UNIVERSIDAD DE CHILE

FACULTAD DE DERECHO

DEPARTAMENTO DE DERECHO PROCESAL

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 Domingo Velásquez Silva

Prof. Guía : Sra. Lorena Donoso A

TESIS PARA OPTAR AL GRADO DE

MAGISTER EN DERECHO DE LA INFORMÁTICA Y DE LAS TELECOMUNICACIONES

SANTIAGO - CHILE

JULIO, 2010.

A mi pequeño Joaquín

Compus sui
Se dueño de ti mismo

Índice general

1. Introducción **1**

1.1. Hipótesis de Investigación 4

1.2. Objetivos Generales y Específicos 5

1.3. Metodología 6

1.4. Contribuciones de la Tesis 7

1.5. Organización de la Tesis 7

2. Orígenes de las telecomunicaciones **9**

2.1. Concepto de señal 10

2.2. Modelamiento de señales 14

2.3. Transmisión de señales 15

2.4. Espectro radioeléctrico 17

2.5. Análogo a digital 18

2.6. Representación binaria 23

2.7. Naturaleza jurídica del servicio de acceso a Internet 25

3. Redes de computadores **31**

3.1. Topologías de redes 32

3.1.1. Bus de datos 32

3.1.2. Bus estrellado 34

3.1.3. Token ring 35

3.1.4. Ethernet 36
3.1.5. Equipos de interconexion 38
3.2. Modelo de Interconexión de Computadores ISO/OSI 42
3.3. Protocolos de comunicación 44
3.4. Medios de transmisión de datos 47
3.5. Las redes de computadores en la legislación 50
3.5.1. Normas y reglamentos para el correcto funcionamiento de las redes de computadores 50
3.5.2. Neutralidad en la Red 52

4. Redes TCP/IP **55**

4.1. Propósito del protocolo 56
4.2. Direcciones IPV4 e IPV6 60
4.3. La dirección IP como dato personal 67
4.4. Ruteo de datagramas 70
4.5. Registro de nombres de dominio 73
4.6. Conectándose a Internet 78
4.6.1. Dial Up 78
4.6.2. ADSL 79
4.6.3. Conexiones inalámbricas 80
4.6.4. Los Internet Service Providers (ISP) 83
4.7. Institucionalidad jurídica de Internet y construcción de protocolos . . . 84

5. Internet y la Web **89**

5.1. Origenes de la NET 90
5.2. La Web 94
5.3. Datos originados en la Web 96
5.4. Otros datos presentes en la Web 101
5.5. Posibilidades de acceso a datos personales en la Web 106

6. Minería de datos de la Web **111**

6.1. Limpieza y preprocesamiento de los web data 112

6.2. Técnicas, algoritmos y métodos usados en web mining 115

6.3. Análisis de la operación de las técnicas de minería de datos 119

6.3.1. Procesamiento de los registros de web log 120

6.3.2. Procesamiento de los contenidos en una página web 121

6.3.3. Procesamiento de la estructura de hipervínculos 123

6.3.4. Análisis de la operación de los sistemas de recomendación 124

6.4. Privacidad y libertad en la navegación desde la perspectiva del web mining 126

7. Aspectos jurídicos del tratamiento de web data **133**

7.1. Marco legal para el análisis de la vida privada en el Web Mining 134

7.2. Privacidad y libertad de navegación en la personalización de la Web . . 141

7.2.1. Comentarios finales 143

7.3. Regulación del derecho al honor, la honra y web mining 146

7.4. Web Mining y legislación sobre delitos informáticos 148

7.4.0.1. Ámbito Nacional. 149

7.4.0.2. Ámbito Internacional. 151

7.5. El Web mining frente a los principios generales del derecho 152

8. Conclusiones **155**

Glosario **167**

Bibliografía **168**

Índice de figuras

2.1. Onda Senoidal 11
2.2. Señal de voz humana 12
2.3. Transformación de una señal de voz en frecuencia 13
2.4. Modelamiento de Señales 14
2.5. Modulación en frecuencia 16
2.6. Modulación en amplitud 17
2.7. Espectro Electromágnetico y Radioeléctrico 19
2.8. Teorema del muestreo 20
2.9. Análogo a digital 21
2.10. Conversión de señal analógica a señal digital 22
2.11. Muestreo básico de una señal senoidal 24

