Web Mining: Análisis sobre la privacidad del tratamiento de datos originados en la Web

Juan D. Velásquez*
Lorena Donoso**

Resumen

Web mining es el concepto que agrupa a todas las técnicas, métodos y algoritmos utilizados para extraer información y conocimiento desde los datos originados en la Web (web data). Parte de estas técnicas apuntan a analizar el comportamiento de los usuarios, con miras a mejorar continuamente la estructura y contenido de los sitios que son visitados. El desarrollo histórico de los sitios web, se puede dividir en tres instantes claves respecto de cómo se presenta el contenido a los usuarios: estático, dinámico y adaptivo. El primero corresponde al origen de la Web, con sitios cuyo contenido era fundamentalmente textual. Luego se dio paso a sitios que incorporaron dinamismo en sus páginas para entrar en estos días en lo que se ha denominado la Web adaptiva o personalizada, es decir, que los contenidos y estructura del sitio se muestran dependiendo del tipo de usuario que lo visita. Detrás de tan altruista idea, es decir, ayudar al usuario a que se sienta lo mejor atendido posible por el sitio web, subyacen una serie de metodologías para el procesamiento de datos, cuya operación es al menos cuestionable, desde el punto de vista de la privacidad de los usuarios de un sitio web determinado. Entonces surge la pregunta ¿hasta donde el deseo por mejorar continuamente lo que se ofrece a través de un sitio web puede vulnerar la privacidad de quien lo visita?. El uso de herramientas de procesamiento de datos poderosas como las que contempla el web mining, puede atentar directamente contra la privacidad de los actos de los usuarios de un sitio web. En este trabajo se analizará el problema del procesamiento de los web data, para concluir con cuáles serían las técnicas que vulneran la privacidad de los

*Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

**Centro de Derecho Informático, Facultad de Derecho, Universidad de Chile, Santiago, Chile.

usuarios que visitan un sitio y cuales no. A la luz de esta información, se podrá orientar mejor a los profesionales de las TICs para aunar esfuerzos en la mejora de la estructura y contenidos de la Web, pero sin vulnerar la privacidad de las personas.

Palabras Clave: *????*, *???*, ???.

1. Introducción

Desde los orígenes de la Web, la creación de un sitio no ha sido un proceso fácil. Muchas veces se requiere de un equipo multidisciplinario de profesionales avocados a una sola misión "asegurar que el contenido y la estructura del sitio le son atractivos al usuario". Lo anterior es la clave del éxito para obtener una adecuada participación en el mercado electrónico, mantener la vigencia del sitio y sobre todo, lograr la tan ansiada y difícil fidelización del cliente digital [9].

La personalización implica que de alguna forma se puede obtener información respecto de los deseos y necesidades de las personas, para luego preparar la oferta correcta en el momento correcto [5]. Lo anterior plantea la necesidad de efectuar estudios previos para analizar la respuesta del consumidor ante un determinado estimulo, por ejemplo, los muy utilizados "focus group", donde un grupo de personas, que son la muestra representativa de un conjunto mayor, entrega su opinión respecto de lo que percibe en un producto o servicio.

Pensando en una esquema como el anterior, tal vez la solución para entender mejor al cliente digital sería someterlo a varias encuestas de opinión vía e-mail o al llenando formularios electrónicos. Sin embargo, la práctica ha demostrado que los usuarios no gustan de llenar formularios, contestar e-mails con preguntas, etc., a menos que se trate de algún amigo o familiar que quiera ayudar en el análisis, lo cuál no sería un caso real.

Cualquier análisis serio que se pretenda hacer respecto del comportamiento de navegación y preferencias que tiene un usuario en la Web, requiere del uso de datos reales, originados por usuarios reales. La pregunta entonces es: "¿de dónde saldrán estos datos?" . La respuesta es simple: de la misma Web. Ahora el cómo extraer estos datos y procesarlos para obtener un nuevo conocimiento acerca de los usuarios, es el gran desafío detrás de la personalización de la Web [2].

¿Hasta dónde este afán por analizar al usuario en la Web no se transforma en una persecución?. Con los web data adecuados, se puede hacer un completo seguimiento a todas las actividades de los usuarios en la Web, es decir, invadir directamente su privacidad, sin que estos se den cuenta de que están siendo vigilados por un "gran hermano" cibernético [4]. Evidentemente, la tecnología tiene dos caras, una de ellas muy siniestra y que la historia ha demostrado que si no se le regula adecuadamente, se pueden caer en excesos que atentan contra los derechos fundamentales de las personas [14].

2. Naturaleza de los datos originados en la Web

La Figura 1 muestra en forma simple el funcionamiento de la Web. El servidor web o web server (1) es un aplicación que está en ejecución continua, atendiendo requerimientos (4) de objetos web, es decir, el conjunto de archivos que conforman el web site (3) y enviándoselos (2) a la aplicación que hace la solicitud, generalmente un web browser (6). En general estos archivos son imágenes, sonidos, películas páginas web que conforman la información visible del sitio. Las páginas están escritas en Hyper Text Markup Language (HTML), que en síntesis es un conjunto de instrucciones, también conocidas como "tags" (5), acerca de cómo desplegar objetos en el browser o dirigirse a otra página web (hyperlinks). Estas instrucciones son interpretadas por el browser, el cual muestra los objetos en la pantalla del usuario [1].

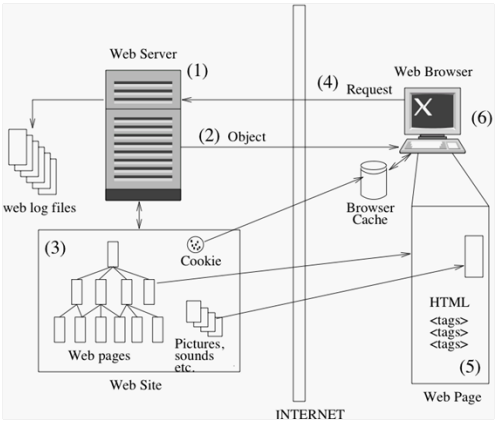

Figura 1: Modelo básico de operación de la Web

Cada uno de los tags presentes en una página, son interpretados por el browser. Algunos de estos tags hacen referencia a otros objetos en el web site, lo que genera una nueva petición en el browser y la posterior respuesta del server. En consecuencia, cuando el usuario digita la página que desea ver, el browser, por interpretación secuencial de cada uno de los tags, se encarga de hacer los requerimientos necesarios que permiten "bajar" el contenido de la página al computador del usuario.

La interacción anterior, ha quedado registrada en archivos conocidos como "web log files" [2], con lo cual es posible saber aproximadamente qué objetos fueron requeridos por un usuario, reconstruir su sesión y en la práctica realizar un verdadero seguimiento a sus actividades de navegación, analizando los contenidos visitados, el tiempo que se ha invertido en ello, qué información no atrae su interés, etc. La Figura 2 muestra un ejemplo del contenido y estruc-

tura de un archivo de web log, comenzando por la dirección IP del visitante del sitio, los parámetros ID y Authority (A), que son una forma de autenticar al usuario, siempre y cuando se especifique esa opción en el sitio web; la fecha y hora de conexión (Time), el método de obtención de la página, estatus de la petición, datos transferidos, de qué página procede el usuario (Referer) y finamente el tipo de software utilizado para navegar (Firefox, Mozilla, Explorer, etc.).

