estado, estados
mensual	0,235481982	0,028676	34	1	mensual
caracteristic	0,217009637	0,026427	54	1	caracteristicas
cliente	0,180861	0,022025	186	1	cliente
anual	0,141425802	0,017222	81	1	anual
fech	0,137609642	0,016758	101	2	fecha, fechas
line	0,100115086	0,012192	251	2	linea, lineas
financ	0,098937159	0,012048	81	2	financiera, financieras

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
gan	1,95897	0,127486	37	6	gana, ganado, ganados, ganando, ganar, ganas
podr	1,293056964	0,08415	78	3	podra, podran, podras
pag	0,761531548	0,049559	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
catalog	0,445230768	0,028975	25	1	catalogo
redcompr	0,40679038	0,026473	21	1	redcompra
par	0,39348171	0,025607	277	1	para
canje	0,374119236	0,024347	54	6	canje, canjeados, canjeando, canjear, canjearlos, canjearse
compr	0,330081093	0,021481	98	5	compra, comprando, comprar, compras, compro
valor	0,283934005	0,018478	45	2	valor, valoradas
sol	0,273087905	0,017772	49	4	sol, sola, solido, solo

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
realiz	1,977441364	0,140964	61	10	realiza, realizada, realizadas, realizados, realizan, realizar, realizara, realizaran, realizarla, realizas
benefici	1,2391	0,08833	92	3	beneficiaras, beneficio, beneficios
corriente	0,283999533	0,020245	117	1	corriente
intereses	0,277407455	0,019775	44	1	intereses
cuent	0,272104138	0,019397	345	3	cuenta, cuentan, cuentas
contactan	0,257426831	0,018351	171	1	contactanos
map	0,255164297	0,01819	173	1	mapa
automat	0,250014584	0,017823	53	4	automatica, automaticamente, automatico, automaticos
home	0,236321917	0,016846	188	1	home
gir	0,231211592	0,016482	31	5	girado, girados, girar, giro, giros

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
cliente	0,286874545	0,025308	186	1	cliente
interes	0,265053162	0,023383	36	2	interes, interesadas
map	0,255164297	0,02251	173	1	mapa
automat	0,250014584	0,022056	53	4	automatica, automaticamente, automatico, automaticos
dat	0,244399246	0,021561	18	1	datos
home	0,236321917	0,020848	188	1	home
primer	0,218144128	0,019244	8	2	primera, primero
adem	0,182557617	0,016105	23	1	ademas
pag	0,17514198	0,015451	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
sol	0,172090727	0,015182	49	4	sol, sola, solido, solo

8.4.1.7.Cluster 7

Vecino Superior

Word	Weight	%	twwc	twtc	Términos
estad	1,4930481	0,136204	76	2	estado, estados
cuent	0,822897699	0,075069	345	3	cuenta, cuentan, cuentas
cartol	0,631168343	0,057579	26	1	cartola
multicredit	0,394684381	0,036005	33	2	multicredito, multicreditos
fond	0,372805975	0,034009	477	2	fondo, fondos
abon	0,258364069	0,023569	29	4	abonado, abonandolo, abono, abonos
deud	0,205435105	0,018741	96	2	deuda, deudas
curs	0,195331465	0,017819	11	4	cursa, cursado, cursados, cursar
solicitud	0,17448241	0,015917	35	1	solicitud
histor	0,173810944	0,015856	15	4	historica, historicamente, historico, historicos

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
anual	1,629858663	0,080464	81	1	anual
corriente	1,406380096	0,069431	117	1	corriente
inclu	1,167437846	0,057635	132	4	incluida, incluidas, incluido, incluir
comision	0,97869375	0,048317	202	1	comision
cuent	0,962247189	0,047505	345	3	cuenta, cuentan, cuentas
desde	0,885447314	0,043714	106	1	desde
esta	0,792740645	0,039137	113	1	esta
premium	0,639395191	0,031566	35	1	premium
intereses	0,440012439	0,021723	44	1	intereses
simul	0,398225306	0,01966	39	3	simula, simulador, simuladores

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
contactan	0,922294	0,045139	171	1	contactanos
map	0,910709	0,044572	173	1	mapa
home	0,822231	0,040241	188	1	home
hazte	0,803339	0,039317	98	1	hazte
autoconsult	0,741854	0,036308	103	1	autoconsulta
atencion	0,712754	0,034883	107	1	atencion
ejecut	0,561529	0,027482	126	3	ejecutan, ejecutivo, ejecutivos
inclu	0,528866448	0,025884	132	4	incluida, incluidas, incluido, incluir
line	0,428218716	0,020958	251	2	linea, lineas
inversion	0,380441477	0,018619	79	2	inversion, inversionistas

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
tod	3,454599768	0,134417	93	4	toda, todas, todo, todos
esta	1,814923334	0,070618	113	1	esta
par	1,255601113	0,048855	277	1	para
mayor	1,09391	0,042564	71	1	mayor
segur	0,611884809	0,023808	112	5	segura, seguras, seguridad, seguro, seguros
fond	0,584383561	0,022738	477	2	fondo, fondos
sistem	0,423358657	0,016473	31	2	sistema, sistemas
form	0,348851089	0,013574	35	3	forma, formar, formas
person	0,347161054	0,013508	135	2	persona, personas
larg	0,313247375	0,012188	48	3	larga, largo, largos

8.4.2. Palabras de vecindades Aplicación SOFM de 32x32

8.4.2.1.Cluster 1

Vecino Superior

Word	Weight	%	twwc	twtc	Términos
plan	2,467211599	0,223177	61	2	plan, plana
cuent	0,920406006	0,083257	345	3	cuenta, cuentan, cuentas
requ	0,343158005	0,031041	37	7	requerida, requerido, requerimiento, requerimientos, requerimos, requiera, requieras
previsional	0,304496058	0,027544	87	1	previsional
estad	0,267148004	0,024165	76	2	estado, estados
maxim	0,240204401	0,021728	84	2	maxima, maximo
period	0,23844382	0,021569	19	2	periodicamente, periodo
minim	0,236971753	0,021436	89	2	minima, minimo
inform	0,195331465	0,017669	11	6	información, informadas, informado, informados, informar, informara
fond	0,173984562	0,015738	477	2	fondo, fondos

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
meses	3,450087319	0,23634	50	1	meses
pag	0,761531548	0,052167	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
simul	0,398225306	0,027279	39	3	simula, simulador, simuladores
numer	0,357863964	0,024515	32	2	numero, numeros
benefici	0,353559655	0,02422	92	3	beneficiaras, beneficio, beneficios
caracteristic	0,312800407	0,021428	54	1	caracteristicas
impuest	0,267732764	0,01834	93	2	impuesto, impuestos
calcul	0,239716065	0,016421	22	8	calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, calculos
cuent	0,219611436	0,015044	345	3	cuenta, cuentan, cuentas
liqu	0,219026199	0,015004	16	3	liquida, liquidar, liquido

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
desde	0,885447314	0,082346	106	1	desde
caracteristic	0,312800407	0,02909	54	1	caracteristicas
net	0,310096438	0,028839	12	1	neto
inclu	0,21039355	0,019566	132	4	incluida, incluidas, incluido, incluir
person	0,203528474	0,018928	135	2	persona, personas
anual	0,178146884	0,016567	81	1	anual
corriente	0,166303473	0,015466	117	1	corriente
tabl	0,157599409	0,014657	32	1	tabla
cuent	0,152578037	0,01419	345	3	cuenta, cuentan, cuentas
comision	0,142338515	0,013237	202	1	comision

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
obten	1,24571	0,101713	61	5	obten, obtener, obtenerlas, obtenidas, obteniendo
excedente	0,41439978	0,033836	9	1	excedente
igual	0,378605542	0,030913	22	1	igual
libre	0,333823416	0,027257	15	1	libre
pension	0,240033576	0,019599	67	4	pension, pensionado, pensionandose, pensionarse
capitalizacion	0,233424652	0,019059	66	1	capitalizacion
promedi	0,213351226	0,01742	24	2	promedio, promedios
rent	0,199183074	0,016263	65	2	renta, rentas
individual	0,198770669	0,01623	20	1	individual
ultim	0,17558439	0,014337	26	5	ultima, ultimas, últimas, ultimo, ultimos

8.4.2.2.Cluster 2

Vecino Superior

Word	Weight	%	twwc	twtc	Términos
abon	0,325433193	0,042431	29	4	abonado, abonandolo, abono, abonos
sobregir	0,315418925	0,041125	38	1	sobregiro
nombre	0,257156026	0,033529	91	1	nombre
cliente	0,227781512	0,029699	186	1	cliente
ahor	0,218144128	0,028442	8	1	ahora
sald	0,201420775	0,026262	80	2	saldo, saldos
line	0,198118776	0,025831	251	2	linea, lineas
cuent	0,187508072	0,024448	345	3	cuenta, cuentan, cuentas
fech	0,173271907	0,022592	101	2	fecha, fechas
person	0,161792589	0,021095	135	2	persona, personas

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
pes	2,042693299	0,176552	136	2	peso, pesos
par	0,739368046	0,063904	277	1	para
onris	0,412756178	0,035675	93	2	onrisa, onrisas
gan	0,345265632	0,029842	37	6	gana, ganado, ganados, ganando, ganar, ganas
person	0,29323308	0,025344	135	2	persona, personas
hac	0,276919395	0,023934	35	4	hacer, hacerlo, hacerse, haciendo
vis	0,268053118	0,023168	79	1	visa
mejor	0,190083048	0,016429	28	5	mejor, mejora, mejoran, mejorar, mejoraran
informacion	0,186010778	0,016077	60	1	informacion
com	0,13509355	0,011676	113	4	coma, comas, comida, como

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
impuest	0,267732764	0,02992	93	2	impuesto, impuestos
meses	0,25252117	0,02822	50	1	meses
benefici	0,245570306	0,027443	92	3	beneficiaras, beneficio, beneficios
gener	0,234864001	0,026247	25	3	generales, generan, generar
liqu	0,219026199	0,024477	16	3	liquida, liquidar, liquido
ingres	0,20415	0,022814	47	8	ingresa, ingresada, ingresan, ingresando, ingresar, ingresas, ingreso, ingresos
pag	0,188617914	0,021079	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
segur	0,161178275	0,018012	112	5	segura, seguras, seguridad, seguro, seguros
cuent	0,115102681	0,012863	345	3	cuenta, cuentan, cuentas
mont	0,114601755	0,012807	165	2	monto, montos

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
contacten	3,1668	0,133752	18	1	contactenos
onris	1,971346771	0,083261	93	2	onrisa, onrisas
pes	1,669595474	0,070517	136	2	peso, pesos
home	1,031830452	0,04358	188	1	home
com	0,838877668	0,035431	113	4	coma, comas, comida, como
hazte	0,803339	0,03393	98	1	hazte
pag	0,744633623	0,03145	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
autoconsult	0,741854	0,031333	103	1	autoconsulta
atencion	0,712754	0,030104	107	1	atencion
ejecut	0,561529	0,023717	126	3	ejecutan, ejecutivo, ejecutivos

8.4.2.3.Cluster 3

Vecino Superior

Word	Weight	%	twwc	twtc	Términos
benefici	2,244769385	0,201156	92	3	beneficiaras, beneficio, beneficios
tod	1,378746666	0,123551	93	4	toda, todas, todo, todos
calcul	0,239716065	0,021481	22	8	calcula, calculadora, calculadoras, calculan, calcular, calcularan, calculo, calculos
par	0,238697304	0,02139	277	1	para
meses	0,220607158	0,019769	50	1	meses
cobr	0,2130108	0,019088	82	8	cobra, cobrado, cobrados, cobrandose, cobrar, cobrara, cobro, cobros
nuestr	0,198545558	0,017792	32	4	nuestra, nuestras, nuestro, nuestros
adicional	0,190160744	0,017041	30	2	adicional, adicionalmente
sol	0,172090727	0,015421	49	4	sol, sola, solido, solo
voluntari	0,124569439	0,011163	144	5	voluntaria, voluntariamente, voluntarias, voluntario, voluntarios

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
benefici	0,245570306	0,040899	92	3	beneficiaras, beneficio, beneficios
dispon	0,150579838	0,025079	69	3	disponer, disponible, disponibles
cuent	0,115102681	0,01917	345	3	cuenta, cuentan, cuentas
mont	0,094616945	0,015758	165	2	monto, montos
remot	0,05998892	0,009991	66	3	remota, remotamente, remotos
cartol	0,05084404	0,008468	26	1	cartola
integral	0,049908625	0,008312	40	1	integral
sobregir	0,046141229	0,007685	38	1	sobregiro
plan	0,040301058	0,006712	61	2	plan, plana
deud	0,038329957	0,006384	96	2	deuda, deudas

