Aplicación de Técnicas de Web Mining sobre los Datos Originados por Usuarios de Páginas Web. Visión Crítica desde las Garantías Fundamentales, especialmente la Libertad, la Privacidad y el Honor de las Personas

Juan D. Velásquez[*]
Lorena Donoso[**]

Resumen

Web mining es la aplicación del data mining a los web data para la extracción y descubrimiento automático de información y conocimiento. Dependiendo del tipo de web data a procesar, web mining se divide en tres grandes categorías: contenido, estructura y uso. El análisis de estos datos permite a las instituciones significativas mejoras en la estructura y contenido de los sitios web corporativos, así como la aplicación de complejos sistemas informáticos destinados a personalizar la experiencia del usuario en el sitio que visita. El presente artículo muestra una revisión científico-técnica de los fundamentos del web mining, sus principales técnicas, métodos y algoritmos, con especial atención en aquellos que permiten extrapolar las preferencias de navegación y contenidos de los usuarios que visitan un sitio web determinado, para finalmente contrastar su operación con la regulación vigente a nivel nacional e internacional.

Palabras Clave: *Web Mining*, *Privacidad*, *Regulación.*

[*]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.

1. Introducción

La World Wide Web o simplemente **La Web** [1] es tal vez el mayor portento tecnológico que el hombre haya desarrollado jamás. Su impacto en nuestra sociedad ha sido tal que se le ha comparado con la invención de la rueda o el descubrimiento del fuego.

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 abocados a una sola misión: asegurar que el contenido y la estructura del sitio le son atractivos al usuario. Lo anterior se ha abordado con relativo 'éxito en el 'ámbito de la *"personalización de la Web"*, concepto que 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.

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 [7]. 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.

Los datos originados en la Web o web data, prácticamente corresponden a todos los datos que se han originado a lo largo de la historia de la computación. En efecto, aquí se encuentran los hipervínculos entre páginas web y sus los contenidos, que pueden ser imágenes, sonidos, videos, texto libre, etc. A lo anterior, se debe agregar datos acerca de la navegación del usuario en los sitios que visita, específicamente la IP desde donde accedió y el tipo de navegador utilizado.

La ley 19.628 sobre protección de la vida privada, establece ciertas restricciones al procesamiento de datos personales, por lo que de entrar los web data en esta categoría, es importante analizar hasta que punto su procesamiento está conforme a la regulación vigente. Si adicionalmente se consideran otros datos que los mismos usuarios pueden develar en blogs, foros o sistemas similares, tales como vinculaciones políticas, vida sexual, origen racial, ideologías

o convicciones religiosas, etc. los web data también entrarán en la categoría de datos sensibles, según lo consigna la letra *"g"* del artículo 2"*o* de la mencionada ley.

En esencia, los algoritmos, técnicas y métodos que comprende el web mining, son utilizados en el procesamiento masivo de datos, lo cual requiere una automatización parcial o total de todas las operaciones a fin de obtener resultados en cuestión de horas o días. En consecuencia, el análisis de los web data utilizando técnicas de web mining cuenta con todos los requisitos necesarios para ser estudiando a partir de la regulación nacional e internacional que hasta el momento se ha desarrollado. En particular, es de suma importancia revisar ¿hasta dónde este afán por analizar al usuario en la Web no se transforma en una persecución? [2].

El presente artículo comienza con la sección 2, la cual aborda el fenómeno Internet y Web desde sus orígenes hasta nuestros días, explicando a grandes rasgos su funcionamiento, los datos que se pueden recolectar y cuales estarían directamente relacionados con información relativa a las personas naturales. A continuación, en la sección 3, se analizan la operación de las técnicas, algoritmos y metodologías propuestas en web mining, del punto de vista de la restricción a la libertad y vulneración de la privacidad de los usuarios de sitios web.

Por su parte, la sección 4 profundiza en los aspectos jurídicos relacionados con el tratamiento de los datos originados en la Web, también conocidos como Web Data. Finalmente, en la sección se presentan las principales conclusiones y recomendaciones que se han obtenido a lo largo de este trabajo.