#	IP	Id	A	Time	Method/URL/Protocol	Statu	Byte	Referer	Agent
1	164.77.129.50	-	-	12/Apr/2003:23:47:44	GET /img/tab.gif HTTP/1.1	200	89	http://www.thebank.cl	MSIE 6.0; Windows 98
2	200.28.206.200	-	20	12/Apr/2003:23:48:31	GET transa/info.htm HTTP/1.1	200	144	/infoeco/info.html	MSIE 4.01; Windows 95
3	200.86.248.170	-	-	12/Apr/2003:23:48:37	GET /img/gen.gif HTTP/1.1	304	0	/ofert/wines/	MSIE 6.0; Windows 98
4	66.249.65.97	-	-	12/Apr/2003:23:48:41	GET /index.htm HTTP/1.1	200	88	-	Googlebot/2.1; google.com/bot.html
5	216.241.8.179	-	31	12/Apr/2003:23:50:03	GET /tx/infoeco/card.htm HTTP/1.1	200	210	/tx/infoeco/prom/	MSIE 6.0; Windows NT 5.1
6	164.77.129.50	-	-	12/Apr/2003:23:48:34	GET /tx/infoeco/ HTTP/1.1	200	186	/tx/infoeco/card.htm	MSIE 6.0; Windows 98
7	200.28.206.200	-	20	12/Apr/2003:23:51:13	GET transa/account.htm HTTP/1.1	200	180	/transa/info.htm	MSIE 4.01; Windows 95
8	216.241.8.179	-	31	12/Apr/2003:23:51:23	GET /tx/infoeco/ind.htm HTTP/1.1	200	300	/tx/infoeco/card.htm	MSIE 6.0; Windows NT 5.1
9	200.86.248.170	-	-	12/Apr/2003:23:51:41	GET /prom/wine.html HTTP/1.1	404	0	/ofert/wines/	MSIE 6.0; Windows 98
10	164.77.129.50	-	44	12/Apr/2003:23:52:04	GET /tx/infoeco/ind.htm HTTP/1.1	200	186	/tx/infoeco/	MSIE 6.0; Windows 98

Figura 2: Parte de un web log file

Como se puede apreciar, cada registro da cuenta de los movimientos de un usuario en un sitio web. En consecuencia, y en forma casi anónima, los datos generados en el sitio web son tal vez la mayor encuesta que podría tener una empresa por sobre sus eventuales clientes, analizando sus preferencias de información, las cuales están directamente relacionadas con las características de los productos y servicios ofrecidos.

El proceso anterior, no está exento de desafíos, siendo el primero de ellos la preparación de los datos para un proceso de extracción de información. En efecto, los web data, como se les conoce, consideran todos los tipos de datos existentes, lo cual dificulta su procesamiento. Adicionalmente, no siempre contienen datos relevantes e incluso algunos de ellos son más bien ruido, por lo cual se requiere de su pre-procesamiento y limpieza antes de que la información salga a la luz. El procesamiento considera la reconstrucción de la sesión del usuario, la limpieza del contenido de las páginas web, para la identificación de elementos relevantes (textos clave, imágenes, sonidos, etc.,) y en general la transformación de los web data en vectores de características que modelen el comportamiento del usuario en un sitio web particular.

3. Limpieza y preprocesamiento de los web data

Los datos originados en la Web o web data, corresponden esencialmente a tres fuentes [2]:

- Contenido: Son los objetos que aparecen dentro de una página web, por ejemplo las imágenes, los textos libres, sonidos, etc.
- Estructura: Se refiere a la estructura de hipervínculos presentes en una página.

- Uso: Son los registros de web logs, que contienen toda la interacción entre los usuarios y el sitio web.

Algunos autores argumentan que también se debieran considerar los datos de "perfil de usuario" como parte de los web data [2] y [3]. Se trata de datos personales como el nombre, edad, sexo, etc., los cuales en rigor no deben ser tratados en un proyecto web mining, correspondiendo su procesamiento al uso de otras herramientas (data mining), por lo que no se les considerará en el presente trabajo.

Los web data deben ser pre procesados antes de entrar en un proceso de web mining, es decir, son transformados en "vectores de características" que almacenan la información intrínseca que hay dentro de ellos [3] y [10].

Aunque todos los web data son importantes, especial atención reciben los web logs, ya que ahí se encuentra almacenada la interacción usuario sitio web, sus preferencias de contenido y en síntesis su comportamiento en el sitio. Por esta razón, y concediendo de que es posible que en los otros web data se pueda albergar información que identifique a los usuarios, nos concentraremos esencialmente en los web logs, como fuente de mayor controversia al momento de analizar el comportamiento de los usuarios.

La primera etapa, entonces, corresponde a la reconstrucción de la sesión del usuario a partir de los datos existentes en los registros de web log. Este proceso se denomina sesionización y puede resumirse en los siguientes pasos [2]:

1. Limpiar los registros de los web logs, dejando sólo aquellos concernientes a peticiones de páginas web, eliminando los que indican peticiones de objetos contenidos en éstas

2. Identificar registros de peticiones hechas por web crawlers. Para ello, existen listas oficiales y no oficiales de crawlers en la Web. Pueden ser identificados a través del campo Agent o, en su defecto, por la IP Address.

3. Agrupar los registros por IP Address y por Agent. Cabe señalar que se está asumiendo que no hay más datos acerca de los usuarios que visitan el sitio.

4. Ordenar los registros de menor a mayor Timestamp, de modo que los registros aparezcan cronológicamente.

5. Identificar las sesiones de navegación, para lo cual existen dos alternativas:

   a) Utilizar un criterio estadístico, es decir, asumir que por lo general las sesiones reales de los usuarios no duran más allá de 30 minutos.

   b) Asumir que ninguna sesión tiene páginas visitadas más de una vez.

6. Por último, reconstruir la sesión usando la estructura de hipervínculos del sitio y completando aquellas páginas que no fueron registradas debido a que se utilizó la memoria caché del web browser o la caché corporativa de un servidor proxy.

Existen dos estrategias a seguir para realizar el proceso de sesionización [2],[3] y [9]:

1. **Estrategia Proactiva**: Consiste en utilizar herramientas invasivas para identificar a los usuarios, generalmente se trata de cookies[1] o spybots[2], las cuales permiten identificar al usuario con un identificador único, de modo que es posible conocer su frecuencia de visitas y cómo ha variado su comportamiento en el tiempo. Con estos datos se puede extraer información muy valiosa, sin embargo, de acuerdo las legislaciones de ciertos países y comunidades como la Unión Europea, la utilización de estas herramientas atenta contra la privacidad de las personas y su uso es condenado. Por otro lado, existen programas especializados en borrar spybots y la mayoría de los navegadores puede eliminar las cookies y no permitir su funcionamiento, por lo que estos métodos están perdiendo su efectividad.

2. **Estrategia Reactiva**: Consiste en la utilización de los web logs como fuente única de datos para la reconstrucción de sesiones, evitando atentar contra la privacidad de los usuarios. Si bien es la estrategia que provee una riqueza potencial de información menor, pues no se puede identificar al usuario y su frecuencia de visitas, es la que puede ser seguida en cualquier sitio.

Cabe señalar que ciertos sitios han cambiado su estructura con el propósito de identificar a sus visitantes [2],[9] y [12]. Una primera estrategia consiste en implementar un sistema username/password, que promueva el registro de los usuarios a cambio de nuevos servicios. Sin embargo, sólo es posible reconstruir perfectamente las sesiones de los registrados, quedando los no registrados en el anonimato. Otra estrategia consiste en utilizar páginas dinámicas en el sitio. Con ellas, cada solicitud de abrir una página genera un identificador único para el usuario, sin embargo, ello obliga a reconstruir el sitio y trae complejidades para identificar qué está realmente viendo el visitante, dadas las direcciones URL dinámicamente generadas [3].

4. Web mining

El concepto web mining, agrupa a todas las técnicas, algoritmos y metodologías utilizadas para extraer información y conocimiento desde los web data.

[1]Fragmento de información que se almacena en el disco duro del visitante de una página web a través de su web browser, a petición del web server de la página.

[2]Programas que entregan directamente la secuencia de visitas realizadas por el usuario en toda la Web.

En este sentido, se podría decir que es la aplicación de teoría del data mining al caso particular que revisten los web data [7].

Web Mining ha permitido estudiar la creciente cantidad de datos disponibles, donde la estadística clásica y la revisión manual ya resultan ineficientes [5]. La importancia de tener datos limpios y consolidados radica en que la calidad y utilidad de los patrones encontrados por estas herramientas dependerán directamente de los datos que sean utilizados. Por este hecho es que muchas veces los procesos de web mining dan lugar a información errónea, pues a diferencia de las herramientas estadísticas, existen herramientas a disposición de cualquier usuario con poco conocimiento de este campo.