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
mont	0,403678402	0,045623	165	2	monto, montos
impuest	0,267717405	0,030257	93	2	impuesto, impuestos
tas	0,227440045	0,025705	83	2	tasa, tasas
pag	0,214860098	0,024283	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
corriente	0,190317575	0,021509	117	1	corriente
informacion	0,186010778	0,021023	60	1	informacion
anual	0,141425802	0,015984	81	1	anual
cuent	0,115102681	0,013009	345	3	cuenta, cuentan, cuentas
contactan	0,071851897	0,008121	171	1	contactanos
map	0,071492451	0,00808	173	1	mapa

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
mastercard	0,293334121	0,026747	20	1	mastercard
pes	0,27412332	0,024995	136	2	peso, pesos
par	0,273157017	0,024907	277	1	para
map	0,255164297	0,023266	173	1	mapa
benefici	0,245570306	0,022391	92	3	beneficiaras, beneficio, beneficios
dispon	0,238921544	0,021785	69	3	disponer, disponible, disponibles
home	0,236321917	0,021548	188	1	home
informacion	0,212913677	0,019414	60	1	informacion
vis	0,212809406	0,019404	79	1	visa
line	0,170034447	0,015504	251	2	linea, lineas

8.4.2.4.Cluster 4

Vecino Superior

Word	Weight	%	twwc	twtc	Términos
minim	0,508988093	0,077808	89	2	minima, minimo
maxim	0,410031252	0,062681	84	2	maxima, maximo
compr	0,350950414	0,053649	98	5	compra, comprando, comprar, compras, compro
product	0,1488446	0,022754	170	2	producto, productos
extranjer	0,077062247	0,01178	95	4	extranjera, extranjero, extranjeros, extranjeros;
hipotecari	0,072695286	0,011113	29	2	hipotecario, hipotecarios
cup	0,065532811	0,010018	25	2	cupo, cupos
mont	0,057355881	0,008768	165	2	monto, montos
vis	0,0565309	0,008642	79	1	visa
comision	0,052600591	0,008041	202	1	comision

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
comisiones	2,115994398	0,07428	57	1	comisiones
anual	1,629858663	0,057215	81	1	anual
relacion	1,48616	0,05217	50	4	relacion, relacionada, relacionadas, relacionados
inclu	1,335801253	0,046892	132	4	incluida, incluidas, incluido, incluir
comision	0,97869375	0,034356	202	1	comision
desde	0,885447314	0,031083	106	1	desde
hazte	0,803339	0,0282	98	1	hazte
fond	0,788074376	0,027665	477	2	fondo, fondos
autoconsult	0,741854	0,026042	103	1	autoconsulta
atencion	0,712754	0,025021	107	1	atencion

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
fond	0,331102328	0,036418	477	2	fondo, fondos
remuneracion	0,19427447	0,021369	126	1	remuneracion
person	0,161792589	0,017796	135	2	persona, personas
ahorr	0,153518667	0,016886	314	7	ahorra, ahorrado, ahorran, ahorrantes, ahorrar, ahorro, ahorros
cuent	0,152578037	0,016782	345	3	cuenta, cuentan, cuentas
voluntari	0,124569439	0,013702	144	5	voluntaria, voluntariamente, voluntarias, voluntario, voluntarios
pag	0,121695661	0,013385	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
banc	0,11079309	0,012186	268	3	banca, banco, bancos
mont	0,104131014	0,011454	165	2	monto, montos
line	0,100115086	0,011012	251	2	linea, lineas

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
deud	1,99569775	0,126765	96	2	deuda, deudas
cuent	0,764026256	0,04853	345	3	cuenta, cuentan, cuentas
par	0,433016685	0,027505	277	1	para
fond	0,314564314	0,019981	477	2	fondo, fondos
cort	0,266821144	0,016948	31	2	corta, corto
contactan	0,257426831	0,016352	171	1	contactanos
benefici	0,245570306	0,015598	92	3	beneficiaras, beneficio, beneficios
home	0,236321917	0,015011	188	1	home
sald	0,230532417	0,014643	80	2	saldo, saldos
cliente	0,227781512	0,014468	186	1	cliente

8.4.2.5.Cluster 5

Vecino Superior

Word	Weight	%	twwc	twtc	Términos
fond	0,709470569	0,074664	477	2	fondo, fondos
desde	0,70281	0,073963	106	1	desde
line	0,61583959	0,06481	251	2	linea, lineas
banc	0,299467688	0,031516	268	3	banca, banco, bancos
serie	0,259372061	0,027296	86	1	serie
pag	0,232285251	0,024445	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
cuent	0,20897838	0,021993	345	3	cuenta, cuentan, cuentas
cheques	0,206632304	0,021746	87	1	cheques
cliente	0,180861	0,019034	186	1	cliente
pas	0,117732865	0,01239	80	4	pasadas, pasado, paso, pasos

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
par	0,238697304	0,041113	277	1	para
tas	0,16761434	0,02887	83	2	tasa, tasas
retir	0,062995468	0,01085	53	2	retiro, retiros
mont	0,057355881	0,009879	165	2	monto, montos
bloque	0,052931252	0,009117	20	4	bloqueada, bloqueara, bloqueo, bloqueos
total	0,044331502	0,007636	52	2	total, totalidad
consum	0,041967175	0,007228	24	2	consumas, consumo
consult	0,039508178	0,006805	41	3	consulta, consultar, consultas
igual	0,034909308	0,006013	22	1	igual
estad	0,034497421	0,005942	76	2	estado, estados

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
deud	1,99569775	0,135659	96	2	deuda, deudas
cuent	0,764026256	0,051935	345	3	cuenta, cuentan, cuentas
par	0,433016685	0,029435	277	1	para
fond	0,314564314	0,021383	477	2	fondo, fondos
cort	0,266821144	0,018137	31	2	corta, corto
contactan	0,257426831	0,017499	171	1	contactanos
benefici	0,245570306	0,016693	92	3	beneficiaras, beneficio, beneficios
home	0,236321917	0,016064	188	1	home
sald	0,230532417	0,015671	80	2	saldo, saldos
cliente	0,227781512	0,015484	186	1	cliente

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
utiliz	0,109945875	0,012797	39	7	utiliza, utilizado, utilizados, utilizando, utilizar, utilizarla, utilizarlos
podr	0,08268126	0,009624	78	3	podra, podran, podras
pag	0,077054223	0,008969	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
contactan	0,071851897	0,008363	171	1	contactanos
map	0,071492451	0,008321	173	1	mapa
home	0,067922577	0,007906	188	1	home
integral	0,062842689	0,007315	40	1	integral
duracion	0,060118189	0,006997	37	1	duracion
remot	0,05998892	0,006982	66	3	remota, remotamente, remotos
line	0,05298412	0,006167	251	2	linea, lineas

8.4.2.6.Cluster 6

Vecino Superior

Word	Weight	%	twwc	twtc	Términos
corriente	0,166303473	0,027194	117	1	corriente
cuent	0,100602285	0,01645	345	3	cuenta, cuentan, cuentas
line	0,100115086	0,016371	251	2	linea, lineas
deposit	0,089252961	0,014594	147	7	depositado, depositados, depositantes, depositar, deposito, depositos, depósitos
banc	0,088126536	0,01441	268	3	banca, banco, bancos
fond	0,086735385	0,014183	477	2	fondo, fondos
inicial	0,076980018	0,012588	58	1	inicial
alfa	0,076022403	0,012431	54	1	alfa
simul	0,075335593	0,012319	39	3	simula, simulador, simuladores
hazte	0,074071816	0,012112	98	1	hazte

Vecino Inferior

Word	Weight	%	twwc	twtc	Términos
carg	3,268856248	0,162333	64	7	carga, cargadas, cargado, cargados, cargar, cargo, cargos
cuent	3,098940393	0,153895	345	3	cuenta, cuentan, cuentas
line	1,04593572	0,051942	251	2	linea, lineas
corriente	0,775453	0,038509	117	1	corriente
banc	0,508577594	0,025256	268	3	banca, banco, bancos
pag	0,482074366	0,02394	347	13	paga, pagado, pagados, pagando, pagar, pagaran, pagaras, pagarla, pagarlo, pagas, pago, pagos, pagues
cartol	0,455335668	0,022612	26	1	cartola
remot	0,399411488	0,019835	66	3	remota, remotamente, remotos
automat	0,339251049	0,016847	53	4	automatica, automaticamente, automatico, automaticos
fond	0,331102328	0,016443	477	2	fondo, fondos

Vecino Izquierdo

Word	Weight	%	twwc	twtc	Términos
cheques	0,260103132	0,032579	87	1	cheques
corriente	0,166303473	0,02083	117	1	corriente
product	0,1488446	0,018644	170	2	producto, productos
cuent	0,126559042	0,015852	345	3	cuenta, cuentan, cuentas
deposit	0,112354156	0,014073	147	7	depositado, depositados, depositantes, depositar, deposito, depositos, depósitos
map	0,102603559	0,012852	173	1	mapa
tod	0,088168083	0,011044	93	4	toda, todas, todo, todos
simul	0,075335593	0,009436	39	3	simula, simulador, simuladores
contactan	0,071851897	0,009	171	1	contactanos
home	0,067922577	0,008508	188	1	home

Vecino Derecho

Word	Weight	%	twwc	twtc	Términos
banc	0,299467688	0,02648	268	3	banca, banco, bancos
larg	0,248743539	0,021994	48	3	larga, largo, largos
home	0,236321917	0,020896	188	1	home
comision	0,225086595	0,019903	202	1	comision
nombre	0,20423816	0,018059	91	1	nombre
carg	0,189910269	0,016792	64	7	carga, cargadas, cargado, cargados, cargar, cargo, cargos
otra	0,189235228	0,016733	12	1	otra
cliente	0,180861	0,015992	186	1	cliente
fech	0,137609642	0,012168	101	2	fecha, fechas
corriente	0,132104159	0,011681	117	1	corriente

Una metodología para mejorar el contenido de un sitio web a partir de la identificación de sus web site keywords

José I. Fernández*
Juan D. Velásquez*

Resumen

Presentamos una metodología para identificar aproximadamente qué palabras atraen la atención del usuario cuando se encuentra visitando páginas de un sitio web. Estas palabras son llamadas "web site keywords" y pueden ser usadas para la creación de contenidos relacionados a un tópico específico con el que se pretende atraer la atención del usuario.

A través de la utilización de las palabras correctas, se puede mejorar gradualmente el contenido de un sitio web, ayudando de esta forma a los usuarios a encontrar la información que buscan, lo cual se considera clave para el éxito y continuidad del sitio.

Aplicando algoritmos de clustering, y asumiendo que existe una correlación entre el tiempo invertido en una página y el interés del usuario, se realiza una segmentación de los usuarios según comportamiento de navegación y preferencias de contenidos. A continuación, se identifican las palabras clave del sitio web. Esta metodología fue aplicada en datos originada desde un sitio web real, mostrando su efectividad.

Palabras Clave: Web site keywords, Clustering, Comportamiento del usuario web.

*Departamento de Ingeniería Industrial, Universidad de Chile

1. Introducción

Para muchas compañías y/o instituciones, ya no es suficiente el desarrollo de un sitio web para ofrecer sus productos y servicios en el mercado digital. Lo que a menudo hace la diferencia entre un éxito o fracaso en un e-business es el potencial del sitio web para atraer o retener usuarios. Este potencial depende del contenido del sitio, diseño y aspectos técnicos como tiempo de descarga de páginas del sitio hacia navegador del usuario, entre otros. En términos de contenido, las palabras usadas en el texto libre en las páginas de un sitio web son muy importantes, por cuanto dicen relación con la información que los usuarios buscan. En efecto, la gran mayoría de lo usuarios recurre a motores de búsqueda, tales como Yahoo! y Google, para realizar consultas respecto de un contenido de su interés, a través de consultas basadas en términos en motores de búsqueda para encontrar información en la Web. Estas consultas son realizadas usando palabras clave, es decir, una palabra o grupo de palabras [14] que caracterizan el contenido de un página web dada o un sitio web.

El correcto uso de las palabras con que se crea el contenido textual de una página web, mejora la información presentada a los usuarios, ayuda a la búsqueda efectiva de información, mientras atrae a nuevos usuarios y retiene a los actuales, mediante actualizaciones continuas del contenido textual de la página. El desafío, entonces, es identificar qué palabras son importantes para los usuarios. Lo anterior tiende a relacionarse con cual es "palabra más frecuentemente usada". Algunas herramientas comerciales[1] ayudan a identificar palabras clave objetivo que los consumidores son más propensos a utilizar mientras realizan sus búsquedas en la Web [6].

Mediante la identificación de las palabras más relevantes en las páginas de los sitios, desde el punto de vista del usuario, las mejoras pueden ser realizadas en el sitio web completo. Por ejemplo, el sitio puede ser reestructurado colocando un nuevo hyperlink relacionado con la palabra clave y por supuesto el contenido textual podría ser modificado utilizando las palabras clave relacionadas con un tópico específico para enriquecer el texto libre en una página Web.