3.1. Bus de datos 33
3.2. Bus estrellado 34
3.3. Token Ring 35
3.4. Configuración clásica de una red Ethernet 36
3.5. Equipos de interconexión usados en la creación de redes 39
3.6. Funcionamiento del Hub 39
3.7. Funcionamiento del Switch 40
3.8. Funcionamiento del router 41
3.9. Modelo de referencia OSI para interconexión de computadores. 43

3.10. Ciclo básico de la comunicación entre dos entidades 45
3.11. Medios de Transmisión de datos. 47

4.1. Protocolo TCP/IP. Elaboración propia. 58
4.2. Interacción de equipos de interconexión e Internet. 61
4.3. Datagrama en versión IPV4. 62
4.4. Tipos de datagramas en IPV4 64
4.5. Datagrama de dirección IP en versión 6. 66
4.6. Funcionamiento del protocolo TCP/IP 72
4.7. Jerarquía de un DNS 75
4.8. Total de nombres en dominio .cl 76
4.9. Envio de un mail vía SMTP 77
4.10. Funcionamiento del Dial Up. 79
4.11. Funcionamiento de ADSL. 80
4.12. Proveedores de Internet en USA 83
4.13. Servicios de conexión que ofrecen distintos ISP en Chile. 84

5.1. Impacto de la falta de Internet si se detuviera 3 días 90
5.2. Red de computadores 91
5.3. Curvas de Adopción de Medios 95
5.4. Modelo básico de operación de la Web 97
5.5. Estructura de un web log file 98
5.6. Sitios Web según estructura de links 101
5.7. Conexión de red LAN con WAN a través de router 103

Índice de tablas

2.1. Algunas medidas de almacenamiento de información 24

2.2. Ejemplo de representación binaria de una señal muestreada 25

3.1. Medios de Transmisión. Elaboración basada en [24] 49

6.1. Mecanismos para identificación de sesiones 120

Resumen

El objetivo de esta tesis, es analizar hasta qué punto las técnicas, algoritmos y métodos comprendido en las herramientas de web mining, aplicados sobre los datos originados en la Web (web data), vulneran las garantías fundamentales, especialmente la libertad, la privacidad y el honor de las personas.

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 primero está relacionado con los contenidos, principalmente textos, que los usuarios o terceros incluyen en las páginas web. El segundo es referente a la estructura de hipervínculos que posee un sitio y el tercero dice relación con los datos de navegación que se almacenan en archivos de web log que posee un sitio. 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.

Dos posiciones históricamente contrapuestas se ven enfrentadas nuevamente en web mining. Por un lado la ciencia busca una aproximación de una verdad que permita un mayor conocimiento del fenómeno Web, y por otro lado las empresas quieren usar este conocimiento para aumentar su participación en el mercado digital. Entonces el procesamiento de los web data se ve expuesto a una serie de interrogantes, siendo las abordadas en esta tesis las que dicen relación con la afectación a la libertad de navegación y vulneración de la privacidad de los usuarios.

Durante este trabajo de tesis, se ha realizado una profunda revisión científico técnica de los fundamentos de las telecomunicaciones, el desarrollo de las redes de computadores para luego revisar los orígenes de Internet y la Web. Acto seguido se analizaron 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, especialmente la ley 19.628 sobre protección de la privacidad y la ley 18.168, de telecomunicaciones

Más importante aun que pensar en redactar leyes y regulaciones específicas para el web mining, es necesario crear conciencia en los científicos y profesionales que desarrollan y aplican web mining, de la importancia de realizar un procesamiento de datos que no implique necesariamente la identificación de los usuarios que visitan un sitio, lo cual es totalmente factible y no necesariamente significaría una merma en las utilidades. Adicionalmente, la creación de sistemas que personalicen la experiencia de visitante de un sitio, al menos deben dar la opción para que no se restrinja la posibilidad de apreciar otras páginas, transformando la imposición de ver un contenido, en una simple sugerencia que el usuario debe finalmente decidir si acepta o no.