Dentro de las técnicas de Data Mining más utilizadas se puede mencionar [7]:

1. *Reglas de asociación*: Consiste en encontrar correlaciones entre conjuntos de datos bajo una probabilidad de ocurrencia (confianza) para una proporción del total de datos (soporte). Por ejemplo: pan queso (soporte = 5 %, confianza = 42 %) quiere decir que 42 % de las personas que compraron pan también compraron queso y que esta combinación se dio en 5 % de las transacciones. Una extensión a ésta es la asociación multidimensional, dónde se asocia más de un atributo a otro. Ejemplo: pan, mantequilla y queso.

2. *Clasificación*: Consiste en clasificar una serie de registros en alguna categoría previamente definida. Para ello, se suele usar un proceso de aprendizaje, en el cual el algoritmo es aplicado sobre datos ya clasificados, de modo que pueda establecer bajo qué valores de los otros atributos del registro, se trata de una categoría u otra. Una vez realizado el aprendizaje, se procede a efectuar la clasificación sobre registros que no fueron utilizados como input en el aprendizaje.

   La efectividad de la clasificación es calculada comparando la categorización dada por el algoritmo versus la real que ya se tenía. Para ejemplificar esto, piense en un estudio para clasificar nuevos clientes en: aquellos sin deuda y aquellos con deuda de acuerdo a sus antecedentes personales. Para ello, se realiza un proceso de aprendizaje sobre un subconjunto de registros de los clientes ya clasificados, para posteriormente efectuar una clasificación sobre otro subconjunto, disjunto al anterior, para determinar la efectividad del algoritmo. Una vez verificada su efectividad, este algoritmo permitirá predecir qué clientes nuevos cumplen un perfil de deudores o no deudores, de acuerdo a sus antecedentes personales.

3. *Clustering*: Consiste en agrupar objetos que tienen características similares, a diferencia de la técnica anterior en el clustering no se conoce la categorización a priori y se espera encontrarla a partir de los datos.

Para ello se utiliza una medida de similitud entre registros, que permite separarlos de acuerdo a sus diferencias en esta medida. Existen principalmente 3 técnicas de clustering:

a) *Particionado*: bajo esta técnica se definen a priori los n clusters en los que se clasificarán los registros.

b) *Jerárquico*: esta técnica construye los clusters mediante una descomposición jerárquica que puede ser de dos tipos: Aglomerativa, en el que se parte con un cluster por cada registro y estos empiezan a agruparse de acuerdo a la medida de similitud utilizada hasta una condición terminal previamente definida y Divisiva, en el que se parte con un sólo cluster que incluye todos los registros y este empieza a separarse de acuerdo a la medida de similitud utilizada hasta una condición terminal previamente definida.

c) *Basado en densidad*: Esta técnica toma prestada la definición de densidad de la Física. Consiste en definir una densidad umbral, que no es más que una cardinalidad predefinida para cada cluster, y un radio, que no es más que una distancia predefinida, de modo que los clusters se van formando por registros a una distancia del centroide del cluster menor al radio, los centroides son redefinidos en cada iteración y el algoritmo para cuando todas las cardinalidades de los clusters encontrados son menores a la densidad umbral previamente definida.

5. Marco legal para el análisis de la privacidad en la web

La legislación internacional ha recogido una de las preocupaciones más relevantes de los constitucionalistas de la nueva sociedad, cual es el grado de afectación a la libertad y a la dignidad humana por el nivel de control personal que se puede alcanzar gracias al almacenamiento y tratamiento de datos personales. Cuando ello alcanza al control de los datos que hoy en día transitan a través de Internet la preocupación se hace más aguda, en tanto que a través de esta red podría llegar a conocerse prácticamente toda la información de la persona, incluso su ubicación y estado de salud o situación emocional, todo ello en tiempo real.

5.1. Antecedentes

Desde la perspectiva jurídica, en Chile y la mayoría de los países es susceptible de ser calificado como "dato personalçualquier información relativa a

personas naturales identificadas o susceptibles de ser identificadas, lo que dependerá del avance de la técnica y de la potencialidad identificativa del dato. En este contexto, los elementos que conforman el concepto de dato personal son los siguientes:

- **Toda información**. Se trata de un concepto amplio, que abarca tanto, imagen, sonido, muestras biológicas, incluso al conjunto de caracteres grafológicos que proporcionen antecedentes que coadyuven al conocimiento de un sujeto.

- **Relativos a una persona Natural**. Si bien hay legislaciones (como la de Argentina) que admiten la protección de datos personales respecto de las personas jurídicas, la generalidad de los países sólo admite la protección de datos personales de personas físicas. Ello principalmente porque hacen derivar la garantía desde los derechos fundamentales emanados de la dignidad humana. Ello no obstante hay otras instituciones que protegen la información que se refiere a personas jurídicas, básicamente reguladas en sede mercantil.

- **Identificada o identificable**. En el ámbito europeo, se ha dicho será identificable "Toda persona cuya identidad pueda determinarse directa o indirectamente, en particular mediante un número de identificación o uno o varios elementos específicos, característicos de su identidad física, fisiológica, psíquica, económica, cultural o social[3]". De su parte, en la legislación española se ha dicho al respecto que estaremos frente a una persona identificable cuando estemos en presencia de "cualquier elemento que permita determinar directa o indirectamente la identidad física, fisiológica, psíquica, económica, cultural o social de la persona física afectada[4]" . En consecuencia el concepto identificable debe adaptarse en cada momento a la realidad tecnológica y social, pues hoy en día existen mecanismos hábiles para la identificación del sujeto, a partir de diversas circunstancias, los que son accesibles al común de las personas y, por tanto, no imponen un esfuerzo exorbitante al responsable del tratamiento o, en su caso, al encargado del tratamiento de datos personales.

Los datos personales podrán ser públicos o sensibles[5], siendo estos últimos aquellos que gozan generalmente de un mayor nivel de protección. Si bien no podemos generalizar respecto de qué datos ocupan una u otra categoría, pues ello dependerá de las condiciones y usos sociales de cada país, lo que queda claro es que el factor que determina la sensibilidad es el riesgo de que el uso

[3] 2 letra a) Directiva 95/46 CE, del Parlamento y del Consejo, de 1995

[4] Así se ha reconocido desde el R.D. 1332/94, Artículo 1, inciso 5.

[5] En la nomenclatura europea, estos son "datos especialmente protegidos". Véase al respecto la Directiva 95/46 CE, del Parlamento y del Consejo, de 1995.

del dato acarree perjuicios a su titular, básicamente por las posibilidades de que terceras personas adopten decisiones arbitrarias a su respecto. De ahí que por regla general y a vía de ejemplo los datos sobre salud, los relativos a las ideologías políticas, al credo religioso y/o la vida sexual sean considerados como datos sensibles.

En este punto cabe preguntarse si los datos personales que se recogen con ocasión del uso de Internet susceptibles de ser calificados como públicos o como sensibles. En particular los interesan tres tipos de datos: a) los "contenidos web", esto es imágenes, texto, sonidos, etc., relativos a una persona; b) Datos de estructura. Esto es, hipervínculos que han sido accionados por la persona; y c) los registros de logs. La pregunta es relevante por cuanto la respuesta condicionará si es factible el tratamiento de datos personales o no, y en la afirmativa, si es necesario solicitar el consentimiento al afectado (titular de los datos personales) o basta con que se le informe sobre el tratamiento de datos personales.

Atendida la complejidad de esta pregunta, la abordaremos en un acápite especial, más adelante. Sólo nos permitiremos adelantar en este punto que la ley chilena considera que es dato sensible aquel que revela los hábitos de una persona. Asimismo, nos permitiremos adelantar que en prácticamente todos los ordenamientos jurídicos se considera que los datos personales de menores de edad habrán de ser tratados como datos sensibles, atendidas las normas de protección derivadas del derecho internacional, destinada a la protección de la infancia.

Otro aspecto conceptual que es importante considerar es que la legislación entiende que el **tratamiento de datos personales** comprende cualquier operación **manual o automatizada**, que se realice sobre los datos personales, desde la recogida y hasta la cancelación definitiva de los datos personales.