En este trabajo se presenta una metodología para analizar el comportamiento de navegación del usuario y sus preferencias de contenido a través de la aplicación de algoritmos de web mining en datos originados en la web, también llamados web data, específicamente registros de un sitio web (web logs) y su contenido textual.

La metodología apunta a identificar aproximadamente cuales palabras atraen

[1] Ver por ejemplo http://www.goodkeywords.com

la atención del usuario cuando esta visitando páginas en un sitio web. Estas palabras son denominadas "palabras clave de un sitio web" [31] y pueden ser utilizadas para la creación de contenidos de texto mejoradas relacionadas con tópicos específicos.

Este paper esta organizado de la siguiente forma: La sección 2 introduce una revisión breve acerca del trabajo relacionado. El proceso de preparación para transformar la web data en vectores de características para ser utilizados como entrada en los algoritmos de web mining es mostrado en la sección 3. En la sección 4, la metodología para identificar las palabras clave de un sitio web es explicada y aplicada en la sección 5. Finalmente, la sección 6 muestra las conclusiones principales de este paper.

2. Trabajos Previos

Cuando un usuario visita un sitio web, datos respecto de qué página visitó son almacenados en archivos de registro llamados web logs. Entonces es directo conocer cuáles páginas son visitadas y cuáles no, e inclusive el tiempo gastado por el usuario en cada una de ellas. Debido a que usualmente las páginas contienen datos acerca de un tópico específico, es posible conocer aproximadamente las preferencias de información de los usuarios. En ese sentido la interacción entre el usuario y el sitio es como una indagación electrónica, entregando los datos necesarios para analizar las preferencias de contenido del usuario en un sitio web particular.

El desafío para analizar las preferencias de texto del usuario en el sitio web es doble. Primero la cantidad de registros en el archivo web log usualmente es enorme, y una parte son datos irrelevantes acerca del comportamiento de navegación del usuario en el sitio. Segundo, el texto libre dentro de las páginas web es comúnmente plano, es decir, sin información adicional que permita conocer directamente cuáles son las palabras que atraen la atención del usuario.

En esta sección se revisan las principales aproximaciones para analizar las web data para extraer patrones significativos relacionados con las preferencias de texto de los usuarios en el sitio web.

2.1. Minando los web data

Las técnicas de web mining emergieron como resultado de la aplicación de teoría de data mining al descubrimiento de patrones desde los web data [8, 16, 25]. El web mining no es una tarea trivial considerando que la Web es una enorme colección de información heterogénea, no clasificada, distribuida, variante en el tiempo, semi estructurada y altamente dimensional. El web mining debe considerar tres importantes pasos: Preprocesamiento, descubrimiento de

patrones y análisis de patrones [27].

Las siguientes terminologías comunes son utilizadas para definir los diferentes tipos de web data.

- Contenido. El contenido de la página web, es decir, imágenes, texto libre, sonidos, etc.
- Estructura. Información que muestra la estructura interna de una página web. En general, tienen etiquetas HTML o XML, alguna de las cuales contienen información acerca de hipervínculos con otras páginas web.
- Uso. Información que describe las preferencias del visitante mientras navega en un sitio web. Es posible encontrar esta información dentro de los archivos web log.
- Perfil del usuario. Colección de información acerca del usuario: Información personal (nombre, edad, etc.), información de uso (por ejemplo, páginas visitadas) e intereses.

Con las definiciones anteriores, y dependiendo de los web data a procesar, las técnicas de web mining pueden ser agrupadas en tres áreas: Minado de contenido web (WCM o Web Content Mining), Minado de la estructura web (WSM o Web Structure Mining), y Minado de la utilización de la web (WUM o Web Usage Mining).

2.1.1. Identificando palabras para la creación de un resumen automático de texto de una página web

La meta es construir automáticamente resúmenes de lenguaje natural de documento [11]. En este caso, una semi estructura relativa es creada por la aplicación de etiquetas HTML desde el contenido textual de una página web, la cual examina temas sin restricción de dominio. En muchos casos, las páginas pueden solamente contener pocas palabras sin elementos textuales (por ejemplo video, imágenes, audio, etc.) [1].

En la investigación de resumen de texto, tres importantes aproximaciones son [18]: basadas en parágrafos, basadas en oraciones y utilización de señales de lenguaje natural en texto.

La primera aproximación consiste en seleccionar un párrafo de un segmento de texto [19] que apunta a un tema en el documento, bajo la suposición que hay varios temas en el texto. La aplicación de esta técnica en una página web no es obvia; los diseñadores web tienen la tendencia de estructurar el texto en párrafos por página. Por lo tanto un documento contiene un solo tema, lo cual hace la aplicación de esta técnica dificultosa.

En la segunda aproximación, las frases más interesantes o frases clave son extraídas y ensambladas en un texto individual [9,37]. Es claro que el texto

resultante puede no ser cohesivo, pero la meta de la técnica es proveer la máxima expresión de información en el documento. Esta técnica es aplicable para páginas web, dado que la entrada puede consistir de pequeñas piezas de texto [6]. La aproximación final es un modelo de discurso basado en la extracción y resumen [14,15] mediante la utilización de señales de lenguaje natural como identificación de nombres propios, sinónimos, frases claves, etc. Este método arma oraciones mediante la creación de una colección de texto con información del documento completo. Esta técnica es más apropiada para documentos dentro de un dominio específico y esto para la implementación en un sitio web es dificultoso.

2.2. Extracción de texto de páginas web y aplicaciones

Las componentes de texto clave son partes de un documento completo, por ejemplo un párrafo, frase y una palabra que contiene información significativa acerca de un tema particular, desde el punto de vista del usuario del sitio web. La identificación de estos componentes puede ser útil para mejorar el contenido textual de un sitio web.

Usualmente, las palabras clave en un sitio web están correlacionadas con las "palabras más frecuentemente utilizadas". En [6], se introduce un método para la extracción de las palabras clave desde un gran conjunto de páginas web. La técnica está basada en la asignación de importancia a las palabras, dependiendo de su frecuencia en todos los documentos. Seguidamente, los párrafos o frases que contienen las palabras clave son extraídos y su importancia es validada a través de pruebas con usuarios reales.

Otro método, en [2], recolecta palabras clave desde un motor de búsqueda. Esto muestra las preferencias globales de palabras de una comunidad web, pero no brinda detalles acerca de un sitio web particular.

Finalmente, en lugar de analizar palabras, en [17] se desarrolla una técnica para extraer conceptos desde el texto de una página web. Los conceptos describen objetos del mundo real, eventos, pensamientos, opiniones e ideas en una estructura simple, como términos descriptivos. Entonces, utilizando el modelo de vector espacial, los conceptos son transformados en vectores de características, permitiendo la aplicación de algoritmos de clustering o clasificación a páginas web.

3. Proceso de preparación de la Web Data

De toda la información web disponible, la más relevante para el análisis del comportamiento y preferencias de navegación del usuario, son los registros (web logs) y las paginas web [33]. Los web logs contienen información acerca

de la secuencia de navegación de páginas y el tiempo gastado en cada página visitada, aplicando el proceso de sesionización. La fuente de la página web es el sitio web en si mismo. Cada página web es definida por su contenido, en particular texto libre. Para estudiar el comportamiento del usuario ambas fuentes - web logs y páginas web - se preparan mediante la utilización de filtros y por la estimación de sesiones reales de usuario. La etapa de preprocesamiento implica, primero, un proceso de limpieza y, segundo, la creación de vectores de características como entrada a los algoritmos de web mining, dentro de la estructura definida por los patrones vistos.

3.1. El proceso de reconstrucción de sesiones

El proceso de segmentación de las actividades de usuarios en sesiones individuales es llamado ***sesionización*** [10] y está basado en los web logs del sitio web. En consideración de los inconvenientes mencionados anteriormente, el proceso no esta libre de errores [26]. La sesionización asume que la sesión tiene un tiempo de duración máximo y que no es posible saber si el visitante ha presionado el botón "volver" (back) en el navegador del sitio web. Si la página esta en el cache del navegador y el visitante vuelve a ella en la misma sesión, podría no quedar registrada en los logs del sitio web. Por esto han sido propuestos el uso de esquemas invasivos como el envío de otra aplicación al browser para capturar el comportamiento de navegación exacto del usuario [3, 10]. Si embargo, este esquema podría ser fácilmente evitado por el visitante.

Muchos autores [3, 10, 20] han propuesto la utilización de heurísticas para la reconstrucción de sesiones por los web logs. En esencia, la idea es crear subgrupos con las visitas de usuarios y aplicando mecanismos sobre los web logs generados para permitir la definición de una sesión como series de eventos entrelazados durante un cierto periodo de tiempo.

La reconstrucción de sesiones apunta a encontrar sesiones de usuarios reales, es decir, cuales páginas fueron visitadas por un ser humano. En ese sentido, cualquiera sea la estrategia utilizada para descubrir las sesiones reales, esta debe satisfacer dos criterios esenciales: las actividades realizadas por una persona real pueden ser agrupadas entre si y el conjunto en actividades que pertenecen a la misma visita (otros objetos requeridos por la página web visitada) también pertenecen al mismo grupo.

Hay varias técnicas para sesionización, las cuales pueden ser agrupadas en dos estrategias mayores: *proactiva y reactiva* [26].

Las **Estrategias Proactivas** intentan identificar el usuario utilizando métodos de identificación como cookies que consisten en una pieza de código asociado al sitio web. Cuando el visitante ingresa al sitio por primera vez, una cookie es enviada al navegador. Luego, cuando la página es revisitada, el navegador muestra el contenido de la cookie al servidor web y automática-

mente la identificación toma lugar. El método tiene problemas desde el punto de vista técnico y también con respecto a la privacidad del usuario. Primero, si el sitio es revisitado después de varias horas, la sesión será considerada muy larga, y será entonces una nueva sesión. En segundo lugar, algunos aspectos de las cookies parecen incompatibles con los principios de protección de datos de algunas comunidades, como la Unión Europea [26]. Finalmente, las cookies pueden ser fácilmente detectadas y desactivadas pro el visitante.

Las Estrategias Reactivas son no invasivas con respecto a la privacidad y hacen uso de la información contenida sólo en los web logs y consiste en el procesamiento de los registros para generar un grupo de sesiones reconstruidas.

En el análisis del sitio web, el escenario general es que los sitios web usualmente no implementan mecanismos de identificación. La utilización de estrategias reactivas puede llegar a ser más útil. Estas pueden ser clasificadas en dos grupos principales [4, 10]:

- Heurísticas orientadas a la navegación: asumen que el visitante llega a páginas a través de hyperlinks desde otras páginas. Si el requerimiento de una página es inalcanzable a través de las páginas previamente visitadas por el usuario, una nueva sesión es iniciada.

- Heurísticas Orientadas al tiempo: se coloca un tiempo máximo de duración, que es usualmente 30 minutos para la sesión completa [7]. Basado en este valor se pueden identificar las transacciones pertenecientes a una sesión específica utilizando filtros programados.

3.1.1. Procesando el contenido textual de una página web

Hay varios métodos para comparar el contenido de dos páginas web, aquí se considera el texto libre dentro de las páginas web. El proceso común es coincidir los términos que componen el texto libre, por ejemplo, mediante la aplicación de un proceso de comparación de palabras. Un análisis más complejo incluye información semántica contenida en el texto libre que involucra también una tarea de aproximación de términos comparados.

La información semántica es fácil de extraer cuando el documento incluye información adicional acerca del contenido del texto, por ejemplo, etiquetas de marcado. Algunas páginas web permiten la comparación de documentos mediante la información estructural contenida en las etiquetas HTML, incluso con restricciones. Este método es utilizado en [28] para comparar páginas escritas en lenguajes diferentes con una estructura HTML similar. La comparación es enriquecida por la aplicación de un proceso de equiparar el contenido textual [29], el cual considera una tarea inicial de traducción a ser completada. El método es altamente efectivo cuando el lenguaje utilizado es el mismo en las páginas que se encuentran en comparación. Una breve encuesta de algorit-

mos para comparar documentos por la utilización de estructuras similares es encontrada en [5].

Las comparaciones son realizadas por una función que retorna un valor numérico mostrando similitudes o diferencias entre dos páginas web. Esta función puede ser utilizada en algoritmos de web mining para procesar un conjunto de páginas web, las cuales pueden pertenecer a una comunidad web o un sitio web aislado. El método de comparación debe considerar un criterio de eficiencia en el procesamiento de contenido de páginas web [13]. Aquí el modelo de vector espacial [24], permite una representación vectorial simple de las páginas web y. mediante el uso de comparación de distancia entre vectores, provee de una medida de las diferencias y similitudes entre páginas.