2. Internet y la Web

Es importante hacer la distinción entre la Web e Internet, ya que son conceptos distintos pero que a menudo se confunden. Internet representa a la red de redes que permite la interconexión de dispositivos que se encuentran a nivel local, con sus similares en una región diferente, a través del envío y recepción de los datos que viajan en paquetes o datagramas. La Web es el conjunto de páginas y objetos relacionados que se vinculan entre si a través de hipervínculos. A un conjunto de páginas web se le denomina sitio web y es administrado por una aplicación conocida como servidor web, la cual utiliza a Internet como lugar físico para transferir las páginas web y otros objetos asociados. De acuerdo a la definición dada por su creador Tim Berners-Lee en 1989, la *"World Wide Web es el universo de información accesible en la red, una encarnación del conocimiento humano"* [1].

2.1. Datos originados en la Web

La Web es el conjunto de archivos (páginas) que se relacionan a través de hipervínculos, almacenados en los servidores ubicados alrededor del mundo, para lo cual se utiliza un mecanismo de direccionamiento global de documentos y de otros recursos conocido como URL. Cada una de estas páginas posee un contenido representado a través de objetos como texto, imágenes, sonidos, películas o vínculos a otros sitios web.

La Fig. 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 los envía (2) a la aplicación que hace la solicitud, generalmente un web browser (6). En general estos archivos son imágenes, sonidos, películas y 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 [3].

Cada uno de los tags presentes en una página, son interpretados por el browser. Algunos de estos 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.

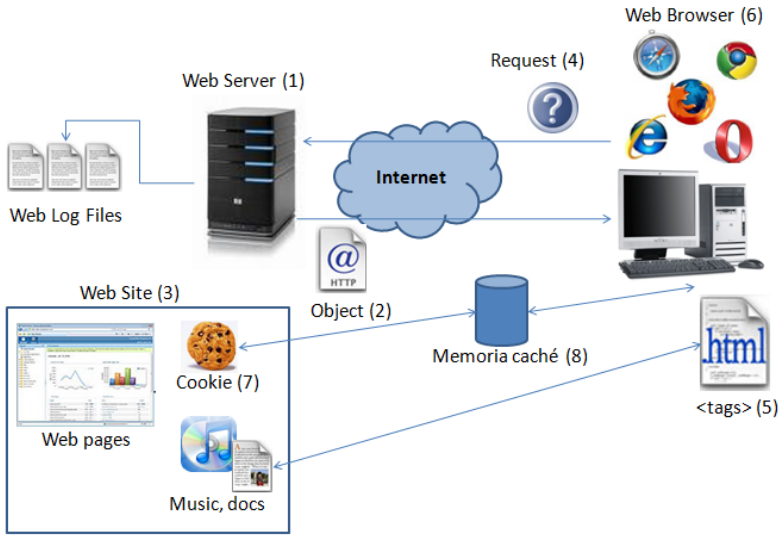

Figura 1: Modelo básico de operación de la Web

La interacción anterior, ha quedado registrada en archivos conocidos como web log files [3], 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 Fig. 2 muestra un ejemplo del contenido y estructura de un archivo de web log.

N°	IP	ID	Access	Time	Method/URL/Protocol	Status	Bytes	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 4.0.1;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 4.0.1; 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: Estructura de un web log file

La estructura estándar del archivo web log queda definida entonces por la dirección IP del dispositivo desde donde se solicita la página, los parámetros IP y Access que permiten una forma de identificación de la sesión, time que especifica la fecha y momento en que se hizo el requerimiento, el tipo de método utilizado para solicitar la página, el estado (status) de la solicitud, la cantidad de bits transmitidos, la URL de la última página visitada y el tipo de aplicación utilizado para navegar (Agent)

Otra forma de capturar el comportamiento del usuario en una sesión es a través de las cookies (7) que se almacenan en el disco duro del cliente a través del Browser o Navegador, y que guardan parte de la información de la página que visitó, usando la memoria de rápido acceso o memoria caché (8). Estas cookies son generadas a pedido del servidor, y se utilizan tanto para el control de usuarios, por ejemplo cuando se pide una contraseña en algún sitio, como para ver el comportamiento de navegación de los usuarios. Hay que dejar en claro que este mecanismo no identifica a una persona en particular, sino a un tipo de usuario que navega en un sitio web en un determinado Browser.

Como se puede ver, 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, permitiendo un análisis profundo de sus preferencias de información, las cuales están directamente relacionadas con las características de los productos y servicios ofrecidos.