Finalmente dentro del contexto es importante destacar que la legislación entiende que el dueño del dato personal es siempre su titular, esto es, la persona a quien se refiere el dato, lo cual no se altera por el hecho de que el titular del banco o base de datos en la cual consta ese dato, le haya agregado valor o lo haya integrado con otros datos personales a través de reglas de negocio específicas y/o procesos tecnológicos de tratamiento de datos.

5.2. Los datos personales y la Web

Uno de los aspectos que se ha desarrollado, gracias a las potencialidades de la llamada Web 2.0, dice relación con la recogida y tratamiento de datos personales a través de la Web. Estas actividades llevan aparejados múltiples desafíos, tanto desde la óptica tecnológica como jurídica. En este último aspecto, el tratamiento de datos personales en/o a través de la web conlleva problemáticas de derecho internacional, en el sentido que podrán producirse

situaciones en las que los servidores se encuentren en un país diferente a aquel en que se encuentra o pertenece el sujeto fuente de datos. Es más, la situación puede ser más compleja en tanto que el titular del banco de datos puede asimismo estar en un país diverso al de localización del servidor. Esto es más evidente conforme avanzan los sistemas y procesos que adscriben al concepto de Cloud computing.

Para los efectos de responder a la interrogante sobre cuál es el estatuto jurídico de estas actividades de tratamiento de datos personales hemos atendido a lo regulado en la Unión Europea y en Estados Unidos. Ello principalmente porque se han pronunciado expresamente sobre estas materias, mientras en nuestro entorno el debate aún no se ha iniciado.

Al respecto es importante considerar que en la Unión Europea, el Grupo del Artículo 29 del tratado constitutivo de la Unión ha sostenido que si el sitio web o motor de búsquedas o recogidas de datos accede a datos de habitantes de Europa, debe aplicárseles la legislación europea de tratamiento de datos[6].

En Estados Unidos, la Children's Online Privacy Protection Act (COPPA) de 1998 extiende su ámbito de aplicación a los sitios web extranjeros que recogen información personal de niños menores de 13 años establecidos en el territorio de los Estados Unidos.

Siguiendo estos criterios, nos queda claro que la tendencia internacional es considerar que la legislación aplicable en esta materia es aquella que corresponde a la del titular de los datos en el momento de la recogida, lo que deberá resolverse de acuerdo a las reglas generales del Derecho internacional.

5.3. Proveedores de medios de conexión y Almacenamiento

Jurídicamente cada uno de los servidores contectados a la red de redes actúa como emisor/receptor de datos (contenidos) y se les denomina nodos de la red. Los servicios de transferencia de información a través de la red suponen la existencia de proveedores de medios de interconexión y de almacenamiento de información. Ambas categorías integran lo que se ha dado en llamar **ISP (Internet Service Providers)** o **PSI (Proveedores de Servicios Internet)**.

En este punto es importante destacar que el marco normativo nacional está contenido en la Resolución Exenta No. 1483, de 22 de octubre del año 1999, que "Fija Procedimiento y Plazo Para Establecer y Aceptar Conexiones Entre ISP". Como podemos apreciar del solo título de la norma ya se desprende que su ámbito de aplicación está restringido a la interconexión entre ISP concebidos en ella como personas naturales o jurídicas que actúan como

[6]Véase al respecto 5035/01/ES/Final WP 56, Documento de trabajo relativo a la aplicación internacional de la legislación comunitaria sobre protección de datos al tratamiento de los datos personales en Internet por sitios web establecidos fuera de la UE

Proveedores de Acceso a Internet[7], que es aquel que "permite acceder a la información y aplicaciones disponibles en la red Internet"[8].

Estos proveedores proporcionan a los usuarios **capacidad de almacenamiento de datos (Hosting), capacidad de conexión**, esto es aquella que permite hacer transitar la información desde su punto de origen hasta su destino, y que en definitiva son aquellos servicios asociados a las redes de telecomunicaciones. De su parte, tratándose de usuarios que no forman parte de una red integrante de Internet, la **posibilidad de conectarse** a la red, a través de un sistema de acceso conmutado sirviéndose de las instalaciones correspondientes a las líneas telefónicas locales, servicio telefónico inalámbrico o cable coaxial de televisión o cualquier otro medio técnicamente autorizado al efecto.

En consecuencia estos servicios podrán clasificarse como: **Proveedor de Acceso**, que permiten la conexión a la red y por tanto el funcionamiento del usuario como un nodo temporal o continuo; **Proveedores de enlace**: que son aquellos que proporcionan los mecanismos de transmisión y finalmente los **Proveedores de Hosting**, que, como su nombre lo indica dan el servicio de proporcionar un espacio en disco que permite almacenar la información, dejándola a disposición de todo aquel que quiera acceder a ella a través de la Red. Estos últimos en todo caso, se califican en general como suministradores de servicios suplementarios. El marco conceptual de este servicio se complementó con las Resolución Exenta No. 698, de 30 de junio del año 2000, de la Subsecretaría de Telecomunicaciones, en que se define una **Página WEB** como un "Documento editado en hipertexto (HTML), publicado, y que puede ser accesado en la red Internet". De su parte en la Resolución Exenta 669 de 2001, también de Subtel, según su texto refundido, fijado por Resolución Exenta 1493 de ese mismo año, conceptualiza en su artículo 1 letra a), se definió conexión conmutada y conexión dedicada diferenciándolas a partir de los medios empleados en la provisión del servicio, en los siguientes términos: a) Conexión conmutada: Forma de acceso a la red Internet donde la conexión es realizada por medio del uso de la Red Pública Telefónica, durante el tiempo que dure dicha conexión. Para estos efectos, se entenderá por Red Pública Telefónica aquella constituida de conformidad a lo dispuesto en el artículo 9º del decreto Supremo Nº425, de 1996, modificado por decreto supremo Nº697, de 2000, ambos del Ministerio de Transportes y Telecomunicaciones, Reglamento del Servicio Público Telefónico.

Como podemos apreciar esta norma no considera que la prestación del servicio de acceso a Internet ADSL pueda ser considerado como acceso conmutado, pues restringe la aplicación de esta condición a aquellas modalidades en que se emplea la RPTF durante la conexión. Esto queda más claro en el

[7]Resolución Exenta 1483, 1999, Subsecretaría de Telecomunicaciones, Artículo 1 letra c)
[8]Resolución Exenta 1483, 1999, Subsecretaría de Telecomunicaciones, Artículo 1 letra a)

artículo 4, relativo a los deberes de información de los indicadores de calidad por los ISP, como veremos más adelante.

En cambio la modalidad de **acceso dedicado**, se define como: una conexión a la red Internet efectuada a través de un enlace de comunicación permanente, que puede ser monousuario o multiusuario.

Teniendo en cuenta lo antes señalado, las escasas bases jurídicas con que se cuenta, debemos sostener que conforme al principio de **neutralidad tecnológica**, habrá de considerarle un servicio de telecomunicaciones[9] y en consecuencia seguir el principio de "aplicación de la misma regulación a los servicios de telecomunicaciones con independencia de la tecnología utilizada para la prestación de los mismos"[10].

El marco básico a aplicar será la ley 18.168, general de telecomunicaciones en lo que respecta al servicio propiamente tal y por la ley 19.628, de tratamiento de datos personales en lo que se refiere a lo datos personales que circulan por las redes con ocasión o gracias a este servicio y la normativa complementaria, por ejemplo en lo que se refiere a la obligación de captura y almacenamiento de datos personales para los efectos, por ejemplo, de investigación criminal. Para esta calificación nos basamos en que los medios empleados para realizar esta transmisión son las redes de telecomunicaciones, las que para estos efectos se intercomunican mediante el protocolo TCP/IP. Considerado así, las comunicaciones a través de las cuales opera Internet responden al concepto de Telecomunicación y en consecuencia la instalación, operación y explotación de los servicios de telecomunicaciones ubicados en el territorio nacional, incluidas las aguas y espacios aéreos sometidos a la jurisdicción nacional y, en lo que les sea aplicable, los sistemas e instalaciones que utilicen ondas electromagnéticas con fines distintos a los de telecomunicación[11], quedan bajo la Subsecretaría de Telecomunicaciones como el organismo técnico cuyas finalidades principales son la aplicación y control de la ley y sus reglamentos, la interpretación técnica de las disposiciones legales y reglamentarias que rigen las telecomunicaciones[12], sin perjuicio de las facultades que establece la legislación respecto de otros órganos administrativos o judiciales.