Las páginas web deben ser limpiadas antes de transformarlas en vectores, tanto para reducir el número de palabras - no todas las palabras tienen el mismo peso - y hacer el proceso más eficiente. Por esto, el proceso debe considerar los siguientes tipos de palabras:

- Etiquetas HTML: En general, estas deben ser limpiadas. Sin embargo, la información contenida en cada etiqueta puede ser utilizada para identificar palabras importantes en el contexto de una página. Por ejemplo, la etiqueta "<titulo>" enmarca el tema central de la página web, es decir, de una noción aproximada del significado semántico de la palabra y, es incluida en la representación vectorial de la página.

- Palabras de detención. (por ejemplo pronombres, preposiciones, conjunciones, etc.).

- Stem de palabras. Después de aplicar el proceso de remoción del sufijo de la palabra (stemización de la palabras [22]), obtenemos la raíz de la palabra o stem.

Para el propósito de representación vectorial, sea R el número total de palabras diferentes y Q el número de páginas en el sitio web. Una representación vectorial del conjunto de páginas es una matriz M de tamaño R x Q.

$$M = (m_{ij}), \text{ con } i = 1, \ldots, R \text{ y } j = 1, \ldots, Q \tag{1}$$

Donde m_{ij} es el peso de la palabra i en la página j.
Basado en *tfxidf-weighting* introducido en [24] los pesos son estimados como:

$$m_{ij} = f_{ij}(1 + sw_i)\log(\frac{Q}{n_i}) \tag{2}$$

Aquí, f_{ij} es el número de ocurrencias de la palabra i en la página j y n_i es el número total de documentos del sitio web que contienen la palabra i.

Adicionalmente, la importancia de las palabras es incrementada por la identificación de palabras especiales, las cuales correspondiente a los términos en la página web que son más importantes que otras, por ejemplo, palabras destacadas (haciendo uso de etiquetas HTML), palabras utilizadas por el usuario en la búsqueda de información y, en general, palabras que implican los deseos y necesidades de los usuarios. La importancia de palabras especiales es almacenada en un arreglo sw de dimensión R, donde sw_i representa un peso adicional para la i-ésima palabra.

El arreglo sw permite al modelo de vector espacial incluir ideas acerca de información semántica contenida en el texto de la página web por la identificación de palabras especiales.

Las fuentes comunes de palabras especiales son:

1. E-Mails: El ofrecimiento de envío de emails por parte del usuario para la plataforma de call center. Este texto enviado es una fuente para identificar las palabras más recurrentes. Sea $ew_i = \frac{w^i_{email}}{TE}$ el arreglo de palabras contenidas en los e-mails, que también están presentes en el sitio web, donde wi email es la frecuencia de la i-ésima palabra y TE es la cantidad total de palabras en el grupo completo del arreglo de palabras de e-mail.

2. Palabras destacadas. En un sitio web, hay palabras con etiquetas especiales, como diferentes fuentes, por ejemplo, itálica, negrita, o palabras pertenecientes al título. Sea $nw_i = \frac{w^i_{marks}}{TM}$ el arreglo de palabras destacadas dentro de las páginas web, donde w^i_{marks} es la frecuencia de la i-ésima palabra y TM es la cantidad de palabras destacadas en el sitio web completo.

3. Palabras de consultas: Un banco, por ejemplo, tiene motores de búsqueda a través de las cuales los usuarios pueden preguntar por asuntos específicos, por la introducción de palabras clave. Sea $aw_i = \frac{w^i_{ask}}{TA}$ el arreglo de palabras usadas por el usuario en el motor de búsqueda y que esta contenida en el sitio web, donde w^i_{ask} es la frecuencia de la i-ésima palabra y TA es la cantidad total de palabras en el grupo completo.

4. Sitios web relacionados. Usualmente un sitio web pertenece a un segmento de mercado, en este caso el mercado de las instituciones bancarias. Luego, es posible recolectar páginas de sitios web que pertenecen a otros sitios en el mismo mercado. Sea $rw_i = \frac{w^i_{rws}}{RWS}$ el arreglo con palabras utilizadas en el mercado de sitios web incluyendo el sitio web bajo estudio,

donde w^i_{rws} es la frecuencia de la i-ésima palabra y RWS es el número total de palabras en todos los sitios web considerados.

La expresión final $sw_i = ew_i + mw_i + aw_i + rw_i$ es la suma simple de los pesos descritos anteriormente.

En la representación vectorial, cada columna de la matriz M es una página web. Por ejemplo, la késima columna mik con i = 1,. . . ,R es la k-ésima página en el grupo completo de páginas.

Definición 1 (Vector de Palabras por página) *es un vector* $WP^k = (wp^k_1, \ldots, wp^k_R) = (m_{1k}, \ldots, m_{rk})$*, con* $k = 1, \ldots, Q$*, es la representación vectorial de la k-ésima página en el grupo de páginas bajo análisis.*

Con las páginas web en representación vectorial, es posible utilizar la medida de distancia para comparar los contenidos de texto. La distancia común es el coseno del ángulo calculado como:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} WP^i_k \cdot WP^j_k}{\sqrt{\sum_{k=1}^{R}(WP^i_k)^2}\sqrt{\sum_{k=1}^{R}(WP^j_k)^2}} \tag{3}$$

La ecuación (3) permite comparar el contenido de dos páginas web, retornando un valor numérico entre [0, 1]. Cuando las páginas son totalmente diferentes, $dp = 0$, y cuando son las mismas, $dp = 1$. Otro aspecto importante es que la ecuación 3 cumple con el requerimiento de ser computacionalmente eficiente, lo cual la hace más apropiada para ser utilizada en algoritmos de web mining.

4. Extrayendo las preferencias de contenido del usuario de las páginas web

Diferentes técnicas son aplicadas para analizar el comportamiento del usuario en el sitio web, desde una simple estadística de uso de una página hasta complejos algoritmos de web mining. En el último caso, la investigación se concentra en predicciones acerca de cuales páginas el usuario visitará y la información que esta buscando. Principalmente por la utilización de la combinación de las aproximaciones de WUM y WCM, el propósito es analizar las preferencias de texto del usuario web y por esta vía, identificar cuales palabras atraen la atención del usuario durante su navegación en el sitio. Previamente a la aplicación de una herramienta de web mining, la data relacionada con el comportamiento del usuario ha sido procesada para crear vectores de características, cuyos componentes dependerán de la implementación particular del algoritmo de web mining a utilizar y la preferencia de patrones ha ser extraídos.

4.1. Modelando el comportamiento del usuario web

La mayoría de los modelos de comportamiento de usuario web examinan la secuencia de páginas visitadas para crear vectores de características que representan el perfil de navegación del usuario web [12, 21, 36]. Estos modelos analizan el comportamiento de navegación del usuario en un sitio web mediante la aplicación de algoritmos que extraen los patrones de navegación. El siguiente paso es examinar las preferencias del usuario, definido como el contenido preferido de la página web por el usuario; y este es el contenido de texto que captura la atención especial, dado que es utilizada para encontrar información interesante relacionada a un tema particular por un motor de búsqueda. Por lo tanto, es necesario incluir una nueva variable como parte de la información del vector de comportamiento del usuario web acerca del contenido y tiempo gastado en cada página web visitada.

Definición 2 (Vector de comportamiento del usuario (UBV)) *Es un vector* $\nu = [(p_1, t_1), \ldots, (p_m, t_n)]$ *, donde son los parámetros que representan la i-ésima página del visitante y el tiempo gastado en ella en la sesión, respectivamente. En esta expresión,* p_i *es el identificador de la página.*

En la definición 2, el comportamiento del usuario en un sitio web es caracterizado por:

1. Secuencia de páginas; la secuencia de páginas visitadas y registradas en los archivos log. Si el usuario retorna a una página almacenada en el caché del browser, esta acción puede no ser registrada.

2. Contenido de la página; representa el contenido que puede ser texto libre, imágenes, sonidos, etc. Para propósitos de este paper, el texto libre es el utilizado principalmente para representar una página web.

3. Tiempo gastado, tiempo utilizado por el usuario en cada página. Para la página, el porcentaje de tiempo gastado en cada página durante la sesión del usuario puede ser directamente calculado.

4.2. Analizando las preferencias de texto de los usuarios

El objetivo es determinar las palabras más importantes para un sitio web dado para los usuarios, mediante la comparación de las preferencias de texto libre, a través del análisis de páginas visitadas y de tiempo gastado en cada una de ellas [34]. Sin embargo, difiere de las propuestas mencionadas anteriormente, dado que el ejercicio es encontrar las palabras clave que atraen y retienen a los usuarios en el uso de data disponible en la web. La expectación esta en involucrar usuarios pasados y actuales en un proceso continuo de determinación de palabras clave.

Las preferencias del contenido web del usuario son identificadas por la comparación de contenido de las páginas visitadas, [34, 33, 35] por la aplicación

del modelo de vector espacial a las páginas web, con la variante propuesta en la sección 3.2, ecuación (2). Los temas principales de interés pueden ser encontrados por el uso de la medición de la distancia entre vectores (por ejemplo, distancia euclidiana).

Desde el vector de comportamiento del usuario (UBV), las páginas más importantes son seleccionadas asumiendo que el grado de importancia esta correlacionado al porcentaje de tiempo gastado en cada página. El UBV se ordena de acuerdo al porcentaje de tiempo total gastado en cada página. Las ι página más importantes, es decir, las primeras ι páginas, son seleccionadas.

Definición 3 (Vector de Páginas Importantes (IPV)). *Es un vector* $\vartheta_\iota(\nu) = [(\rho_1, \tau_1), \ldots, (\rho_i, \tau_i)]$ *, donde* (ρ_ι, τ_ι) *es el componente que representa la* ι*-ésima página más importante y el porcentaje de tiempo gastado en ella por la sesión.*

Sean α y β dos UBV. La medida de similitud propuesta entre los dos IPV es introducida en la ecuación 4 como:

$$st(\vartheta_\iota(\alpha), \vartheta_\iota(\beta)) = \frac{1}{\iota} \sum_{k=1}^{\iota} \text{mín}\{\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}\} dp(\rho_k^\alpha, \rho_k^\beta) \tag{4}$$

El primer elemento en (4) indica el interés del usuario en las páginas visitadas. Si el porcentaje de tiempo gastado por los usuarios α y β en la k-ésima página visitada es cercano a la otra, el valor de la expresión min.,. será cercano a 1. En el caso opuesto, será cercano a 0. El segundo elemento en (4) es dp, la distancia entre páginas representada en forma vectorial, introducida en (3). En (4) el contenido de las páginas más importantes es multiplicado por el porcentaje de tiempo total gastado en cada página. Esto permite a las páginas con contenidos similares ser distinguidas por intereses diferentes de usuarios.

4.3. Identificando palabras clave del sitio web

Una palabra clave de un sitio web (o web site keyword) es definido como "una palabra o posiblemente un grupo de palabras que hacen de una página web más atractiva para un usuario eventual durante su visita al sitio web" [32]. Es interesante notar que las mismas palabras clave del sitio web pueden ser utilizadas por el usuario en un motor de búsqueda, cuando este está en busca de contenido web.

Para encontrar palabras clave de un sitio web, es necesario seleccionar las páginas web con el contenido textual que es significativo para los usuarios. La suposición es que existe una relación entre el tiempo gastado por el usuario en una página web y su interés en el contenido [31]. La relación es almacenada por el vector de páginas importantes (IPV), dando la información necesaria para extraer las palabras clave de un sitio web a través de la utilización de una herramienta de web mining.

Entre estas técnicas de web mining, se debe colocar especial atención a los algoritmos de clustering. La suposición es, dado un grupo de clusters extraídos de la información generada durante la formación de las sesiones de los usuarios en el sitio en, es posible el extraer las preferencias de los usuarios mediante el análisis de los contenidos del cluster. Los patrones en cada cluster detectado podrían ser suficientes para extrapolar el contenido que él o la usuario esta buscando [20, 23, 30].

En cada IPV, el componente página tiene una representación vectorial presentada por la ecuación (2). En esta ecuación, un paso importante es el cálculo de pesos considerados en el arreglo de palabras especiales swi. Las palabras especiales son diferentes a las palabras normales en el sitio, dado que pertenecen a una fuente alternativa y relacionada o ellas tienen una información adicional mostrando su importancia en el sitio, por ejemplo, una etiqueta HTML que enfatiza una palabra.