2.2. Otros datos presentes en la Web

Para acceder a los datos existentes en la Web, primero se requiere una conexión a Internet, la cual se realiza a través de diversos métodos, siendo

uno de los más populares el de la contratación de un Internet Service Provider (ISP), empresas que utilizando un medio de comunicación, como puede ser el espectro radioeléctrico, TV-cable, telefonía convencional, etc. brindan el acceso de un abonado a la gran red [12].

Independiente del método, toda navegación o interacción del usuario con la Web, se realiza a partir del envío de paquetes de datos, que por construcción poseen la dirección IP de origen y destino de la comunicación. Esta estructura permite un monitoreo muy eficaz y eficiente de las sesiones que establecen los usuarios que navegan por la Web.

Variados son los dispositivos y aplicaciones usados para brindarle conectividad a los usuarios y también para restringir su navegación en la Web. Algunos de estos son:

- Router: Dispositivo por excelencia que permite la conectividad de una Red de Área Local (LAN) a la Red Externa (WAN).

- Proxy. Se trata de un espacio de memoria, comúnmente discos duros en un computador dedicado, para el almacenamiento de todas las páginas web que se visitan desde una red local.

- Firewall. El dispositivo de seguridad por excelencia de una red local. Se trata de un conjunto de aplicaciones de seguridad que se ejecutan preferentemente en un computador dedicado a estos menesteres, que por arquitectura aisla a la red local de la Internet.

Los dispositivos antes analizados, permiten recopilar datos sobre toda la navegación que los usuarios realicen en la Web, con lo cual es posible realizar un trabajo muy profundo de monitoreo de sus actividades. Sin embargo, estos datos quedan en posesión del dueño del dispositivo, es decir, si se desea acceso a la navegación de los usuarios que tienen contratado un determinado ISP, habrá que solicitar estos datos al dueño del servicio, permiso que, salvo una orden judicial, es muy difícil de obtener.

En vista de estas restricciones, se han desarrollado otros métodos para recolectar parte de los datos originados en la Web:

- Web Crawlers. También conocido como *Spyder Robots*, son aplicaciones que recorren la Web de forma automática y sistemática, almacenando información de los sitios web mediante el uso de sus hipervínculos.

- Spyware: Es una aplicación cuya función es espiar las actividades de navegación de un usuario desde un computador conectado a Internet. Esta aplicación, claramente no se instala con la venia de los usuarios, lo cual la transforma en un ente extraordinariamente invasivo, amén que ocupa los recursos del computador para su funcionamiento (RAM,

CPU, etc.). Su instalación se realiza cuando el usuario sin saber navega por sitios que contienen spywares.

Estos métodos se orientan en parte a capturar toda la información que se pueda de las transacciones y navegación que realiza un usuario. De un punto de vista comercial, entender el comportamiento de compra de los clientes a través de la Web es vital, en términos de mejorar el contenido del sitio utilizado para el negocio, ya sea Business to Business (B2B), Business to Consumer (B2C), Peer to Peer (P2P) o cualquier otra variación que requiera del análisis de los registros de un visitante vía web.

En el caso particular de los usuarios que realizan compras por Internet, generalmente se almacenan datos personales, ya sea de una cuenta de usuario o tarjetas que permitan realizar este tipo de compras. Las personas han hecho uso gradual de este medio como un sistema de compras, dada la inseguridad (intangibilidad del producto), verificación de la compra, detalles del producto a entregar, plazos, etc. Por esto es que la idea es adaptarse a las necesidades de los clientes, haciéndolos sentir que la compra es personalizada (por ejemplo se puede hacer sugerencias basándose en el historial del cliente), fácil, rápida y que agrega valor. Este último es muy importante por ejemplo para las empresas, ya que a través de Internet se ha eliminado parte de la cadena de suministro. Un ejemplo claro es Amazon, ya que no tiene necesidad de tener un inventario en exhibición de sus libros, acortando la cadena, eliminando a los elementos que no aportan y disminuyendo de esta forma los costos asociados. Otro ejemplo son las transacciones que realizan los bancos e instituciones financieras, que al ser vía electrónica ahorran en impresiones, personal y tiempo tanto para la empresa como para los clientes.

3. La dirección IP como dato personal

Hemos dicho que en la ley se define datos personales como cualquier información relativa a personas naturales identificadas o identificables. Al respecto cabe señalar que esta definición es similar a las que se han adoptado en otros países, por lo que nos podrá ayudar a dilucidar si una dirección IP puede ser considerada un dato personal no obstante no haya pronunciamientos expresos en nuestra legislación o jurisprudencia.