[9]El Artículo 1 de la ley 19.628 define servicio de telecomunicaciones en los siguientes términos: Para los efectos de esta ley, se entenderá por telecomunicaciones toda transmisión, emisión o recepción de signos, señales, escritos, imágenes, sonidos e informaciones de cualquier naturaleza, por línea física, radioelectricidad, medios ópticos u otros sistemas electromagnéticos.

[10]Comentarios de Aniel en Relación con la Comunicación de la Comisión Europea, COM (1999) 539, "Revisión 1999 del Sector de las comunicaciones" en línea en http://europa.eu.int/ISPO/infosoc/telecompolicy/review99/comments/aniel22b.htm [Consulta: 28 Ago. 2004]. En el mismo sentido Directiva 2002/20/CE del Parlamento Europeo y del Consejo, relativa a la autorización de redes y servicios de comunicaciones electrónicas, de 7 de marzo de 2002 (Directiva autorización)

[11]4 Ley 18168, incisos 1 y 2.

[12]Art. 6 incisos 1 y 2 ley 18168

Además la Subsecretaría, como órgano creado al efecto debe velar porque todos los servicios de telecomunicaciones y sistemas e instalaciones que generen ondas electromagnéticas, cualquiera sea su naturaleza sean instalados, operados y operados de modo que no causen lesiones a personas o daños a cosas, ni interferencias perjudiciales a los servicios de telecomunicaciones nacionales o extranjeros o interrupciones en su funcionamiento, además de controlar y supervigilar el funcionamiento de los servicios públicos de telecomunicaciones y la protección de los derechos de los usuarios[13], sin perjuicio de las acciones judiciales y administrativas a que éstos tengan derecho.

Como podemos apreciar de estas normas, en general la interpretación de la Ley, así como su aplicación práctica, ya sea al momento de constituirse una empresa como prestador de servicios de telecomunicaciones, al momento de entrar en operaciones y luego explotar el servicio corresponde a la Subsecretaría, quien puede intervenir en las relaciones con ocasión de la prestación de servicios de telecomunicaciones, se produzcan entre los prestadores de servicios y en aquellas que se susciten entre los prestadores y los usuarios de los servicios de telecomunicaciones.

Respecto de estos últimos, a la Subsecretaría le corresponde velar por la protección de los derechos de los suscriptores y usuarios en lo que se refiere al servicio de telecomunicaciones, sin perjuicio de las demás acciones judiciales o administrativas que les asistan en protección de sus derechos. En lo que nos interesa, la Subsecretaría debiera velar porque las personas que utilizan los servicios de telecomunicaciones, y en particular los servicios de acceso a Internet no se vean afectados ilegítimamente en lo que respecta a sus datos personales por actos de los ISP.

En efecto, si bien la estrategia normativa adoptada en este momento fue la **mínima intervención** a fin de propiciar el desarrollo de este mercado en nuestro país, la **mínima alteración de las categorías normativas vigentes** a efectos de garantizar la certeza en el conocimiento del marco jurídico por los operadores y que dado el avance del servicio en nuestro país a esa época, ello no ha significado abstraer al sector de la normativa sobre tratamiento de datos personales, sino todo lo contrario, la consolidación de una intervención desde la óptica de la mínima intervención coadyuva a propiciar un marco de seguridad jurídica que de certeza a los actores sociales respecto de los efectos jurídicos de sus actos y asimismo garantizar la eficacia y coherencia del ordenamiento jurídico en su conjunto, esto sin perjuicio de intervenir el sistema normativo para ajustar aquellos mínimos que sean necesarios para la operación de los servicios de la sociedad de la información, considerando que éstos normal y progresivamente serán servicios convergentes. Como señalamos, uno de los ejes esenciales de la regulación ha sido la necesidad de dar protección a los usuarios de Telecomunicaciones, quienes además de tener un acceso efi-

[13] Art. 7 incisos 1 y 2 ley 18168

ciente y equitativo a los servicios de telecomunicaciones, habrán de contar con las estrategias y mecanismos de protección suficientes al efecto de garantizar su derecho fundamental a comunicarse con pleno respeto a sus derechos como titulares de datos personales.

Este principio viene a plasmar normativamente una realidad, cual es el desequilibrio en que se encuentran los prestadores y los usuarios de los servicios de la sociedad de la información. Conforme a él la normativa de telecomunicaciones, considerada en forma integral, debe establecer mecanismos claros de protección y promoción a los derechos de los usuarios. Este es el espíritu del artículo 7 de la Ley General de Telecomunicaciones, cuando dispone que "El Ministerio de Transportes y Telecomunicaciones velará porque todos los servicios de telecomunicaciones y sistemas e instalaciones que generen ondas electromagnéticas, cualquiera sea su naturaleza, sean instalados, operados y explotados de modo que no causen lesiones a personas o daños a cosas ni interferencias perjudiciales a los servicios de telecomunicaciones nacionales o extranjeros o interrupciones en su funcionamiento. Además, le corresponderá controlar y supervigilar el funcionamiento de los servicios públicos de telecomunicaciones y la protección de los derechos del usuario, sin perjuicio de las acciones judiciales y administrativas a que éstos tengan derecho"[14].

En el contexto que nos interesa, los proveedores del servicio de acceso a Internet son responsables de la custodia de datos personales a los que tienen acceso y que son objeto de tratamiento con ocasión del servicio prestado. Estos serán principalmente aquellos relativos a las navegaciones, comunicaciones efectuadas y en general, los logs generados en el proceso comunicativo. Asimismo, en términos generales los PSI quedan expresamente exonerados de monitorizar la red o el servidor que administran, por entender que dicha labor es imposible ante el volumen de información existente.

Ahora bien, en cuanto al papel de los proveedores de medio en la cadena de responsabilidad por los contenidos, se ha reconocido que ellos "no controlan de manera directa el contenido disponible en Internet ni qué parte deciden consultar sus clientes"... puede que sea preciso cambiar o clarificar la legislación para ayudar a los suministradores de acceso y los suministradores de servicio de ordenador central, cuya ocupación primordial es prestar servicio al cliente, a abrir un camino que evite, por un lado las acusaciones de censura, y por otro, los riesgos de acciones judiciales"[15].

[14]Modificado como aparece por ley 19302, Art.1°, letra c.-

[15]Comunicación de la Comisión Europea al Parlamento Europeo, al Consejo, al Comité Económico y Social y al Comité de las Regiones"de fecha 16 de octubre de 1996, (COM(96)487), Acápite Nos. 1 y 2.

5.4. Regulación y web mining

La idea básica que subyace detrás del web mining es la extracción de información y conocimiento desde un conjunto de web data. Dependiendo del tipo de web data a minar, el algoritmo de web mining puede estar altamente relacionado con los datos personales del usuario. Lo anterior plantea muchas interrogantes, sobre todo en lo referente a la privacidad del usuario.

Partamos analizando el o los archivos de web log, en especial la dirección IP desde donde accedió el usuario al sitio web. Este parámetro en combinación con otros datos existentes en el registro de web log, ha sido frecuentemente utilizada para identificar la sesión del usuario. Debido a la posibilidad de que se pueda relacionar o identificar a la persona a través de la dirección IP que utiliza para navegar por la Web, es que en la UE se está comenzando a considerar a la IP como un dato personal[16]. En España, la Ley Orgánica 15/1999, en su artículo 3a define al dato personal como "cualquier información concerniente a personas físicas identificadas o identificables".

El TCP/IP versión 4, que es el protocolo con que en la actualidad opera Internet, fue concebido para identificar un computador conectado a la red. Hay que recordar que Internet (Interconection Network) es una "red de redes", así que para identificar un computador, primero se identifica a qué red pertenece. De esta forma, las direcciones IP están compuestas de cuatro números (rango entre 0 y 255 cada uno) con los que se identifica la red y el computador dentro de ésta.