El algoritmo de clustering es utilizado para agrupar IPV similares por comparación de la cada componente de tiempo y página del vector, siendo importante el uso de la medida de similitud presentada en la ecuación (4). El resultado debería ser un grupo de clusters cuya calidad debe ser chequeada mediante el criterio de aceptación / rechazo. Un camino simple es aceptar los clusters cuyas páginas comparte un tema principal similar, y en otro caso, rechazar el cluster. En este punto, es necesario conocer que páginas en el sitio son cercanas con los vectores del cluster. Debido a que conoceremos la representación vectorial de las páginas web del sitio y utilizando la ecuación (3) podemos identificar la página más cercana de un cluster dado y de esta forma obtener las páginas adecuadas a un cluster para revisar si las páginas comparten un tema principal en común.

Para cada cluster aceptado y recordando que los centroides contienen páginas donde los usuarios gastan más tiempo durante su sesión respectiva y en la representación vectorial tienen los pesos más altos, el procedimiento de identificación de palabras clave del sitio web es aplicar una medida, descrita en la ecuación (5) (miembro geométrico) para calcular la importancia de cada palabra

$$kw[i] = \sqrt{\prod_{p \epsilon \varsigma} m_{ip}}, \tag{5}$$

donde $i = 1, \ldots, R$ y kw es un arreglo que contiene los pesos para cada palabra relativa a un cluster dado y ς el grupo de páginas representando el cluster. Las palabras clave del sitio web son el resultado del ordenamiento de kw y de la detección de palabras con los pesos más altos, por ejemplo, las 10 palabras con mayor peso.

5. Extrayendo patrones de los datos originados en un sitio web real

Para propósitos experimentales, el sitio web seleccionado debe ser complejo con respecto a varias características: número de visitas, actualización periódica (preferiblemente mensual con el fin de estudiar la reacción de los usuarios a los cambios) y ser rico en contenido textual. La página web de un banco virtual Chileno (sin sucursales físicas, todas las transacciones realizadas electrónicamente) cumple con estos criterios. Cabe destacar que para efectos de privacidad de los datos usados en la investigación, se firmó un acuerdo con el banco, por lo cual so nombre no pude ser mencionado.

Las principales características del sitio web del banco son las siguientes; presentado en Español, con 217 páginas web estáticas y aproximadamente ocho millones de filas en los registros de web log para un periodo de estudio entre Enero y Marzo del 2003.

El comportamiento del usuario en el sitio web del banco es analizado en dos formas. Primero, mediante la utilización de los archivos de registro que contienen información acerca del visitante y del comportamiento de navegación del cliente. Esta información requiere de una reconstrucción previa y limpieza antes de que las herramientas de web mining sean aplicadas. Segundo, la web data en el sitio web en si mismo, específicamente el contenido textual de las páginas web - esto también necesita de un preprocesamiento y limpieza.

5.1. Proceso de reconstrucción de sesiones

La Fig. 5.1 muestra parte de los registros del sitio web bancario e incluye tanto a clientes identificados como visitantes anónimos.

Figura 1: Extracto de un archivo de web log generado en el sitio web de un banco

El acceso de los clientes al sitio es a través de una conexión segura, utilizando un protocolo SSL que permite el almacenamiento de un valor de identificación en el parámetro de autentificación de usuario en el archivo de registros

web. Otro modo de identificación de usuarios es mediante cookies, pero algunas veces estas son desactivadas por los usuarios en sus navegadores. En este caso sería necesario el reconstruir la sesión del visitante.

Durante el proceso de reconstrucción de sesiones, se aplican filtros a los registros del sitio web. En este caso particular, solo se utilizan registros de requerimiento de páginas web para analizar el comportamiento específico del usuario en el sitio. También es importante la limpieza de sesiones anormales, por ejemplo, web crawlers, como es mostrado en la Fig. 1, línea 4, donde un robot perteneciente a Google es detectado.

Las filas de los registros log del sitio web contienen cuatro meses de transacciones, con aproximadamente 8 millones de registros. Sólo se consideran los registros relacionados con páginas web para la reconstrucción de sesiones y análisis del comportamiento del usuario; la información que apunta a otros objetos como imágenes, sonidos, etc, son limpiadas.

5.2. Preprocesamiento del contenido de una página web

Mediante la aplicación de filtros a los textos de las páginas web, se ha encontrado que en el sitio completo contiene R=2034 palabras diferentes para ser utilizadas en el análisis.

Considerando los pesos de las palabras y la especificación de palabras especiales, fue utilizado el procedimiento presentado en la sección 3.2, con el fin de calcular sw_i, en la ecuación 2. Las fuentes de datos fueron:

1. Palabras destacadas. Dentro de las páginas web, se encontraron 743 palabras diferentes después de la aplicación del paso de preprocesamiento y limpieza.

2. Sitios web relacionados: Cuatro sitios web fueron considerados, cada un de ellos con aproximadamente 300 páginas.

El número total de palabras diferentes fue de 9253, con 1842 de ellas contenidas en el contenido del sitio web.

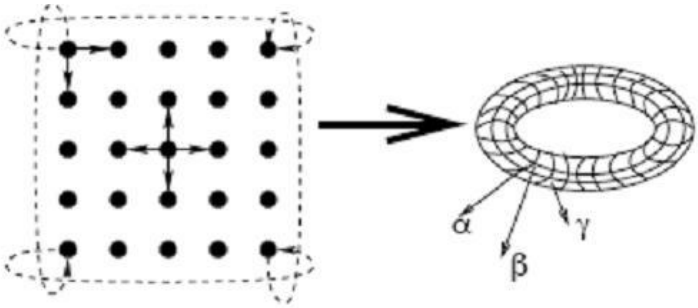
Después de la identificación de palabras especiales y sus respectivos pesos, es posible calcular el peso final para cada palabra en la totalidad del sitio web, por la aplicación de la ecuación (2). Luego, se obtiene la representación vectorial para todas las páginas del sitio

5.3. Analizando las preferencias de texto del usuario

Dos redes neuronales fueron aplicadas al web data para la identificación de clusters. La red neuronal artificial del tipo Kohonen (Self Organizing Feature Map; SOFM) y K-means.

Esquemáticamente, una red SOFM es una red neuronal artificial no supervisada, correspondiente a un arreglo de neuronas de dos dimensiones. Cada neurona esta constituida por un arreglo bidimensional de vectores de n dimensiones cuyos componentes son los pesos sinápticos. Por construcción, todas las neuronas reciben el mismo input en un momento determinado. La noción de vecindad entre neuronas define diversas topologías. Para el caso de este trabajo, se utilizó la topología toroidal [38] que significa que las neuronas localizadas de un borde, son cercanas al borde opuesto. La ventaja de la topología radica en que mantiene la continuidad de los clusters o cuando la data corresponde a secuencias de eventos.

Figura 2: de Vectores de Páginas Importantes en un SOFM con topología toroidal.

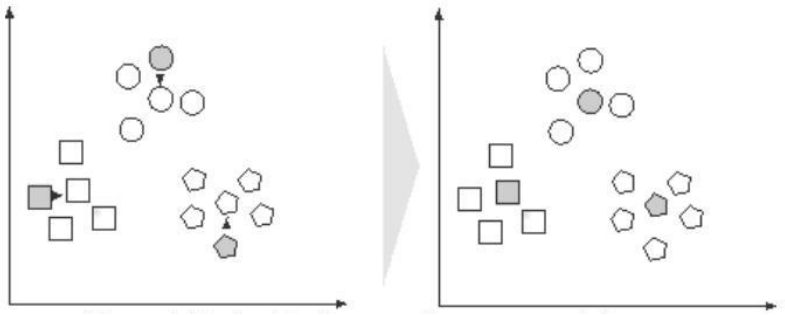


K-means es una red de aprendizaje supervisada y predefiniendo el número de centroides, genera las agrupaciones de vectores llamados miembros en torno a ellos. K-means para detectar las pertenencias a sus centroides tradicionalmente utiliza la distancia euclideana para discernir que centroide es más representativo para un vector. Puesto que la investigación se centra en un vector de comportamiento del usuario con una estructura diferente a la euclideana, se hace modificación de esta red de aprendizaje y se utiliza la medida de similitud presentada en la ecuación (4) para establecer las pertenencias a los centroides correspondientes. Para el caso de esta investigación, el principal input de este algoritmo - los K centroides - será originado por el resultado que entregue la red SOFM que será inicialmente utilizada para el análisis del comportamiento del usuario y que parte de los resultados que retorne serán los clusters detectados. La Fig. 3 muestra el comportamiento de los centroides a medida que se van encontrando mejores representantes.

5.4. Analizando las preferencias del usuario con una red SOFM

Se ha fijado en 3 el número máximo de dimensiones del vector. Luego, un SOFM con 3 neuronas de entrada y 32*32 neuronas de salida fue utilizado

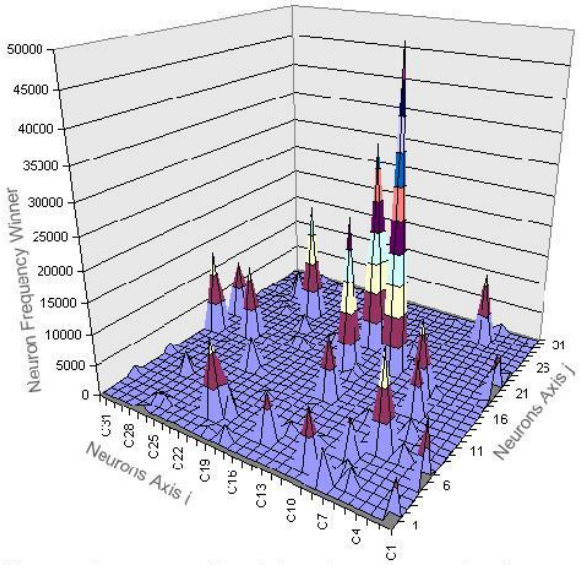
Figura 3: Evolución de centroides en una red K-means



para encontrar los clusters de vectores de páginas importantes.

La Fig. 4 muestra las posiciones de las neuronas en el SOFM en los ejes x e y. El eje z es la frecuencia normalizada de veces que una neurona gana durante el entrenamiento.

Figura 4: Clusters de vectores de páginas importantes desde una red SOFM.



La Fig. 4, muestra 8 cluster principales que contienen información acerca de las páginas más importantes del sitio web. Sin embargo, sólo 5 fueron aceptadas. El criterio de aceptación / rechazo es simple; si las páginas de un centroide de cluster tienen el mismo tema principal, entones el cluster es aceptado, de otra forma se rechaza.

Los centroides de los clusters son mostrados en el Cuadro 1. La segunda columna contiene las neuronas centrales (neuronas ganadoras) para cada cluster y representa las páginas visitadas más importantes.

Cuadro 1: Vectores de páginas importantes obtenidos con SOFM.

Cluster	Páginas Visitadas
1	(171, 130, 159)
2	(76, 58, 130)
3	(175, 78, 10)
4	(78, 32, 130)
5	(130, 171, 159)

5.5. Analizando las preferencias del usuario con una red K-means

Desde el resultado obtenido con la aplicación de SOFM, se inicializa el entrenamiento de la red K-means. La cantidad de centroides es de acuerdo al número final aceptado. Luego, el proceso se inicializó con 5 centroides. La detención de asignación de miembros al clusters se produce cuando los mejores representantes de cada centroide van variando de menor manera llegando en un punto a quedar prácticamente establecido el centroide que será el representante de los miembros.

Para cada centroide se obtiene las páginas visitadas representativas de los grupos. En el Cuadro 2 se muestran las páginas visitadas de los representantes de los clusters y la cantidad de miembros identificados en ellos.

Cuadro 2: Vectores de páginas importantes obtenidos con K-means.

Cluster	Páginas Visitadas
1	(117,192,19)
2	(21,10,179)
3	(205,128,210)
4	(55,18,41)
5	(24,104,95)

5.6. Identificación de web site keywords

Se requiere un paso final para obtener las palabras clave de un sitio web: analizar cuales son las palabras que tienen una mayor importancia relativa con respecto al sitio web completo.

Las palabras clave y su importancia relativa en cada cluster son obtenidas por la aplicación de la ecuación (5). Por ejemplo, si el cluster es ($\varsigma = \{171, 130, 159\}$), entonces $kw[i] = \sqrt[3]{m_{i171}, m_{i130} m_{i159}}$, con $i = 1, \ldots, R$.

Finalmente, ordenando las kw de forma descendente, podemos seleccionar las k palabras más importantes para cada cluster, por ejemplo $k = 5$.

No se nos permite mostrar las palabras clave específicas debido a la cláusula de confidencialidad con el banco, por esta razón las palabras son numeradas. El Cuadro 3 muestra las palabras encontradas con el método propuesto.

El Cuadro 4 muestra un grupo seleccionado de palabras clave de todos los clusters. Las palabras clave en si, sin embargo, no tienen mucho sentido. Estas necesitan un contexto de página web donde ellas podrían ser utilizadas como palabras especiales, por ejemplo, palabras destacadas para enfatizar un concepto o como palabras vinculadas a otras páginas.