Conceptualizando una dirección IP, la RFC 791[1] , la dirección Internet es *"una dirección de origen o destino de 4 octetos (32 bits) formada por un campo de Red y un campo de Dirección Local"*. Ello nos denota que la finalidad del establecimiento de estas direcciones es el reconocer máquinas interconectadas a través del protocolo IP, y no necesariamente a las personas que están

[1] http://www.rfc-es.org/rfc/rfc0791-es.txt [consulta: 31.03.2010]

operándolas. Sin embargo, con la masificación de Internet y su desarrollo como sistemas de intercambio de la información, se ha cuestionado la posibilidad de que estas direcciones sean consideradas un dato personal.

Una primera aproximación al respecto nos llama a reparar en que la definición de dato personal exige que la persona a quien se refiere el dato personal sea al menos *"identificable?* esto es, que exista la posibilidad de identificarla, sin importar las dificultades técnicas y/o económicas que la determinación implique. Asimismo, se ha sostenido que la identificabilidad es un atributo cambiable en el tiempo, en donde tiene mucho que decir el avance científico y tecnológico. En efecto, hace unos años atrás no era posible identificar a una persona específica a través de una muestra biológica, en cambio hoy en día es perfectamente factible hacerlo. Con esto queremos señalar que un dato, por ejemplo de dirección IP, si hoy, por las condiciones de mercado y/o tecnológicas no es atribuible a una persona determinada o determinable, es posible que mañana si pueda ser considerado como tal. Siendo así la pregunta natural es ¿qué criterio debemos aplicar al respecto?. En derecho comparado, ya en 2003, la Agencia Española de Protección de datos (AEPD), mediante informe 327/03 sostuvo que las direcciones IP, tanto fijas como dinámicas, son datos de carácter personal, decisión que basa en los siguientes argumentos:

a) Es factible identificar por medios razonables a los usuarios a los que se asigna una dirección IP fija o dinámica, por parte de los proveedores de acceso a Internet y los administradores de redes locales.

b) Con la asistencia de terceras partes responsables de la asignación de la dirección IP se puede identificar a un usuario de Internet por medios razonables.

c) Existe la posibilidad de relacionar la dirección IP del usuario con otros datos de carácter personal, de acceso público o no, que permitan identificarlo, especialmente si se utilizan medios invisibles de tratamiento para recoger información adicional sobre el usuario, tales como cookies con un identificador único o sistemas modernos de minería de datos.

Este pronunciamiento si bien ha sido controvertido desde la óptica técnico/económica, en el sentido que la aplicación del estatuto jurídico de los datos personales, implica que se les deba aplicar medidas de seguridad al menos de nivel básico, ha sido en general acatada y sostenida en el tiempo. Claro está en este entorno, en todo caso, que no serán considerados dato personal las IP disociadas ya sea porque han sido sometidas al proceso de disociación y/o que se hayan generado disociadas, esto es, que no sea posible por ningún medio atribuirla a una persona determinada y/o determinable.

De su parte, en el seno de la Unión Europea, el grupo del artículo 29 (que es aquel referido al tratamiento de datos personales), el año 2000 se

había pronunciado en este mismo sentido a través del documento de trabajo 5063/00/ES/Final (wp37), titulado **Privacidad en Internet: enfoque integrado comunitario de la protección de datos en línea**. Esta opinión fue ratificada a través de dictamen 04/2007 sobre el concepto de datos personales y recientemente fue aplicada en un caso concreto que ha suscitado bastante polémica, relativo a las actividades de tratamiento de datos de IP de algunos buscadores de Internet. En este documento se enfatiza que *"a menos que el prestador de servicios de Internet sepa con absoluta certeza que los datos corresponden a usuarios que no pueden ser identificados, tendrá que tratar toda información IP como datos personales para guardarse las espaldas?.*

Un tercer documento relevante de la Unión Europea nos lleva a las mismas conclusiones. Se trata de la Directiva de comunicaciones electrónicas 2002/58 del Parlamento y del Consejo, a cuyo respecto el Grupo del Artículo 29 propone su modificación a través de dictamen 2/2008, en la que profundiza aún más sobre las consecuencias de la consideración de las direcciones IP como datos personales, proponiendo que en la directiva en comento se incluya la obligación de los ESP de notificar los incidentes de seguridad de datos personales a los usuarios *"interesados"*. Proponen que esta obligación se extienda no sólo a los proveedores del servicio de acceso a Internet, sino a todos los proveedores de servicios de la sociedad de la información. A su turno, respecto de los interesados, el grupo 29 propone que sean considerados como tales no sólo los abonados, sino todas aquellas personas cuyos datos se han visto efectivamente comprometidos por la violación de seguridad.