Entonces, por construcción la dirección IP no fue creada para identificar a la persona detrás del computador. Mucho menos ahora que existen sistemas que permiten a varios usuarios acceder a Internet, usando la misma IP y que los ISP entregan direcciones dinámicas, es decir, sólo relacionan una sesión de usuario mientras que está conectado. Incluso más, es posible que durante la sesión, el usuario experimente cambios en la IP asignada. Sin embargo, si se realizan los cruces de datos adecuados, se puede llegar a una aproximación respecto de quien sería la persona que en un determinado momento, estaba conectada desde un computador, usando una IP específica. Si lo anterior es probatorio en un tribunal, es un tema que escapa a las pretensiones de este estudio, pero hay que dejar en claro de que, desde el punto de vista técnico, no habría una certeza del 100 % de que una persona accedió a un sitio desde una IP determinada.

El escenario anterior debería cambiar una vez que la nueva generación del protocolo TCP/IP, la versión 6, entre en total funcionamiento en Internet, por cuanto se podrán implementar otros recursos de identificación de la persona, por ejemplo transmisión de datos cripto-segurizados y con firma digital.

Asumiendo, entonces, que a través de la dirección IP sólo se puede identi-

[16] http://www.habeasdata.org/Ipcomodatopersonal

ficar la sesión y no a la persona que hay detrás, los algoritmos de web mining se orientan a extraer información desde los web data para analizar comportamientos de usuarios en determinados momentos del día, es decir, un mismo usuario se puede comportar diferente en momentos diferentes, con lo cual se argumenta que no se estaría analizando a la persona, sino más bien a grupos de personas para extrapolar comportamientos colectivos [7].

Como en todo aquello donde el ser humano no encuentra consenso, aparecen posturas divididas. Por una parte, los especialistas que aplican algoritmos de web mining para lograr un mejor entendimiento de las preferencias de los usuarios de un sitio web, asumen que todo el procesamiento de los web data es inocuo para el usuario. Por otro lado, si el usuario se entera de que todos sus movimientos en el sitio están siendo monitoreados, tal vez decide no visitar el sitio, porque siente que a través del uso de estas herramientas se produce una intromisión directa en su privacidad.

Ahora bien, ¿qué es la privacidad?. La RAE define el término como "ámbito de la vida privada que se tiene derecho a proteger de cualquier intromisión". Y en Internet, ¿este concepto tiene sentido?. En este trabajo no vamos a ahondar en el contexto filosófico de la privacidad, sino que se fijarán limites sólo en lo referente al control de la información respecto de uno mismo, es decir, la capacidad que tiene el individuo de proteger los datos que se refieren a su persona. Entonces, la privacidad puede ser violada cuando los datos personales son obtenidos, usados, procesados y diseminados, especialmente sin el consentimiento de su titular. En este contexto, es donde el web mining tendría su mayor accionar, ya que el usuario no tendría la más mínima idea de que información referente a su persona puede estar siendo procesada.

A partir del uso de algoritmos de web mining, se pueden extraer patrones respecto del comportamiento de grupos de usuarios en la Web. En este sentido el valor del "individualismo" podría verse afectado. Este concepto se relaciona con el de privacidad por cuanto muchos sistemas que usan los patrones extraídos a través del web mining, tienden a clasificar a los usuarios y a tomar decisiones en base a cuan parecido es su comportamiento respecto de un grupo. Por ejemplo, este usuario se comporta como aquellos que pertenecen al grupo de los amantes del rock, entonces las páginas a mostrarle en su navegación son sólo las referentes a ese tipo de música. Lo anterior claramente coarta toda posibilidad al usuario de que pueda tomar decisiones respecto de lo que en realidad quiere ver [13].

Si se analizan los web data utilizados para la extracción de patrones de navegación y preferencias de lo usuarios, estos se pueden agrupar en [6]:

- *Datos explícitos.* Son provistos por los usuarios en forma directa, por ejemplo, su nombre, edad, nacionalidad, etc.

- *Datos implícitos.* Se infieren a partir del comportamiento del usuario en

un sitio web, por ejemplo, qué paginas visitará, historia de compras, etc.

¿Qué parte de los web data es público y privado? La respuesta depende casi exclusivamente del país donde se haga la pregunta. Algunos países industrializados, han abordado la privacidad de los datos creando regulaciones específicas, por ejemplo en EEUU se crea la Self-Regulatory Principles for Online Preference Marketing by Network Advisers. NetworkAdvertising Initiative del año 2000, donde la legislación cubre muy pocos tipos de datos, pero se espera que esto cambie rápidamente.

Desde su creación, tanto los sistemas de información como los de recuperación de esta, siempre han contado con mecanismos de consultas hacia las bases de datos. La diferencia entre hacer muchas consultas y lo que entregan las técnicas de data mining, está justamente en la capacidad de estas últimas para extraer patrones a partir de grandes volúmenes de datos. La utilización de estos patrones para la toma de decisiones, puede entrar en conflicto con algunas reglas comerciales, por ejemplos las formuladas por la OECD[17] . Estas reglas se derivan directamente de la Directiva 95/46/EC del parlamento Europeo, donde se consigna el propósito para el cual fueron recolectados los datos (informar), en que se van a usar, etc. En este punto, por la naturaleza de las herramientas de data mining, no se puede saber a ciencia cierta a que se llegará con el procesamiento de datos. Cabe recordar que se trata de un proceso de "descubrimiento de conocimiento", es decir, no se puede anticipar qué se va a descubrir, pues el solo decirlo ya implica que está descubierto.

La situación anterior es totalmente expandible al web mining, con la salvedad de que esta vez el usuario ni siquiera tiene la opción de solicitar de que datos acerca de su navegación no sean recolectados. En efecto, por construcción y operación un sitio Web debe mantener una bitácora de sus visitantes, así que si un usuario no está de acuerdo con esta "norma", entonces no puede visitar un determinado lugar en la Web.

El dueño o mantenedor de un sitio, es amo y señor de los registros de web logs que se generen producto de la navegación de los usuarios. Perfectamente podría comercializar estos datos, pero sería muy extraño, pues estaría abriendo al mundo la mayor ventaja competitiva que puede tener una empresa en el mundo digital "conocer el comportamiento de sus clientes virtuales" [8], [9] y [11].

Visto lo anterior, la sola visita de un usuario a un sitio expone la privacidad de su navegación al dueño de los web logs. Lo mismo sucede cuando entramos en una tienda con cámaras de vigilancia. El fin altruista puede ser proteger al cliente ante los robos, pero igual este pierde intimidad al ser filmado, toda vez que su comportamiento de compra también puede ser estudiado para luego

[17]En 1980, la OECD (Organization for Economic Cooperation Development) desarrollo unos principios internacionalmente aceptados para la recolección, uso y divulgaciónn de los datos personales.

formular reglas de fidelización, promociones, etc. Desde el derecho la respuesta es clara, las filmaciones son operaciones de tratamiento de datos personales y por tanto deberán respetar los estándares legislativos correspondientes. Por tanto esas filmaciones sólo podrán ser utilizadas en el marco de la finalidad para las que fueron recogidas.

Otro punto muy importante a dejar en claro, es que los registros de web log no pueden identificar a una persona, pero si a un usuario web, es decir, un ente que posee una dirección IP desde donde se conecta, fecha y hora de visita, las páginas que visitó etc. En este sentido, en forma directa no se estaría trabajando con datos personales, por lo que la regulación que existe al respecto, podría ser insuficiente para los web data.

La utilización de mecanismos de identificación, tales como las conocidas cookies, podría establecer una relación directa entre el ser humano y el usuario web. Sin embargo, es posible que un tercero use el computador de una persona y sin desearlo la suplante en el sitio web que visita, ya que estaría usando la misma cookie que su antecesor.