Cuadro 3: Las 5 palabras más importantes por cluster

C	Palabras Clave	Peso ordenado
1	$(w_{2032}, w_{1233}, w_{287}, w_{1087}, w_{594})$	(2.35,1.93,1.56,1.32,1.03)
2	$(w_{1003}, w_{449}, w_{895}, w_{867}, w_{1567})$	(2.54,2.14,1.98,1.58,1.38)
3	$(w_{1005}, w_{948}, w_{505}, w_{1675}, w_{1545})$	(2.72,2.12,1.85,1.52,1.31)
4	$(w_{501}, w_{733}, w_{385}, w_{684}, w_{885})$	(2.84,2.32,2.14,1.85,1.58)
5	$(w_{200}, w_{1321}, w_{206}, w_{205}, w_{1757})$	(2.33,2.22,1.12,1.01,0.93)

Cuadro 4: Parte de las palabras descubiertas.

#	Palabras Clave
1	Cuenta
2	Fondo
3	Inversión
4	Tarjeta
5	Hipotecario
6	Seguro
7	Cheques
8	Crédito

La recomendación específica es utilizar las palabras clave como “palabras para escribir” en un sitio web, es decir, los párrafos escritos en la página deberían incluir algunas palabras clave y algunas podrían ser un enlace a otras páginas.

Además es posible sobre la base de este ejercicio el realizar recomendaciones de contenidos de texto. Sin embargo, para reiterar, las palabras clave no funcionan de forma separada sino que requieren de un contexto que las utilice. Revisando el Cuadro 2, para cada cluster, la palabra clave descubierta podría ser utilizada para reescribir un párrafo o una página web completa. Adicionalmente, es importante insertar palabras clave para destacar conceptos específicos.

Las palabras clave también son utilizadas como palabras índice para un motor de búsqueda, es decir, algunas podrían ser utilizadas para personalizar

el crawler que visita sitios web y carga páginas. Luego, cuando un usuario esta buscando por una página en específico en un motor de búsqueda, la probabilidad de obtener el sitio web se incrementa.

5.7. Mejorando el contenido textual el sitio web

Las palabras clave son conceptos para motivar los intereses de los usuarios y hacerlos visitar el sitio web. Están para ser jugadas dentro de su contexto como palabras aisladas que pueden tener un pequeño sentido , dado que los clusters representar contextos diferentes. La recomendación específica es utilizar palabras clave como "palabras para escribir" en el sitio web.

En cuanto cada página contiene un contenido de texto específico, es posible asociar las palabras clave de un sitio web a un contenido de la página; y desde esta sugerir la revisión o reconstrucción de un nuevo contenido en el sitio web. Por ejemplo, si la nueva versión de la página es relacionada con "tarjetas de crédito", entonces las palabras clave del sitio web "crédito, puntos y promociones" deben ser asignadas para la reescritura del contenido textual de la página.

5.8. Testeo de la efectividad de las recomendaciones de texto

La detección y aplicación de web site keywords no garantizan el éxito de aplicación en un contenido textual. Incluso, el riesgo de utilizarlas puede generar disgusto en un usuario habitual y por lo tanto abandonar o dejar de utilizar el sitio web. Como medida precautoria, se realizaron test de efectividad de las web site keywords. Sobre el contenido del sitio web se extrajeron 10 párrafos que contenían para el caso de 5 de ellos web site keywords y otros no las contenían. El resultado se realizó sobre un universo de 10 personas con el fin de conocer la recepción que ellos tenían respecto a párrafos que contenían las palabras detectadas, según el contexto de si entregaban información relevante en un sitio bancario. El resultado de este test es el que se muestra en el Cuadro 6.

Como se puede apreciar, aquellas palabras que contenían web site keywords eran para el usuario mucho más interesantes e importantes en el contexto de navegación en que estaban inmersos, versus aquellos párrafos en que no había presencia de dichos web site keywords. Las web site keywords atraen la atención del usuario y pueden ser una muy buena guía en el diseño de contenidos específicos de un sitio web. Esta combinación de elementos que se alinean a los que el usuario busca puede otorgar un mejor resultado en la satisfacción de los clientes.

Cuadro 5: Párrafos testeados para análisis de keywords.

#	Incluye web site keyword	Párrafo
1	Si	Orientado a empresas que deseen manejar excedentes de caja, así como a Personas que quieran mantener parte de sus recursos **invertidos** en un **fondo mutuo**, cuya cartera esté compuesta exclusivamente por instrumentos de **deuda** nacional, obteniendo rentabilidad y liquidez a corto plazo.
2	Si	Solicitándolos con un día hábil bancario de antelación, se pagarán mediante **cheques** nominativos, **vales** vista o **depósitos** en **cuentas** corrientes, de acuerdo a sus instrucciones.
3	Si	Este plan busca otorgar a tus **Ahorros Previsionales** Voluntarios acumulados a esta fecha y los futuros, una atractiva y segura **rentabilidad** que te permitirá poder mejorar considerablemente tus **ahorros** para una mejor **pensión.**
4	Si	Para obtener información de tu **Cuenta Corriente** y de tu **Línea** de **Sobregiro** debes seguir los siguientes pasos
5	Si	El Servicio de Mensajería es un servicio de entregas y **retiros** de dinero, especies valoradas y documentos que podrás utilizar siendo **cliente**
6	No	Para solicitar tu Plan debes completar la siguiente información y se contactarán contigo.

6. Conclusiones

Cuando un usuario visita un sitio web, hay una correlación entre el máximo de tiempo gastado por sesión en una página y su contenido de texto libre. Esto permite modelar las preferencias del usuario a través del "Vector de Páginas Importantes (IPV)", el cual es la estructura de datos básica de almacenamiento de páginas donde el usuario gasta más tiempo durante e su sesión. Mediante la utilización de IPV como entrada en un SOFM y K-means, se pueden identificar clusters que contienen la navegación del usuario e información de sus preferencias de contenido.

El criterio de aceptación / rechazo de un cluster es simple: si las páginas dentro de cada cluster están relacionadas con un tema principal similar, entonces el cluster es aceptado, en caso contrario, se rechaza. Aplicando este

Cuadro 6: párrafos testeados para análisis de keywords.

		Opinión de aceptabilidad				
#	Incluye web site Keyword	Irrelevante	Moder. irrelevante	Algo relevante	Moder. relevante	relevante
1	Si				8	2
2	Si			4	4	2
3	Si			4	2	4
4	Si				7	3
5	Si			1	2	7
6	No	1	3	5	1	
7	No	3	2	5		
8	No	6	4			
9	No	5	2	1	2	
10	No	7	2	1		

criterio, 5 clusters son aceptados y el patrón contenido en cada una de ellas fue utilizado para extraer las palabras clave del sitio web.

El texto contenido en las páginas web puede ser mejorado utilizando las palabras clave del sitio web, y por esta vía atraer la atención del usuario cuando están visitando un sitio web. Sin embargo, es necesario recordar que estas palabras no pueden ser utilizadas de forma individual, de hecho necesitan de un contexto, el cual es provisto por un ser humano.

Como validación de las palabras detectadas, se realizó un testeo de párrafos que contenían dichos web site keywords versus otros que no contenían. El resultado fue satisfactorio corroborando la importancia de las palabras pues aquellos contenidos con web site keywords parecían más relevantes que otras que no contenían dichas palabras, por lo tanto el interés del usuario en contenidos con los keywords se hace mayor y de ahí la importancia de dar uso a estas palabras en los párrafos del contenido.

Como trabajo futuro, se aplicará la metodología en otros web data, por ejemplo las imágenes y objetos no textuales, con el fin de identificar cuales elementos atraen la atención del usuario en el sitio web.

Agradecimientos: Este trabajo fue parcialmente financiado por el Instituto Milenio Sistemas Complejos de Ingeniería

Referencias

[1] Green, Paul E. and V. Srinivasan (1990), "Conjoint Analysis in Marketing Research: New Developments and Directions", Journal of Marketing 54, 4, 3-19.

[2] E. Amitay and C. Paris. "Automatically summarizing web sites: Is there

any way around it?" In Procs. of the 9th Int. Conf. on Information and Knowledge Management, pages 173-179, McLean, Virginia, USA, 2000.

[3] R. Baeza-Yates. "Web usage mining in search engines", chapter Web Mining: Applications and Techniques, pages 307-321. Idea Group, 2004.

[4] B. Berendt, A. Hotho, and G. Stumme. "Towards semantic web mining". In Proc. in First Int. Semantic Web Conference, pages 264-278, 2002.

[5] B. Berendt and M. Spiliopoulou. "Analysis of navigation behavior in web sites integrating multiple information systems". The VLDB Journal, 9:56-75, 2001.

[6] D. Buttler. "A short survey of document structure similarity algorithms". In Procs. Int. Conf. on Internet Computing, pages 3-9, 2004.

[7] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. "Focused web searching with pdas". Computer Networks, 33(1- 6):213-230, June 2000.

[8] L. D. Catledge and J. E. Pitkow. "Characterizing browsing behaviors on the world wide web". Computers Networks and ISDN System, 27:1065-1073, 1995.

[9] G. Chang, M. Healey, J. McHugh, and J. Wang. "Mining the World Wide Web". Kluwer Academic Publishers, 2003.

[10] W. Chuang and J. Yang. "Extracting sentence segment for text summarization? a machine learning approach". In Procs. Int. Conf. ACM SIGIR, pages 152-159, Athens, Greece, 2000.

[11] R. Cooley, B. Mobasher, and J. Srivastava. "Data preparation for mining world wide web browsing patterns". Journal of Knowlegde and Information Systems, 1:5-32, 1999.

[12] U. Hahn and I. Mani. "The challenges of automatic summarization". IEEE Computer, 33(11):29-36, 2000.

[13] A. Joshi and R. Krishnapuram. "On mining web access logs". In Proc. of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 63- 69, 2000.

[14] A. P. Jr and N. Ziviani. "Retrieving similar documents from the web". Journal of Web Engineering, 2(4):247-261, 2004.

[15] D. Lawrie, B. W. Croft, and A. Rosenberg. "Finding topic words for hierarchical summarization". In Proc. 24th Int SIGIR Conf. on Research and Development in Information Retrieval, pages 349-357, New Orleans, Louisiana, USA, 2001. ACM Press.

[16] E. Liddy, K. McVearry, W. Paik, E. Yu, and M. McKenna. "Development, implementation and testing of a discourse model for newspaper texts". In Procs. Int. Conf. on ARPA Workshop on Human Language Technology, pages 159-164, Princeton, NJ, USA, 1993.

[17] G. Linoff and M. Berry. "Mining the Web". Jon Wiley & Sons, New York, 2001.

[18] S. Loh, L. Wives, and J. P. M. de Oliveira. "Concept based knowledge discovery in texts extracted from the web". SIGKDD Explorations, 2(1):29-39, 2000.

[19] I. Mani and M. Maybury. "Advances in automatic text summarization". MIT Press, Cambridge, Mass., 1999.

[20] S. Mitra, S. K. Pal, and P. Mitra. "Data mining in soft computing framework: A survey". IEEE Transactions on Neural Networks, 13(1):3-14, 2002.

[21] B.Mobasher, R. Cooley, and J. Srivastava. "Creating adaptive web sites through usage-based clustering of urls". In Procs.Int Conf IEEE Knowledge and Data Engineering Exchange, November 1999.

[22] B. Mobasher, R. Cooley, and J. Srivastava. "Automatic personalization based on web usage mining". Communications of the ACM, 43(8):142-151, 2000.

[23] M. F. Porter. "An algorithm for suffix stripping". Program; automated library and information systems, 14(3):130-137, 1980.

[24] T. A. Runkler and J. Bezdek. "Web mining with relational clustering". International Journal of Approximate Reasoning, 32(2-3):217-236, Feb 2003.

[25] G. Salton, A. Wong, and C. S. Yang. "A vector space model for automatic indexing". Communications of the ACM archive, 18(11):613-620, November 1975.

[26] M. Spiliopoulou. "Data mining for the web". In Principles of Data Mining and Knowledge Discovery, pages 588-589, 1999.

[27] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. "A framework for the evaluation of session reconstruction heuristics in web-usage analysis". INFORMS Journal on Computing, 15:171-190, 2003.

[28] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. "Web usage mining: Discovery and applications of usage patterns from web data". SIGKDD Explorations, 1(2):12-23, 2000.

[29] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. "Recovering traceability links in multilingual web sites". In Procs. Int Conf. Web Site Evolution, pages 14-21. IEEE Press, 2001.

[30] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. "Restructuring multilingual web sites". In Procs. Int. Conf. Software Maintenance, pages 290-299. IEEE Press, 2002.

[31] J. D. Velázquez and V. Palade. "A knowledge base for the maintenance of knowledge extracted from web data". Journal of Knowledge-Based Systems, 20(3):238-248, 2007.