Concluyendo, con independencia de las consideraciones técnicas que podamos realizar, nos parece meridianamente claras las siguientes conclusiones:

a) Siempre que no sea posible sostener la imposibilidad de que una dirección IP sea atribuible a una persona, habrá de dársele el tratamiento de un dato personal, aplicando la legislación correspondiente.

b) Será de responsabilidad de quien trata estos datos el acreditar que no existen medios razonables para atribuir a una persona esos datos personales. En consecuencia, a este sujeto le corresponde probar la disociación del dato, lo que implica invertir la carga de la prueba.

4. Minería de datos de la Web

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.

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 [13, 14]. Entonces surge la pregunta de ¿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?.

4.1. Limpieza y preprocesamiento de los web data

Los datos originados en la Web o web data, corresponden esencialmente a tres fuentes [3]:

1. **Contenido**: Son los objetos que aparecen dentro de una página web, por ejemplo las imágenes, los textos libres, sonidos, etc.
2. **Estructura**: Se refiere a la estructura de hipervínculos presentes en una página.
3. **Uso**: Son los registros de web logs, que contienen toda la interacción entre los usuarios y el sitio web.

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 [6, 16].

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** [3].

Cabe señalar que ciertos sitios han cambiado su estructura con el propósito de identificar a sus visitantes [17]. 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 [6].

4.2. Técnicas, algoritmos y métodos usados en web mining

El concepto web mining, agrupa a todas las técnicas, algoritmos y metodologías utilizadas para extraer información y conocimiento desde los web data, entre los cuales se cuentan [9, 17]:

1. **Self Organizing Feature Maps (SOFMs)**: Esta herramienta tiene una estructura semejante a las redes neuronales, pero en este caso el aprendizaje se da de manera competitiva, es decir, las neuronas compiten para ser activadas, y sólo lo hace una a la vez. La idea de este aprendizaje es que se compara un elemento con la red con el fin de encontrar la neurona más similar, o neurona ganadora. A partir de lo anterior se generar grupos de neuronas o clusters cuyas características son similares.

2. **K-Means**: Este algoritmo se basa en la determinación de grupos o clusters dentro de un conjunto de datos. Para su funcionamiento se necesita como parámetro el número esperado de grupos (k). Cada uno de estos clusters estará representado por un centroide, que es el elemento cuyas características se parecen más a las de su conjunto (Obtenido mediante una medida de similitud). Este método tiene una alta performance, por lo que es posible repetirlo varias veces con distintos parámetros.

3. **Árboles de Decisión**: Esta técnica se basa en la estimación de un resultado y toma de decisiones a partir de datos conocidos. La idea es identificar los atributos mínimos con los cuales se pueda deducir un resultado, clasificando los datos en una estructura de árbol y moviéndose a través de las ramas.

4. **Support Vector Machines (SVMs)**: En comparación con las redes neuronales, tiene la ventaja de ser menos propensos al sobre aprendizaje, por lo tanto pueden mantener un gran número de características y datos sin preocuparse de la complejidad del problema. La idea básica de esta herramienta es trabajar con ciertas funciones efectivas (Funciones de Kernel) que permitan tratar los datos a otro nivel dimensional y de esta forma trabajar con modelos complejos.

5. **Algoritmos Inspirados en la Vida**. Se trata de una nueva familia de algoritmos cuya operación está basada en cómo ciertas especies, bacterias y la misma evolución con cambios genéticos, tratan de sobrevivir y perpetuarse en la vida.

4.3. 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 [16]. Los otros web data: contenido y estructura, pueden ser usados como complemento para hacer más certera la aplicación de técnicas como clustering, clasificación y la estadística.

El resultado que más interesa a las empresas dueñas de sitios web orientados al comercio electrónico, es la creación de sistemas que permitan mejorar la experiencia de los usuarios en el sitio a partir de la personalización de su navegación, lo cual se logra fundamentalmente a través de recomendaciones en línea, respecto de qué deben ver o por donde deberían dirigir su navegación. Lo anterior no elimina la posibilidad de que también se hagan recomendaciones a los administradores del sitio respecto de modificaciones que se deben hacer durante su mantención, es decir, sin usuarios concurrentes.