Otro caso de vinculación usuario web/persona se produce en los ISP[18]. En efecto, cuando contratamos el servicio Internet, datos personales respecto de nosotros quedan consignados en un contrato. Luego para una determinada sesión, el ISP sabe al menos a través de que conexión el cliente está navegando por la Web. Sin embargo, dado que una conexión puede ser compartida, es decir, varias personas saliendo por un mismo lugar, nuevamente no es posible vincular una determinada sesión a un usuario.

La ley alemana para la "legítima interceptación" obliga a los ISPs a mantener todas las transacciones que han realizado los usuarios a través de sus sistemas, en el caso de que el gobierno las necesite para realizar una investigación criminal.

También se da el caso de que los ISP pueden ser restringidos en su operación, por ejemplo, en Holanda, este servicio es considerado como una telecomunicación más, es decir, tienen que obedecer lo estipulado en la nueva ley de Telecomunicaciones de 1998[19], el cual estipula que los ISP están obligados a borrar o hacer anónimo, todos los datos relacionados con el tráfico generado por sus suscriptores una vez que estos finalizan la llamada. La aplicación de cualquier técnica o algoritmo de extracción de información por sobre los datos generados por a través ISP, sólo se puede realizar previa autorización expresa del cliente.

La tendencia mundial en mejores prácticas para el tratamiento de los web data, especifica que se debe [6]:

- Informar al usuario que está entrando en un sistema informático el cual

[18]Internet Service Provider
[19]El Dutch Telecomunications Act de 1998.

por construcción almacenará datos respecto de su navegación en el sitio y que dichos datos pueden ser usados para hacer estudios posteriores.

- Obtener el consentimiento explícito del usuario para realizar una operación de personalización del sitio web que visita. Por ejemplo "¿desea usted que le enviemos sugerencias de navegación?".
- Proveer una explicación sobre las políticas de seguridad que se aplican para mantener los web data que se generen en el sitio.

Estas prácticas, son un marco mínimo de requerimientos para asegurar una adecuada privacidad del usuario en el tratamiento de los web data.

En Chile, la ley 19.628 sobre datos personales, consagra como tales a "los relativos a cualquier información concerniente a personas naturales, identificadas o identificables". En su sentido amplio, los web data no estarían contemplados como dato personal, salvo los referentes a las direcciones IP que podrían ser utilizadas, en combinación con otros datos para identificar a la persona detrás de la sesión del usuario. Entonces, el tratamiento de los web data podría estar regulado por la citada ley, siendo el responsable del banco de datos, el administrador o dueño del sitio web que el usuario visita.

6. Privacidad en la personalización de la Web

La personalización de la Web es la rama de la investigación en "Web Intelligence" dedicada a ayudar al usuario a que pueda encontrar lo que busca en un sitio web [5] y [9]. Para esto, se han desarrollado sistemas informáticos que ayudan a los usuarios a través de sugerencias de navegación, contenidos, etc. y más aun, entregan información valiosa a los dueños y administradores de sitios para que realicen cambios en su estructura y contenido, siempre con la idea de mejorar la experiencia del usuario, haciéndolo "sentir" como si fuese el visitante más importante del sitio, con una atención personalizada. Para lograr lo anterior, se han desarrollado múltiples esfuerzos tendientes a extraer información desde los web data que se generan con cada visita del usuario a un sitio, siendo los trabajos en web mining, los que han concentrado la mayor atención de empresas e investigadores en los últimos años.

Primero que todo, hay que dejar en claro el fin último que persigue el uso del web mining: aprender del comportamiento de los usuarios en la Web, para mejorar la estructura y contenido de un determinado sitio, personalizando la atención del usuario [11] y [13].

Como se puede apreciar, el fin es bastante altruista, siempre orientado a satisfacer al usuario y en el fondo a ayudarlo a encontrar lo que busca. Ahora bien, el exceso de "ayuda" no sólo puede molestar al usuario, sino que además,

para ayudarlo mejor, se requiere de más y más datos, conocer sus preferencias y en buenas cuentas, intrometerse en su privacidad.

Existe evidencia empírica que los sitios que incorporan sistemas de personalización de sus contenidos, logran establecer una relación de lealtad con sus visitantes [6]. Sin embargo, el precio a pagar es permitir que el sistema se inmiscuya en aspectos relacionados con las actividades del usuario en el sitio, sus hábitos anteriores de navegación o de pares parecidos, etc. En algunos casos, el usuario puede llegar a experimentar una verdadera sensación de invasión su privacidad, lo que se traduce en otra razón más por la cual un usuario no visita un sitio web que personaliza la información que muestra a sus visitantes, es decir, el "remedio fue peor que la enfermedad", por lo que el desarrollo de este tipo de sistemas se está tomando con cautela, más allá de las implicancias legales que puede traer el vulnerar la privacidad de los actos de los visitantes de un sitio.

Entonces ¿hasta que punto la personalización de la Web es invasiva de la privacidad de los usuarios? [3].[4] y [6]. La percepción dependerá mucho de las características culturales de cada país o más aun, comunidad de individuos. La solución a la cual más se ha recurrido, es realizar encuestas de opinión a los usuarios de los sitios, pero que van más de acorde a las bondades que trae la personalización, sin explicar en detalle el cómo se logra.

La creación de sistemas para personalizar la navegación en la Web, limita el libre albedrío, por cuanto asume que el usuario no es lo suficientemente avezado como para encontrar información por si solo y necesita ayuda, que al final se transforma en una imposición sublime sobre qué debe ver. Ahora bien, la gran queja de los usuarios es que visitan un sitio y nunca encuentran lo que buscan, pese a que muchas veces el contenido si estaba. La "culpa" es compartida, por cuanto si el usuario no encuentra nada, tal vez se deba a su poca experiencia en la Web, y también es muy posible que el sitio esté mal estructurado y en realidad oculte información en vez de mostrarla [8].

Entonces, ¿dónde está el punto de balance entre vulnerar privacidad y ayudar al usuario?. Tal vez la solución sea muy simple, y todo pase por preguntarle al usuario si necesita apoyo y explicarle que para ayudarlo se requiere involucrarse un poco más en su vida privada.

Lamentablemente lo anterior en la Web es complicado, ya que muchas preguntas cansan al usuario y es ineficaz.

7. 7. Análisis de la operación de las técnicas de minería de datos

Las técnicas de web mining analizadas, utilizan como entrada de datos los

web data preprocesados y en forma de vectores de características. Como ya se ha comentado antes, de todos los posibles web data, son los registros de log los que más información aportan para realizar un análisis del comportamiento de los usuarios en un sitio web [10].

Previo al uso de estos registros, se requiere aplicar un proceso de reconstrucción de la sesión de los usuarios: la sesionización. Al respecto, la Tabla 1 muestra un resumen de las técnicas más utilizadas para sesionizar registros de log.

Desde un punto de vista de la privacidad de los web data, todo apunta a que el análisis del comportamiento del usuario debe hacerse utilizando estrategias de reconstrucción de la sesión que no liguen directamente a un ser humano con el usuario web [2] y [3]. En tal sentido, las técnicas más comúnmente aceptadas son las dos primeras de la Tabla 1. Sin embargo, la extracción de patrones de navegación y preferencia de los usuarios, siempre puede ser utilizada como una forma indirecta de extrapolar el comportamiento de un visitante en un sitio web, que a través de la personalización de sus contenidos, puede atentar contra el libre albedrío del usuario, toda vez que la información que verá no será toda la que puede ver, de eso la lógica informática del sitio se va a encargar, tal como "el gran hermano" que vela por lo bueno y lo malo que se le permite ver a las personas.

Luego, asumiendo que sólo se trabajará con datos que identifican sesiones pero no personas, se construyen los vectores de características. Los más frecuentemente usados, contienen información sobre la página visitada, el tiempo que el usuario gasta por página y sesión, más alguna referencia al objeto que se está visitando [12] y [13].

Las técnicas de web mining más frecuentemente usadas, apuntan a la identificación de grupos de usuarios con preferencias de navegación y contenidos similares (uso de clustering), las cuales no permiten identificar en forma directa a la persona detrás de la sesión.