[32] J. D. Velásquez, S. Ríos, A. Bassi, H. Yasuda, and T. Aoki. "Towards the identification of keywords in the web site text content: A methodological approach". International Journal of Web Information Systems, 1(1):11-15, March 2005.

[33] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. "A methodology to find web site keywords". In Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service, pages 285-292, Taipei, Taiwan, March 2004.

[34] J. D. Velásquez, H. Yasuda, and T. Aoki. "Combining the web content and usage mining to understand the visitor behavior in a web site". In Procs. 3th IEEE Int. Conf. on Data Mining, pages 669-672, Melbourne, Florida, USA, November 2003.

[35] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. "Using the kdd process to support the web site reconFig.tion". In Procs. IEEE/WIC Int. Conf. on Web Intelligence, pages 511-515, Halifax, Canada, October 2003.

[36] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. "A new similarity measure to understand visitor behavior in a web site". IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization, E87-D(2):389-396, February 2004.

[37] J. Xiao, Y. Zhang, X. Jia, and T. Li. "Measuring similarity of interests for clustering web-users". In ADC '01: Proceedings of the 12th Australasian conference on Database technologies, pages 107-114, Washington, DC, USA, 2001. IEEE Computer Society.

[38] K. Zechner. "Fast generation of abstracts from general domain text corpora by extracting relevant sentences". In Procs. Int. Conf. on Computational Linguistics, pages 986-989, 1996.

[39] J. D. Velásquez, H. Yasuda, T. Aoki, R. Weber and E. Vera (2003) "Using self organizing feature maps to acquire knowledge about visitor behavior in a web site". Lecture Notes in Artificial Intelligence, 2773(1): 951-958

Hacia la identificación de palabras importantes desde el punto de vista del usuario web.

Juan D. Velásquez and José I. Fernández
Departamento de Ingeniería Industrial, Universidad de Chile,[1]
E-mail: *jvelasqu@dii.uchile.cl, josferna@ing.uchile.cl*

Extracto
Presentamos una metodología para identificar aproximadamente cuales palabras atraen la atención del usuario cuando se encuentra visitando páginas de un sitio web. Estas palabras son llamadas "web site keywords" o "palabras claves de un sitio web" y pueden ser usadas para la creación de contenidos relacionados a un tópico específico de una forma más aproximada. A través de la utilización de las palabras correctas, podemos ayudar a los usuarios a encontrar lo que ellos buscan. Aplicando algoritmos de clustering, y asumiendo que existe una correlación entre el tiempo invertido en una página y el interés del usuario, segmentamos a los usuarios por comportamiento de navegación y preferencias de contenidos. A continuación, identificamos las palabras claves del sitio web. Esta metodología fue aplicada en data originada desde un sitio web real, mostrando su efectividad.

1. Introducción.

Para muchas compañías y/o instituciones, ya no es suficiente el tener un sitio web con productos de alta calidad y servicios. Lo que a menudo hace la diferencia entre un éxito o fracaso en un e-business es el potencial del sitio web en atraer o retener usuarios. Este potencial depende en el contenido del sitio, diseño y aspectos técnicos como tiempo de descarga de páginas del sitio web al navegador del usuario, entre otros. En términos de contenido, las palabras usadas en el texto libre en las páginas de un sitio web son muy importantes, ya que la mayoría de los usuarios realizan consultas basadas en términos en motores de búsqueda para encontrar información en la Web. Estas consultas son realizadas por palabras claves, es decir, una palabra o grupo de palabras [14] que caracterizan el contenido de un página web dada o un sitio web.

Mediante la identificación de las palabras más relevantes en las páginas de los sitios, desde el punto de vista del usuario, las mejoras pueden ser realizadas en el sitio web completo.

En trabajo presenta una metodología para analizar el comportamiento de navegación del usuario y sus preferencias de contenido a través de la aplicación de algoritmos de web mining en datos originados en la web, también llamados web data, específicamente registros de un sitio web (web logs) y su contenido textual apuntando a identificar cuales palabras atraen la atención del usuario cuando esta visitando páginas en un sitio web. Estas palabras son denominadas "palabras claves de un sitio web" [31] y pueden ser utilizadas para la creación de contenidos de texto mejoradas relacionadas con tópicos específicos.

Este trabajo se organiza de la siguiente forma: La sección 2 introduce una revisión breve acerca del trabajo relacionado. En la sección 3 se muestra el proceso de preparación para transformar la web data en vectores de características para ser utilizados como entrada en los algoritmos de web mining. En la sección 4, se explica la metodología para identificar las palabras claves de un sitio web y es aplicada en la sección 5. Finalmente, la sección 6 muestra las conclusiones principales de este trabajo.

2. Trabajo Relacionado.

Cuando un usuario visita un sitio web, la información de páginas visitadas es almacenada en archivos de registro llamados web logs. Entonces es directo conocer cuales páginas son visitadas y cuales no, e inclusive el tiempo gastado por el usuario en cada una de ellas. Debido a que usualmente las páginas contienen información de un tópico específico, es posible conocer aproximadamente la información de preferencia de los usuarios.

El desafío para analizar las preferencias de texto del usuario en el sitio web es doble. Primero la cantidad de registros en el archivo web log usualmente es enorme, y una parte

[1] Av. República 701, oficina 301, Santiago, CHILE, P.C. 837-0720

importante de ellas contiene información irrelevante acerca del comportamiento de navegación del usuario en el sitio. Segundo, el texto libre dentro de las páginas web es comúnmente plana, es decir, sin información adicional que nos permita conocer directamente cuales son las palabras que atraen la atención del usuario.

2.1. Minando la web data.

Las técnicas de web mining emergieron como resultado de la aplicación de teoría de data mining en el descubrimiento de patrones desde la web data [8, 16, 25]. El web mining no es una tarea trivial considerando que la web es una enorme colección de información heterogénea, no clasificada, distribuida, variante en el tiempo, semi estructurada y altamente dimensional. El web mining debe considerar tres importantes pasos: Preprocesamiento, descubrimiento de patrones y análisis de patrones [27].

2.2. Extracción de texto de páginas web y aplicaciones.

Las componentes de texto clave son partes de un documento completo, por ejemplo un párrafo, frase y una palabra que contiene información significante acerca de un tema particular, desde el punto de vista del usuario del sitio web. La identificación de estos componentes pueden ser útiles para mejorar el contenido textual de un sitio web.

Usualmente, las palabras claves en un sitio web están correlacionadas con las “palabras más frecuentemente utilizadas”. En [17] se desarrolla una técnica para extraer conceptos desde el texto de una página web. Los conceptos describen objetos del mundo real, eventos, pensamientos, opiniones e ideas en una estructura simple, como términos descriptivos. Entonces, utilizando el modelo de espacio vectorial, los conceptos son transformados en vectores de características, permitiendo la aplicación de algoritmos de clustering o clasificación a páginas web y así extraer conceptos.

3. Proceso de preparación de la Web Data.

De toda la información web disponible, la más relevante para el análisis del comportamiento y preferencias de navegación del usuario, son los registros (web logs) y las paginas web [33]. Los web logs contienen información acerca de la secuencia de navegación de páginas y el tiempo gastado en cada página visitada, aplicando el proceso de sesionización. La etapa de preprocesamiento implica, primero, un proceso de limpieza y, segundo, la creación de vectores de características como entrada a los algoritmos de web mining, dentro de la estructura definida por los patrones vistos.

3.1. El proceso de reconstrucción de sesiones.

El proceso de segmentación de las actividades de usuarios en sesiones individuales es llamado ***sesionización*** [10]. Este proceso es basado en los web logs del sitio web y en consideración de los inconvenientes mencionados anteriormente, el proceso no esta libre de errores [26]. Por esto han sido propuestos el uso de esquemas invasivos como el envío de otra aplicación al browser para capturar el comportamiento de navegación exacto del usuario [3, 10]. Si embargo, este esquema podría ser fácilmente evitado por el visitante.

La reconstrucción de sesiones apunta encontrar sesiones de usuarios reales, es decir, cuales páginas fueron visitadas por un ser humano.

En el análisis del sitio web, el escenario general es que usualmente no implementan mecanismos de identificación. La utilización de estrategias reactivas puede llegar a ser más útil. Estas pueden ser clasificadas en dos grupos principales [4, 10]: Heurísticas orientadas a la navegación y Heurísticas Orientadas al tiempo. Esta última es la utilizada en este trabajo y considera un tiempo máximo de duración de la sesión.

3.2. Procesando el contenido textual de una página web.

El modelo de vector espacial [24], permite una representación vectorial simple de las páginas web y. mediante el uso de comparación de distancia entre vectores, provee de una medida de las diferencias y similaridades entre páginas. Las páginas web deben ser limpiadas antes de transformarlas en vectores, tanto para reducir el número de palabras – no todas las palabras tienen el mismo peso – y hacer el proceso más eficiente.

Para el propósito de representación vectorial, sea R el número total de palabras diferentes y Q el número de páginas en el sitio web. Una representación vectorial del conjunto de páginas es una matriz M de tamaño R x Q.

$$M = (m_{ij}), i = 1,..., R \text{ y } j = 1,..., Q \qquad (1)$$

Donde m_{ij} es el peso de la palabra i en la página j. Basado en *tfxidf-weighting* introducido en [24] los pesos son estimados como:

$$m_{ij} = f_{ij}(1 + sw(i)) * \log(Q/n_i) \quad (2)$$

Aquí, f_{ij} es el número de ocurrencias de la palabra i en la página j y n_i es el número total de veces que la palabra i aparece en el sitio web completo. Adicionalmente, la importancia de las palabras es incrementada por la identificación de palabras especiales. La importancia de palabras especiales es almacenada en un arreglo *sw* de dimensión R, donde *sw(i)* representa un peso adicional para la i-ésima palabra.

El arreglo *sw* permite al modelo de vector espacial incluir ideas acerca de información semántica contenida en el texto de la página web por la identificación de palabras especiales.

Las fuentes comunes de palabras especiales son: E-Mails: el texto enviado es una fuente para identificar las palabras más recurrentes; palabras destacadas, palabras con etiquetas especiales; palabras de consultas: palabras usadas por el usuario en el motor de búsqueda y que esta contenida en el sitio web; sitios web relacionados: palabras de páginas de sitios web que pertenecen a otros sitios en el mismo mercado.

En la representación vectorial, cada columna de la matriz M es una página web. Por ejemplo, la k$_{ésima}$ columna m_{ik} con $i = 1,\ldots,R$ es la k-ésima página en el grupo completo de páginas.

Definición 1 (Vector de Palabras por página) *es un vector* $WP^k = (wp_1^k,..., wp_R^k) = (m_{1k},..., m_{Rk})$, *k = 1,..., Q, es la representación vectorial de la k-ésima página en el grupo de páginas bajo análisis.*

Con las páginas web en representación vectorial, es posible utilizar la medida de distancia para comparar los contenidos de texto. La distancia común es el coseno del ángulo calculado como:

$$dp(WP^i, WP^j) = \sum_{k=1}^{R} wp_k^i wp_k^j \Big/ \left(\sqrt{\sum_{k=1}^{R} (wp_k^i)^2} \sqrt{\sum_{k=1}^{R} (wp_k^j)^2} \right) \qquad (3)$$

La ecuación (3) permite comparar el contenido de dos páginas web, retornando un valor numérico entre [0, 1]. Cuando las páginas son totalmente diferentes, *dp = 0*, y cuando son las mismas, *dp = 1*.

4. Extrayendo las preferencias de contenido del usuario de las páginas web.

La investigación se concentra en predicciones acerca de cuales páginas el usuario visitará y la información que esta buscando.

4.1. Modelando el comportamiento del usuario web.

El siguiente paso es examinar las preferencias del usuario, definido como el contenido preferido de la página web por el usuario. El comportamiento del usuario es caracterizado por: Secuencia de páginas, Contenido de la página y Tiempo gastado.

Definición 2 (Vector de comportamiento del usuario (UBV)) *Es un vector* $\upsilon = \left[(p_1, t_1)...(p_n, t_n)\right]$, *donde* (p_i, t_i) *son los parámetros que representan la i-ésima página del visitante y el tiempo gastado en ella en la sesión, respectivamente. En esta expresión,* p_i *es el identificador de la página.*

4.2. Analizando las preferencias de texto de los usuarios.

Las preferencias del contenido web del usuario son identificadas por la comparación de contenido de las páginas visitadas, [34, 33, 35] por la aplicación del modelo de espacio vectorial a las páginas web, con la variante propuesta en la sección 3.2, ecuación (2).

Desde el vector de comportamiento del usuario (UBV), las páginas más importantes son seleccionadas asumiendo que el grado de importancia esta correlacionado al porcentaje de tiempo gastado en cada página. El UBV se ordena de acuerdo al porcentaje de tiempo total gastado en cada página. Las t página más importantes, es decir, las primeras t páginas, son seleccionadas.