4.3.1. Procesamiento de los registros de web log

Previo al uso de estos registros, se requiere aplicar un proceso de reconstrucción de la sesión de los usuarios: la sesionización.

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 [3, 6]. 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 [10], 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 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 [18].

4.3.2. Procesamiento de los contenidos en una página web

En una página web se pueden encontrar todos los contenidos desarrollados a lo largo de la historia de la computación, con una variada posibilidad de

formatos. Entonces, el análisis de estos datos se vuelve un proceso no trivial, que requiere de un preprocesamiento y representación de la información previo.

El primer tipo de dato a analizar es el texto libre, el cual corresponde a todo lo que esté escrito y que se haya consignado en una página web, ya sea a través de un archivo enlazado o dentro de la misma página. Estos textos deben ser transformados a un formato numérico, el cual considera que existen palabras más importantes que otras, que se puede reducir un conjunto de palabras a la idea central y que es posible prescindir de algunas estructuras morfológicas [15].

Los otros tipos de datos a analizar, corresponden a imágenes, sonidos y videos. Por lo pronto el desarrollo de herramientas de web mining para estos formatos se encuentra en sus primeras etapas de investigación, siendo necesario recurrir a los metadatos, esto es datos por sobre los datos, que permitan procesar el entorno de estos objetos. Por ejemplo, si se está buscando información sobre una persona, específicamente su fotografía o la escena donde aparece en un video en la Web, será necesario conocer datos adicionales, consignados en los textos que acompañan al objeto, tal como su nombre, edad, etc.

4.3.3. Procesamiento de la estructura de hipervínculos

El análisis de la estructura de hipervínculos, apunta principalmente a la extracción de información respecto de la importancia de una página en la Web, la identificación de comunidades y el rankeo de la información recuperada por alguno de los motores de búsqueda. Con esta conocimiento es posible mejorar notablemente la búsqueda de información que realiza el motor para su usuario.

Por construcción, los motores de búsqueda realizan periódicamente una actualización de su base de datos de páginas web, esto es, revisar la Web y recuperar los objetos que han variado en un sitio desde su última visita, respetando la política de seguridad que se haya configurado en el servidor web que mantiene al sitio, es decir, si un objeto no tiene permiso para ser recuperado por el motor, entonces dicha operación no se lleva a cabo.

Bajo la premisa anterior, se puede pensar que el motor de búsqueda sólo puede realizar operaciones de análisis de las páginas que hayan podido ser recuperadas de manera directa, es decir, sin la necesidad de recurrir a algún mecanismo de seguridad como puede ser la aplicación de una clave de acceso. Entonces, la regla es que sólo aquello que es público puede ser buscado en la Web, por lo que la responsabilidad de la publicidad de los contenidos de un sitio queda expresamente consignada a quienes lo mantienen.

Durante todo el período de la Web, también conocido como 1.0, eran los dueños y administradores de los sitios los encargados de publicar la información que se haría publica en el ciberespacio. Sin embargo, con el advenimiento de la Web 2.0, algo cambió radicalmente. Ahora son los usuarios los que han

tomado el control de la publicación de información que muchas veces les es privada, pero que quieren mostrar al mundo, por ejemplo a través de un blog, foro, facebook, etc. . De inmediato surgen varias interrogantes:

- ¿De quién son los datos?¿Del usuario que lo publicó o del dueño del sitio?.
- Si un usuario quiere borrar algo que el mismo publicó, ¿existen los canales directos para hacerlo?.
- Si alguien publicó en un sitio información que daña la honra de una persona ¿a quién se le obliga a eliminar la página y dar las compensaciones necesarias?¿al dueño del sitio o al usuario?.

La premisa de que la responsabilidad de publicación de información en un sitio es de quien lo mantiene, no está clara, por cuanto ahora son los usuarios los que pueden crear sus propios contenidos [4].

4.3.4. Análisis de la operación de los sistemas de recomendación

La próxima generación de sitios web, estará fuertemente influenciada por la capacidad que estos tengan para adaptar su estructura y contenido a las necesidades de información que tenga el usuario, ya sea durante su navegación o luego que esta se haya realizado [17]. Este nuevo tipo de sistemas incorpora módulos de personalización del sitio, los cuales tienen su realización práctica en la recomendación de qué visitar, buscar o simplemente observar que se le hace a los usuarios de un sitio.