La aplicación del web mining potencialmente puede atentar contra la privacidad y libertad de navegación de los usuarios que visitan un sitio, por lo cual su uso debe ser realizado considerando buenas prácticas en la preparación de los datos, la creación de perfiles de usuario, la identificación de personas y la generación de sistemas de recomendación en línea de lo que debe o no visitar el usuario. Lo anterior no requiere de una normativa específica, sino más bien de crear conciencia en lo referente a lo perjudicial que pueden llegar a ser estas herramientas si se hace un mal uso de ellas.

Capítulo 1

Introducción

The journey of a thousand miles begins with one step.
Lao Tzu

La World Wide Web o simplemente **La Web** [4] 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 [22].

La Web es considerado un canal de difusión e intercambio de información a escala global. Esta realidad, ha incentivado a muchas compañías a replantear cómo enfrentar sus negocios de cara a las nuevas condiciones que impone el mercado digital, siendo el sitio web corporativo la nueva "*tarjeta de presentación virtual*" con que muestran información a todo aquel que visite el sitio [57].

Para muchas compañías e instituciones, ya no es suficiente contar con un sitio web que sólo muestre información respecto de los productos que ofrece. La verdadera diferencia, que puede provocar una ventaja competitiva en el mercado digital, está dada por el potencial que tiene el sitio para atraer y retener a sus visitantes. Dicho potencial está determinado por el contenido, estructura y diseño del sitio, además de otros aspectos técnicos como lo es el tiempo de respuesta frente a similares requerimientos,

comparados con otros sitios.

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 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 [25].

La personalización implica que de alguna forma se puede obtener información respecto de los deseos y necesidades de las personas [2], para luego preparar la oferta correcta en el momento correcto [26]. 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 [8].

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, vídeos, 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.

En el artículo 2^o letra *"f"* de la citada ley, se define a los datos personales como los *"los relativos a cualquier información concerniente a personas naturales, identificadas o identificables"*. Analizando la naturaleza de los web data, se desprende que es posible identificar aproximadamente o en su totalidad al usuario humano responsable de una navegación. En efecto, utilizando la la dirección IP de acceso, más la combinación de otros datos como el tipo de navegador utilizado y contenidos relativos a las preferencias particulares de un usuario se puede lograr una muy buena aproximación a su identificación y en muchos casos, la certeza absoluta. En consecuencia, los web data entran, aunque sólo una parte de ellos, en la categoría de datos personales.

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 mencionado artículo.

Por último, el procesamiento de los web data, como se verá en el capítulo 6 utilizando web mining, está comprendido dentro de la definición que la ley establece, ya que corresponden a *"cualquier operación o complejo de operaciones o procedimientos técnicos, de carácter automatizado o no, que permitan recolectar, almacenar, grabar, organizar, elaborar, seleccionar, extraer, confrontar, interconectar, disociar, comuni-*

car, ceder, transferir, transmitir o cancelar datos de carácter personal, o utilizarlos en cualquier otra forma".

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 afín de obtener resultados en cuestión de horas o días. Lo anterior permite que utilizando computadoras del alto rendimiento, se puedan realizar una cantidad importante de operaciones y correlaciones que a la postre entregarán información y conocimiento que no se puede obtener a partir con la estadística o el simple análisis ocular. Dicho de otra forma, se puede llegar a niveles de extrapolación de información que vulneren directamente la privacidad de los usuarios y que pueda ser utilizado para coartar su libertad de navegación en la Web.

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? [6, 18]. 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 y de paso coartar su libertad de visitar los contenidos y sitios que les plazca, sin que estos se den cuenta de que están siendo vigilados por un "*gran hermano*" cibernético [25]. 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 puede caer en excesos que atentan contra los derechos fundamentales de las personas [50].

1.1. Hipótesis de Investigación

Las técnicas, algoritmos y herramientas utilizadas en el análisis del comportamiento del usuario en la Web agrupadas en la disciplina conocida como web mining,

pueden ser usadas inapropiadamente para vulnerar la privacidad de la persona que navega en un sitio o coartar su libertad para acceder a la información ahí contenida.

1.2. Objetivos Generales y Específicos

Objetivo General. Analizar hasta qué punto el uso de las técnicas, algoritmos y metodologías presentes en web mining restringen la libertad de navegación del usuario y vulneran su privacidad durante su visita a un sitio.