El paso siguiente es analizar cómo usar los patrones que se pueden extraer desde los grupos de usuarios con características a fines. Desde un punto netamente informático, este "cómo" se transforma en reglas "if-then-else", que junto con los patrones, configuran el conocimiento extraído desde los web data [11].

La personalización de la Web se logra a partir del uso del conocimiento extraído, el cual permite aplicar técnicas de clasificación de los usuarios por similaridades con los grupos identificados [10] y [11]. En síntesis, cuando un usuario visita el sitio, inmediatamente se procede a analizar su navegación para luego identificar a que grupo de usuarios pertenece. Luego, aplicando las reglas "if-then-else" se procede a crear la recomendación de navegación y preferencia que se enviará al usuario, siempre manteniendo la "sugerencia" por omisión que es el estado de "no sugerencia" es decir, sino hay nada bueno que

Método	Descripción	Vulneración de la Privacidad	Ventajas	Desventajas
IP + Agente	Asume que cada par IP/Agente único es una sesión	Baja	Siempre disponible. No se requiere tecnología adicional	No garantiza unicidad. Las Ips pueden cambiar
Identificadores de sesión	Uso de páginas web dinámicas para asociar ID a cada hipervínculo	Baja a Media	Siempre disponible independiente de la IP	No puede capturar visitas repetidas. Sobre carga por generación de páginas dinámicas
Registro	Uso explícito de los logs del sitio web	Media	Se pueden seguir las acciones de la persona detrás del browser	Muchos usuarios no se registran.
Cookie	Almacena una ID en el computador del cliente	Medio a Alto	Se puede hacer seguimiento de varias visitas de un mismo browser	El usuario las puede deshabilitar
Robots espías	Código que se carga en el browser y envía toda la navegación del usuario a un destinatario	Alta	Se pueden seguir todos los movimientos de un usuario en el sitio	Existen software que permiten limpiar los robot espías del alojados en el browser

Tabla 1: Técnicas de reconstrucción de la sesión de un usuario.

recomendar, entonces no se perturba al usuario con información anexa.

Es en la preparación de la recomendación donde más se puede vulnerar la privacidad del usuario, ya que se requiere de un seguimiento de sus acciones en el sitio, para poder clasificarlo en el grupo adecuado y preparar la recomendación de navegación que más se ajuste a lo que el sistema cree que el usuario anda buscando en el sitio.

En concreto, la etapa de preparación de los web data para ser usados en un proceso de web mining, puede ser realizada sin identificar a la persona detrás del browser, a través de técnicas no invasivas [2] y [3]. De hecho, son las más comúnmente aceptadas en investigación y que generan menos controversia en el tratamiento de los web data.

Las técnicas usadas en web mining para analizar el comportamiento del usuario en la Web, trabajan con miles de sesiones, sin importar quién es la persona que generó una determinada sesión. Aquí se aplica el principio estadístico de que el comportamiento de una persona es aleatorio, por lo tanto no sirve para conjeturar nada. Sin embargo, el comportamiento colectivo siempre marca una tendencia, por lo que se puede extrapolar y usar como un estimador

probabilístico aceptado.

Respecto de las técnicas de web mining utilizadas para personalizar la Web, involucran que de alguna forma se pueda hacer un análisis del usuario que en ese momento visita el sitio, incluso se podría hasta llegar a la identificación de la persona detrás del usuario (buenos días señora Lorena, ayer compró un libro de web mining electrónico, pero aun no lee nada, ¿algún comentario?), con lo cual se estaría vulnerando abiertamente la privacidad del visitante.

Finalmente, la preparación de la acción de personalización claramente limita el libre albedrío del usuario que visita el sitio, por cuanto implica limitar su exposición a contenidos que "tal vez no le son de interés". En la práctica, esta limitación no ha sido mal recibida por los usuarios, lo cual no quita que igual sea una invasión en la privacidad del visitante del sitio. Sin embargo, en el ciberespacio, ¿existe el libre albedrío?, claramente somos dueños de ir donde queramos, pero en la mayoría de los casos lo hacemos influenciados por una recomendación de un motor de búsqueda, así que al menor podemos decir que el libre albedrío estaría limitado a lo que "el gran hermano" tecnológico quiera mostrarnos.

8. Conclusiones

La identificación de una persona a partir de los web data que se recolectan en un sitio web, no es factible en su totalidad. Utilizando el actual protocolo de comunicaciones de Internet: IP versión 4. A lo más se puede identificar la sesión de un usuario, es decir, un ente que en un determinado momento está navegando en un sitio web. Cuando esté en operación el próximo protocolo de Internet: IP versión 6, será posible establecer mecanismos de autenticación de los usuarios, por ejemplo firma digital avanzada, y asociar una determinada dirección IP a una persona. En ese momento, tal vez sería conveniente analizar si es factible que este guarismo sea tratado como un dato personal.

El uso de las técnicas de web mining para la extracción de información y conocimiento desde los web data, puede vulnerar la privacidad de los usuarios que visitan un sitio web. Ahora bien, existen formas de minimizar esta vulneración hasta lo estrictamente necesario y con el consentimiento del usuario para ayudarlo en su búsqueda de información en un sitio, comenzando por cómo se pre procesan los web data y finalizando en la forma en que se entregan las recomendaciones de navegación y preferencias de contenido.

La personalización de la Web es el área de investigación en Web Intelligence que por lejos ha acaparado más seguidores en investigación y en el comercio. Se le considera el siguiente paso en la creación de sitios web, por lo que el desarrollo tecnológico a su alrededor ha sido exponencial. Como suele suceder

cuando aparece una nueva tecnología, el marco regulatorio no existe y su creación demora, comparativamente hablando, "siglos" en estar listo, y cuando esto ocurre, ya está obsoleto. Sin embargo, existen ciertos principios universales que se podrían cautelar y que son independientes del cambio tecnológico. Uno de ellos es la privacidad de la navegación del usuario en un sitio web. Al menos debería quedar claro que si el usuario acepta que su visita sea personalizada, entonces existirá un seguimiento de sus acciones y que los datos que genere su sesión serán usados para establecer líneas de acción en los cambios que experimentará el sitio web.

Agradecimientos:

Referencias

[1] http://wi.dii.uchile.cl/

[2] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, 1(1):5-32, 1999.

[3] M. Eirinaki and M. Vazirgannis. Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1):1-27, 2003.

[4] M. Kilfoil, A. Ghorbani, W. Xing, Z. Lei, J. Lu, J. Zhang, and X. Xu. Toward an adaptive web: The state of the art and science. In Procs. *Annual Conference on Communication Networks & Services Research*, pages 119-130, Moncton, Canada, May 2003.

[5] W. Kim. Personalization: Definition, status, and challenges ahead. *Journal of Object Technology*, 1(1):29-40, 2002.

[6] A. Kobsa. Tailoring privacy to usersñeeds. *In In Procs. of the* 8^{th} *International Conference in User Modeling*, pages 303-313, 2001.

[7] Z. Markov and D. T. Larose. Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage. *John Wiley & Sons*, 2007.

[8] M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence*, 118(1-2):245-275, April 2000.

[9] J.D. Velásquez and V. Palade. Adaptive Web Site. *IOS Press*, Netherlands. 2008.

[10] J.D. Velásquez and V. Palade. Building a Knowledge Base for Implementing a Web-Based Computerized Recommendation System. *International Journal of Artificial Intelligence Tools*, 2007.

[11] J.D. Velásquez and V. Palade. A Knowledge Base for the maintenance of knowledge extracted from web data. *Knowledge-Based Systems*, 20(3):238-248.

[12] S.A. Ríos, J.D. Velásquez, H. Yasuda and T. Aoki, Web site off-line structure reconfiguration: A web user browsing analysis. *Lecture Notes in Artificial Intelligence*, 4252(1):371-378, 2006.

[13] J.D. Velásquez, R. Weber, H. Yasuda and T. Aoki. Acquisition and maintenance of knowledge for web site online navigation suggestions. *IEICE Transactions on Information and Systems*, E88-D(5):993-1003, 2005.

[14] A. Vedder. KDD: The Challenge to Individualism. *Ethics and Information Technology*, 1: 275-281, 1999.