Es en la preparación de la recomendación donde más se puede vulnerar la privacidad del usuario [13], 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.

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.

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 [11, 5].

5. Aspectos jurídicos del tratamiento de web data

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.

5.1. Marco legal para el análisis de la privacidad en la Web

Partamos analizando 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 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. 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 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 esta.

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 este está conectado e 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.

Asumiendo, entonces, que a través de la dirección IP sólo se puede identificar 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 [9].

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 artículo no se ahondará en el contexto filosófico de la privacidad, sino que se fijarán límites 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 le son propios de 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 mas 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 que en realidad quiere ver [16].

Otro punto muy importante a dejar en claro, es que los registros web log, identifican a usuarios y no necesariamente a personas determinadas, no obstante pudieran ser determinables y por esta vía ingresan al estatuto jurídico de los datos personales. Es decir, lo que los web log identifican directamente es un ente que tiene acceso a una dirección IP desde la que se conecta, fecha y hora de visita, las páginas que visitó, etc., para los efectos de determinar si, cuando trabajamos con estos datos lo hacemos con datos personales deberemos previamente cuestionarnos sobre si existen medios razonables para identificar efectivamente a las personas *"detrás"* de esos usuarios. Si la respuesta es sí: estaremos frente a un dato personal.

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. 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. En Chile, el decreto 142 de 2005 de la Subtel señala que los ISPs deben mantener un registro, no inferior a seis meses, de las conexiones a Internet que realicen sus abonados.

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, 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 [8]:

- Informar al usuario que está entrando en un sistema informático el cual 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*[2]. En su sentido amplio, los web data no estarían con-

[2]ey sobre protección de datos de carácter personal. Ver http://www.bnc.cl/, 2002

templados 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.

5.2. Privacidad y libertad de navegación 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 [7, 17]. 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 [16].

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 y limitar la cantidad de contenidos que puede ver de un sitio.

Existe evidencia empírica que los sitios que incorporan sistemas de personalización de sus contenidos, logran establecer una relación de lealtad con sus visitantes. 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. Adicionalmente, los sistemas de personalización tienden a mostrar lo que se cree es bueno e interesante para el usuario, coartándole su libertad de navegación y restringiendolo sólo a lo que el sistema considera que es importante o necesario que vea.

Entonces ¿hasta qué punto la personalización de la Web es invasiva de la privacidad de los usuarios? [6], ¿Se coarta la libertad de navegación al ocultar o sólo mostrar ciertos contenidos, dependiendo de lo que el sistema estime es conveniente para el usuario?. 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. Entonces, ¿dónde está el punto de balance entre vulnerar privacidad, coartar la libertad de navegación y ayudar efectivamente 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.

5.3. Comentarios finales

En la Ley 19.628, artículo 2^o letra *"e"* se define como dato estadístico a aquel que *"en su origen, o como consecuencia de su tratamiento, no puede ser asociado a un titular identificado o identificable"*. En este sentido, las técnicas de preprocesamiento y limpieza de web data, que son aplicadas como paso previo al web mining, pueden eliminar cualquier indicio que identifique o permita identificar a los usuarios, transformando de esta forma al web data en un dato estadístico. El problema es que si se realiza esta práctica, se minimiza el beneficio potencial que las empresas pueden obtener respecto del uso que los visitantes les dan a sus sitios web corporativos.

Del punto de vista de la investigación científica, el uso de web data de corte estadístico no es un problema, por cuanto lo que se estudia es el comportamiento de grupos de usuarios utilizando los datos consignados en las sesiones y los contenidos de las páginas web, con lo que se salvaguardaría la privacidad de los usuarios.

Respecto de la afectación a la libertad de navegación, esta viene dada principalmente por los sistemas de personalización y adaptación de la Web. Si bien es cierto que tanto del punto de vista científico como de negocio no es necesa-

ria la identificación del usuario para dar una recomendación de navegación o la reestructuración en línea de los sitios, de todas maneras se producirá una reducción de los posibles contenidos que el usuario podrá ver. En este sentido, es necesario que al menos se de la posibilidad al usuario para que libre y soberanamente decida si quiere ser ayudado por un sistema informático a encontrar lo que busca o si desea hacerlo por su propia cuenta.