Objetivos Específicos:

1. Establecer una base técnica conceptual que permita comprender el funcionamiento de las Tecnologías de Información y Comunicaciones relacionadas con el funcionamiento de Internet y la Web.
2. Analizar el funcionamiento de Internet y la Web, del punto de vista de los datos que se crean, transmiten y procesan, con especial énfasis en aquellos que dicen relación con el comportamiento de navegación de los usuarios.
3. Revisar la regulación nacional e internacional, respecto de los datos originados en la Web, en particular, aquellos que se relacionan con los usuarios humanos que visitan un sitio web.
4. Estudiar la operación de las técnicas, métodos y algoritmos de minería de datos que se utilizan en el análisis del comportamiento del usuario en la Web para establecer su impacto en la restricción de la libertad de navegación de los usuarios y vulneración de su privacidad.
5. Establecer un conjunto de *"buenas prácticas"* en el minado de datos originados en la Web, que permita el estudio científico de los fenómenos ahí presentes, salvaguardando la privacidad de los usuarios y estableciendo condiciones necesarias para la restricciones a la libertad de navegación.
6. Desarrollar conciencia en los encargados de la creación de las futuras leyes, normativas, etc. sobre el tratamiento de datos originados en la Web, respecto del impacto positivo/negativo que hay detrás del uso de TIC avanzadas como lo es el web mining.

1.3. Metodología

Para establecer una base técnica teórico/práctica respecto del funcionamiento de las TICs relacionadas con Internet y la Web, se revisará literatura y apuntes de cursos sobre telecomunicaciones, redes de computadores, teoría de la señal, electromagnetismo y los fundamentos del funcionamiento del protocolo TCP/IP y HTTP.

A partir del análisis del funcionamiento de Internet y de la Web, se establecerán cuáles son los posibles datos originados por los usuarios que podrían ser usados en un proceso de extracción de conocimiento usando técnicas de web mining. Con esta información, será posible establecer un paralelo respecto de la regulación nacional e internacional que diga relación con los web data.

Es importante establecer un marco de referencia para revisar la operatoria de los algoritmos, técnicas y métodos considerados en web mining para luego indagar su verdadero impacto en la restricción de navegación de los usuarios y afectación de su privacidad. De esta forma, luego de una revisión bibliográfica profunda, se acudirá a la consulta de expertos en el área del web mining, para desentrañar la forma y el fondo de su operación.

Ya clarificados qué se entiende por web data y la operatoria de los algoritmos, técnicas y métodos presentes en web mining, es posible establecer un conjunto de buenas prácticas que permita a los usuarios salvaguardar sus derechos a la libre navegación y privacidad de sus actos, y por otra parte a los científicos, analistas, dueños de sitios web, etc. poder extraer el tan necesario conocimiento para mantener su oferta de productos, servicios e información en general, siempre atractiva para sus eventuales visitantes y clientes. Dichas buenas prácticas serán contrastadas con la experiencia de los expertos y con lo que se ha desarrollado en otras latitudes en lo referente al desarrollo de la nueva generación de sitios web.

Finalmente, siendo uno de los objetivos de esta tesis la generación de conciencia respecto de las implicancias del web mining en el desarrollo futuro de una sociedad

del conocimiento altamente influenciada por las TICs, se procederá a la publicación y presentación de los resultados de este trabajo en conferencias nacionales y en la publicación de artículos científicos.

1.4. Contribuciones de la Tesis

Los capítulos de la presente tesis, han contribuido a la generación de nuevos conocimientos:

- Publicación en revista nacional. J.D. Velásquez y Lorena Donoso, Web Mining: Análisis sobre la privacidad del tratamiento de datos originados en la Web, Revista Ingeniería de Sistemas, 23(1):5-26, 2009.
- Publicación en revista nacional. J.D. Velásquez y Lorena Donoso, 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, Revista Ingeniería de Sistemas, 24(1):27-49, 2010.
- Apuntes sobre introducción a las Tecnologías de Información y Comunicaciones para el Magister en Derecho de la Informática y de las Telecomunicaciones de la Universidad de Chile

1.5. Organización de la Tesis

El propósito principal de esta tesis es proveer una introducción a un tema que ha sido tímidamente abordado a nivel mundial, como lo es hasta qué punto se afectan las libertades y se vulnera la privacidad de lo usuarios que navegan en la Web, a partir de la aplicación de técnicas, algoritmos y metodología de web mining, sobre los datos que se originan durante cualquier visita a un sitio.