Definición 3 (Vector de Páginas Importantes (IPV)). *Es un vector* $\vartheta_i(\upsilon) = [(\rho_1, \tau_1),...,(\rho_i, \tau_i)]$, *donde* (ρ_i, τ_i) *es el componente que representa la i-ésima página más importante y el porcentaje de tiempo gastado en ella por la sesión.*

Sean α y β don UBVs. La medida de similaridad propuesta entre los dos IPVs es introducida en la ecuación 4 como:

$$st(\vartheta_t(\alpha), \vartheta_t(\beta)) = \frac{1}{t}\sum_{k=1}^{t} \min\left\{ \tau_k^{\alpha} \Big/ \tau_k^{\beta}, \tau_k^{\beta} \Big/ \tau_k^{\alpha} \right\} * dp(\rho_k^{\alpha}, \rho_k^{\beta}) \quad (4)$$

El primer elemento en (4) indica el interés del usuario en las páginas visitadas. Si el porcentaje de tiempo gastado por los usuarios α y β en la k-ésima página visitada es cercano a la otra, el valor de la expresión $\min\{\cdot,\cdot\}$ será cercano a 1. En el caso opuesto, será cercano a 0. El segundo elemento en (4) es *dp*, la distancia entre páginas representada en forma vectorial, introducida en (3). En (4) el contenido de las páginas más importantes es multiplicado por el porcentaje de tiempo total gastado en cada página. Esto permite a las páginas con contenidos similares ser distinguidas por diferentes intereses de usuarios.

4.3. Identificando palabras claves del sitio web.

Una palabra clave de un sitio web (o web site keyword) es definida como *"una palabra o posiblemente un grupo de palabras que hacen de una página web más atractiva para un usuario eventual durante su visita al sitio web"* [32].

El procedimiento de identificación de palabras claves del sitio web es aplicar una medida, descrita en la ecuación (5) (miembro geométrico) para calcular la importancia de cada palabra

$$kw[i] = \sqrt{\prod_{p \in \varsigma} m_{ip}} \quad (5)$$

Donde *i = 1,…,R, kw* es un arreglo que contiene los pesos para cada palabra relativa a un cluster dado y ς el grupo de páginas representando el cluster. Las palabras claves del sitio web son el resultado del ordenamiento de *kw* y de la detección de palabras con los pesos más altos, por ejemplo, las 10 palabras con mayor peso.

5. Extrayendo patrones de la data originada en un sitio web real.

Para propósitos experimentales, el sitio web seleccionado debería ser complejo con respecto a varias características: número de visitas, actualización periódica y ser rico en contenido textual. La página web de un banco virtual Chileno (sin sucursales físicas, todas las transacciones realizadas electrónicamente) cumplen con estos criterios.

Las principales características del sitio web del banco son: sitio en Español, con 217 páginas web estáticas y aproximadamente ocho millones de filas en los web log para un periodo de estudio entre Enero y Marzo del 2003.

5.1 Proceso de reconstrucción de sesiones.

Durante el proceso de reconstrucción de sesiones, se aplican filtros a los registros del sitio web. En este caso particular, solo se utilizan registros de requerimiento de páginas web para analizar el comportamiento específico del usuario en el sitio. También es importante la limpieza de sesiones anormales, por ejemplo, web crawlers. Las filas información que apunta a otros objetos como imágenes, sonidos, etc, son limpiadas.

5.2 Preprocesamiento del contenido de una página web.

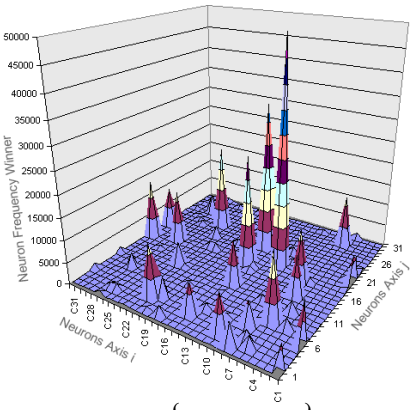
Mediante la aplicación de filtros a los textos de las páginas web, se ha encontrado que en el sitio completo contiene R=2034 palabras diferentes para ser utilizadas en el análisis.

Después de la identificación de palabras especiales y sus respectivos pesos, es posible calcular el peso final para cada palabra en la totalidad del sitio web, por la aplicación de la ecuación (2). Luego, se obtiene la representación vectorial para todas las páginas del sitio.

5.3 Analizando las preferencias de texto del usuario.

Se ha fijado en 3 el número máximo de dimensiones del vector. Luego, un SOFM con 3 neuronas de entrada y 32 neuronas de salida fue utilizado para encontrar los clusters de vectores de páginas importantes.

La figura 2 muestra las posiciones de las neuronas en el SOFM en los ejes x e y. El eje z es la frecuencia normalizada de veces que una neurona gana durante el entrenamiento donde se detectaron 8 clusters principales que contienen información acerca de las páginas más importantes del sitio web. Sin embargo, sólo 5 fueron aceptadas. El criterio de aceptación / rechazo es simple; si las páginas de un centroide de cluster tienen el mismo tema principal, entones el cluster es aceptado, de otra forma se rechaza.



Las palabras clave y su importancia relativa en cada cluster son obtenidas por la aplicación de la ecuación (5). Por ejemplo, si el cluster es $\varsigma = \{16,159,173\}$, entonces $kw[i] = \sqrt[3]{m_{i16}, m_{i159} m_{i173}}$, con $i = 1,..., R$.

Finalmente, ordenando las *kw* de forma descendente, podemos seleccionar las *k* palabras más importantes para cada cluster, por ejemplo *k = 5*. La tabla 2 muestra un grupo seleccionado de palabras clave de todos los clusters.

Tabla 2. Parte de las palabras descubiertas.

#	Palabra clave	
1	*Cuenta*	*Account*
2	*Fondo*	*Fund*
3	*Inversión*	*Investment*
4	*Tarjeta*	*Credit Card*
5	*Hipotecario*	*House credit*

La recomendación específica es utilizar las palabras clave como “palabras para escribir” en un sitio web, es decir, los párrafos escritos en la página deberían incluir algunas palabras claves y algunas podrían ser un enlace a otras páginas. Las palabras clave no funcionan de forma separada sino que requieren de un contexto que las utilice.

5.4 Mejorando el contenido textual el sitio web.

En cuanto cada página contiene un contenido de texto específico, es posible asociar las palabras claves de un sitio web a un contenido de la página; y desde esta sugerir la revisión o reconstrucción de un nuevo contenido en el sitio web. Por ejemplo, si la nueva versión de la página es relacionada con “tarjetas de crédito”, entonces las palabras claves del sitio web “crédito, puntos y promociones” deben ser asignadas para la reescritura del contenido textual de la página.

6. Conclusiones

Cuando el usuario visita un sitio web, hay una correlación entre el máximo de tiempo gastado por sesión en una página y su contenido de texto libre. Luego creamos el “Vector de Páginas Importantes (IPV)”, el cual es la estructura de datos básica de almacenamiento de páginas donde el usuario gasta más tiempo durante la duración de su sesión. Mediante la utilización de IPV como input en un SOFM, podemos identificar clusters que contienen la navegación del usuario e información de las preferencias.

El texto contenido en las páginas web puede ser mejorado utilizando las palabras claves del sitio web, y por esta vía atraer la atención del usuario cuando están visitando un sitio web.

Sin embargo, es necesario recordar que estas palabras no pueden ser utilizadas de forma individual, de hecho necesitan de un contexto, el cual es provisto por un ser humano.

Referencias

[1] E. Amitay and C. Paris. Automatically summarizing web sites: Is there any way around it? In *Procs. of the 9th Int. Conf. on Information and Knowledge Management*, pages 173–179, McLean, Virginia, USA, 2000.

[2] R. Baeza-Yates. *Web usage mining in search engines*, chapter Web Mining: Applications and Techniques, pages 307–321. Idea Group, 2004.

[3] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In *Proc. in First Int. Semantic Web Conference*, pages 264–278, 2002.

[4] B. Berendt and M. Spiliopoulou. Analysis of navigation behavior in web sites integrating multiple information systems. *The VLDB Journal*, 9:56–75, 2001.

[5] D. Buttler. A short survey of document structure similarity algorithms. In *Procs. Int. Conf. on Internet Computing*, pages 3–9, 2004.

[6] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Focused web searching with pdas. *Computer Networks*, 33(1- 6):213–230, June 2000.

[7] L. D. Catledge and J. E. Pitkow. Characterizing browsing behaviors on the world wide web. *Computers Networks and ISDN System*, 27:1065–1073, 1995.

[8] G. Chang, M. Healey, J. McHugh, and J. Wang. *Mining the World Wide Web*. Kluwer Academic Publishers, 2003.

[9] W. Chuang and J. Yang. Extracting sentence segment for text summarization? a machine learning approach. In *Procs. Int. Conf. ACM SIGIR*, pages 152–159, Athens, Greece, 2000.

[10] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowlegde and Information Systems*, 1:5–32, 1999.

[11] U. Hahn and I. Mani. The challenges of automatic summarization. *IEEE Computer*, 33(11):29–36, 2000.

[12] A. Joshi and R. Krishnapuram. On mining web access logs. In *Proc. of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 63– 69, 2000.

[13] A. P. Jr and N. Ziviani. Retrieving similar documents from the web. *Journal of Web Engineering*, 2(4):247–261, 2004.

[14] D. Lawrie, B. W. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *Proc. 24th Int SIGIR Conf. on Research and Development in Information Retrieval*, pages 349–357, New Orleans, Louisiana, USA, 2001. ACM Press.

[15] E. Liddy, K. McVearry, W. Paik, E. Yu, and M. McKenna. Development, implementation and testing of a discourse model for newspaper texts. In *Procs. Int. Conf. on ARPA Workshop on Human Language Technology*, pages 159–164, Princeton, NJ, USA, 1993.

[16] G. Linoff and M. Berry. *Mining the Web*. Jon Wiley & Sons, New York, 2001.

[17] S. Loh, L. Wives, and J. P. M. de Oliveira. Concept based knowledge discovery in texts extracted from the web. *SIGKDD Explorations*, 2(1):29–39, 2000.

[18] I. Mani and M. Maybury. *Advances in automatic text summarization*. MIT Press, Cambridge, Mass., 1999.

[19] S. Mitra, S. K. Pal, and P. Mitra. Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 13(1):3–14, 2002.

[20] B.Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In *Procs.Int Conf IEEE Knowledge and Data Engineering Exchange*, November 1999.

[21] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[22] M. F. Porter. An algorithm for suffix stripping. *Program; automated library and information systems*, 14(3):130–137, 1980.

[23] T. A. Runkler and J. Bezdek. Web mining with relational clustering. *International Journal of Approximate Reasoning*, 32(2-3):217–236, Feb 2003.

[24] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM archive*, 18(11):613–620, November 1975.

[25] M. Spiliopoulou. Data mining for the web. In *Principles of Data Mining and Knowledge Discovery*, pages 588–589, 1999.

[26] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, 15:171–190, 2003.

[27] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[28] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. Recovering traceability links in multilingual web sites. In *Procs. Int Conf. Web Site Evolution*, pages 14–21. IEEE Press, 2001.

[29] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. Restructuring multilingual web sites. In *Procs. Int. Conf. Software Maintenance*, pages 290–299. IEEE Press, 2002.

[30] J. D. Velázquez and V. Palade. A knowledge base for the maintenance of knowledge extracted from web data. *Journal of Knowledge Based Systems (Elsevier)*, page to appear, 2007.

[31] J. D. Velásquez, S. R´ıos, A. Bassi, H. Yasuda, and T. Aoki. Towards the identification of keywords in the web site text content: A methodological approach. *International Journal of Web Information Systems*, 1(1):11–15, March 2005.

[32] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. A methodology to find web site keywords. In *Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, pages 285–292, Taipei, Taiwan, March 2004.

[33] J. D. Velásquez, H. Yasuda, and T. Aoki. Combining the web content and usage mining to understand the visitor behavior in a web site. In *Procs.* 3th *IEEE Int. Conf. on Data Mining*, pages 669–672, Melbourne, Florida, USA, November 2003.

[34] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. Using the kdd process to support the web site reconfiguration. In *Procs. IEEE/WIC Int. Conf. on Web Intelligence*, pages 511–515, Halifax, Canada, October 2003.

[35] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396, February 2004.

[36] J. Xiao, Y. Zhang, X. Jia, and T. Li. Measuring similarity of interests for clustering web-users. In *ADC '01: Proceedings of the 12th Australasian conference on Database technologies*, pages 107–114, Washington, DC, USA, 2001. IEEE Computer Society.

[37] K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Procs. Int. Conf. on Computational Linguistics*, pages 986–989, 1996.