El desarrollo actual y futuro de Internet y la Web, estará estrechamente relacionado con los avances en telecomunicaciones. Entonces se hace necesario analizar la normativa vigente para las telecomunicaciones, por cuanto los datos originados en la Web, se transmiten a través de redes de computadores, cuyos medios justamente son regulados por ley. En efecto, en la ley General de Telecomunicaciones 18.168 [3], se entiende por telecomunicación a *"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"*.

Lo primero antes de establecer la normativa a aplicar en el caso de los web data, desde el ámbito de la ley 18.168, es establecer a qué servicio correspondería su transmisión dentro de Internet. Yendo al meollo mismo, habría que analizar que tipo de servicio es el acceso a la gran red, pues es ahí donde comienza la generación de los datos. En base a la definición de **Servicio Complementario**, que se consagra en la ley como *"servicios adicionales que pueden ser prestados por concesionarios de servicio público o terceros, mediante la conexión de equipos a redes públicas. No requieren autorización previa de ninguna autoridad"*, se puede argumentar que el acceso a Internet es un servicio complementario de telecomunicaciones, por lo que las normas y principios de esta rama del derecho le son plenamente aplicables, especialmente los principios de libertad y secreto de las comunicaciones, neutralidad de red, acceso universal, protección de usuarios, etc., respecto de los cuales corresponde velar a la Subsecretaría de Telecomunicaciones (Subtel), la cual a su vez debe estar encargada de velar por la protección de los derechos de los abonados que le pagan a una empresa por usar el enlace que le permite llegar a la gran red, sin perjuicio de otras instancias que le permitan al abonado estampar algún reclamo.

6. 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

[3] Ley general de telecomunicaciones. Ver http://www. bnc.cl/, 1982

está navegando en un sitio web.

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.

Finalmente, la pregunta que motiva este artículo tiene como respuesta de que efectivamente, las herramientas de web mining pueden ser el soporte tecnológico como para que se vulnere la privacidad de los usuarios y se coarte su libertad de navegación a través de sistemas que buscan la personalización de su experiencia en un sitio web. Por lo tanto, lo que se debe hacer es promover un conjunto de buenas prácticas para hacer un trabajo limpio, ético y que salvaguarde las garantías fundamentales de todos los involucrados. No se recomienda bajo ningún precepto la creación de una nueva regulación que sólo actuaría en casos puntuales, que en muy poco tiempo quedaría obsoleta y lo que es peor, detendrá el desarrollo científico en un área tan importante como lo es el futuro de la Web.

Agradecimientos

El primer autor agradece el aporte del Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FBO16).

Referencias

[1] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret. The world wide web. *Communications of ACM*, 37(8):76–82, 1994.

[2] P. Carrasco-Jiménez. *Análisis Masivo de Datos y Contraterrorismo*. Tirant lo Blanch, Valencia, España, 2009.

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

[4] M. Corripio-Gil-Delgado. *Regulación Jurídica de los Tratamientos de Datos Personales realizados por el Sector privado en Internet*. Agencia de Protección de Datos, Madrid, España, 2000.

[5] P.L. Murillo de la Cueva and J. L. Piñar-Mañas. *El Derecho a la Autodeterminación Informativa*. Fundación Coloquio Jurídico Europeo, Madrid, España, 2009.

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

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

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

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

[10] M.D. Mulvenna, S.S. Anand, and A.G. Büchner. Personalization on the net using web mining. *Communications of the ACM*, 43(8):123–125, 2000.

[11] M. Muñoz-Campos and H. Soto-Arroyo. *Derecho de Autodeterminación Informativa*. Editorial Jurídica Continental, San José, Costa Rica, 2005.

[12] W. Stallings. *SNMP, SNMPv2, and CMIP: the practical guide to network management*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1993.

[13] H.T. Tavani. Informational privacy, data mining, and the internet. *Ethics and Information Technology*, 1:137–145, 1999.

[14] A. Vedder. Privacy and confidentiality. medical data, new information technologies, and the need for normative principles other than privacy rules. *Law and Medicine*, 3:441–459, 2000.

[15] J.D. Velásquez and P. González. Expanding the possibilities of deliberation: The use of data mining for strengthening democracy with an application to education reform. *The Information Society*, 26(1):1–16, 2010.

[16] 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, 2007.

[17] J.D. Velásquez and V. Palade. *Adaptive Web Site*. IOS Press, Amsterdam, Netherland, 2008.

[18] 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.