El capítulo 2 aborda el fenómeno de las telecomunicaciones, desde sus fundamentos teórico/prácticos hasta la regulación que históricamente se ha ido tejiendo entorno a su desarrollo tecnológico.

Las redes de computadores y la legislación que existe al respecto, son revisadas en el capítulo 3, lo cual sirve de antesala para analizar el fundamento del protocolo de comunicaciones TCP/IP usado en Internet, en sus versiones 4 y 6, lo que se estudiará en el capítulo 4.

El capítulo 5, 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.

Los capítulos anteriores han servido de fundamentación teórica/practica y legal para poder analizar 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, lo cual se aborda en el capítulo 6.

Por su parte, el capítulo 7 profundiza en los aspectos jurídicos relacionados con el tratamiento de los datos originados en la Web, también conocidos como Web Data.

Finalmente, en el capítulo 8 se presentan las principales conclusiones y recomendaciones que se han obtenido a lo largo de la tesis.

Capítulo 2

Orígenes de las telecomunicaciones

En el comienzo, Dios creo los cielos y la tierra.
(Genesis 1:1)

Desde que el hombre comenzó a organizarse en torno a comunidades, se hizo necesario establecer un mecanismo que permitiera comunicar ideas, el cual en un principio fue una combinación de realizaciones guturales y gestuales, para posteriormente configurar un lenguaje hablado complejo. El paso siguiente fue la codificación del lenguaje en una representación estandarizada, para una comunidad, que permitiera la expresión de ideas, con prevalecencia en el tiempo. Nació entonces la escritura.

La transmisión de ideas entre las personas, ya sea a través de gestos, signos, expresiones orales y escritas, es lo que se entiende por una comunicación. Si adicionalmente se agregara que esta interacción se realiza a través de medios *"telegráficos, telefónicos o radiotelegráficos y demás análogos"* entre dos lugares geográficamente separados, se estará en presencia de una telecomunicación [16].

En esencia, lo que posibilita que se produzca una telecomunicación, es que de alguna forma, entre el emisor y el receptor, existe una señal que es capaz de transmitirse por un medio, y en cuyas características va adosada la información que se desea enviar.

Aparte de ciertos casos donde existe un dueño definido, la gran mayoría de los medios de telecomunicaciones utilizados tienen una característica intrínseca de bien público escaso, por lo que su uso siempre tendrá una alta demanda, lo cual motiva una regulación que permita el avance en el desarrollo científico y tecnológico asociado, pero salvaguardando las garantías fundamentales consagradas en las leyes y la constitución de cada país. En este sentido, el artículo 7^o inciso primero de la Ley General de Telecomunicaciones 18.168 establece que le corresponde al Ministerio de Transportes y Telecomunicaciones velar 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"*. Adicionalmente, en el mismo artículo inciso 2^o le entrega atribuciones a dicho ministerio para *"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"*.

2.1. Concepto de señal

En su definición más amplia, se entiende por señal a la medida de una magnitud física que caracteríza la evolución en el tiempo de un sistema real [52].

Las señales aportadas por un sistema físico puede ser de variada naturaleza: eléctrica, química, acústica, óptica, mecánica, etc. No obstante el desarrollo tecnológico hace preferible el formato eléctrico para los procesos de codificación y transmisión de la información.

De manera formal, una señal es función de valores reales o complejos que dependen de una variable real *"t"*, que usualmente se identifica con el tiempo. La Fig. 2.1 muestra la representación gráfica de una señal de tipo sinusoidal $a(t)$ y de período

T, es decir, la forma de la señal se repetirá identicamentente cada intervalo de tiempo T, o dicho de otra forma, su frecuencia de repetición estará dada por $\frac{1}{T}$. Otra característica es su amplitud A_0 y la fase β que marca el comienzo en la gráfica.

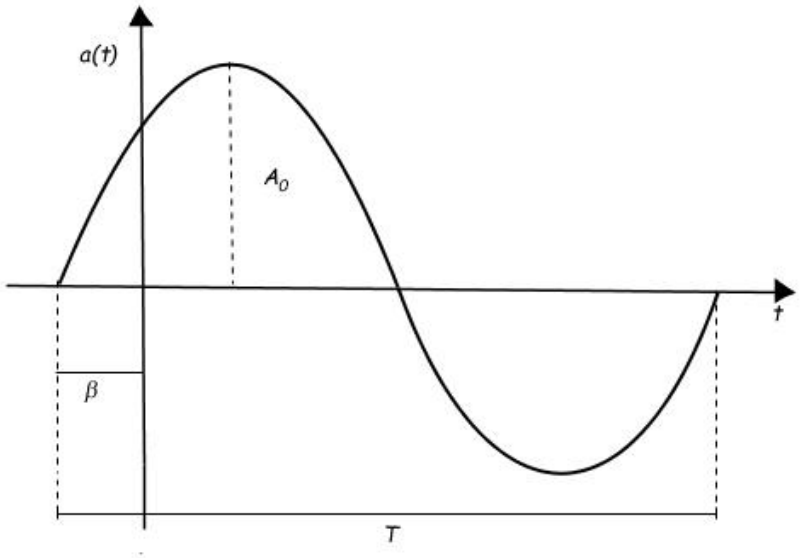

Figura 2.1: Onda Senoidal

Si se desea transmitir la señal de la Fig. 2.1 hacia un receptor, no es necesario enviar todas sus infinitas variaciones en el tiempo, sino que sólo cuatro parámetros: que es una sinusoide, la amplitud A_0, la fase β y el período T. Su modelamiento como expresión matemática es $a(t) = A_0 \sin(\omega t)$.

Como es evidente, no todas las señales pueden ser representadas de una forma tan simple. La Fig. 2.2 se muestra un trozo de una señal de voz humana. Su representación matemática no es posible de obtener con la mera observación. Sin embargo, basándose en la teoría de la señal, se puede desarrollar un modelo aproximado a través de la superposición (matemáticamente como una sumatoria) de varias señales simples.

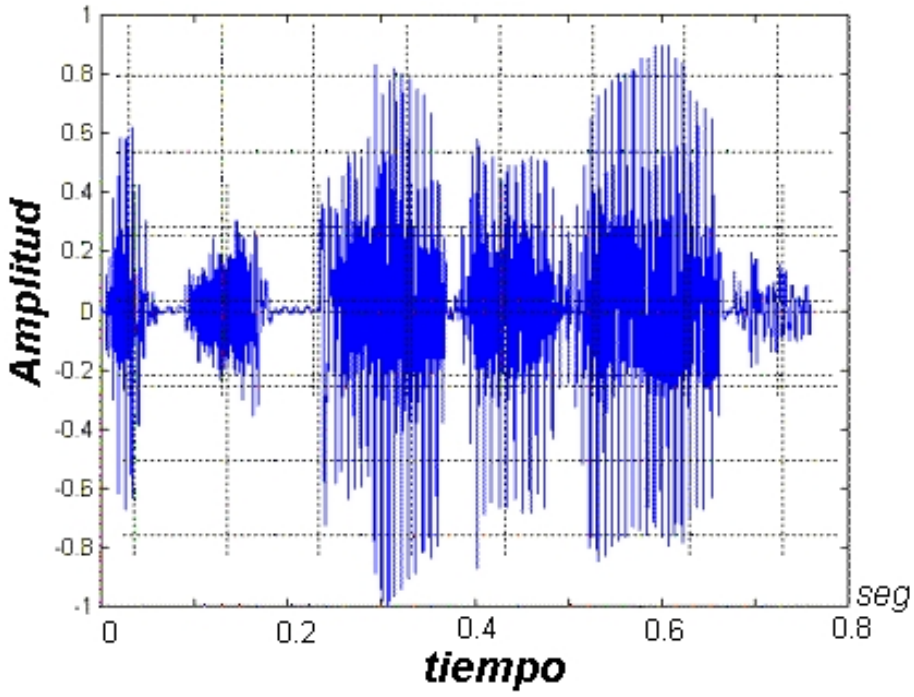

Figura 2.2: Señal de voz humana

La Eq. 2.1 muestra una forma de representar una señal de voz

$$v(t) = g_1(t) + g2(t) + \cdots + g_n(t) \tag{2.1}$$

donde $g_i(t),\ i = 1 \ldots n$ son señales de la forma $g_i(t) = A_i \sin(t + \beta_i)$.

Cada una de estas señales $g_i(t)$ posee su propio período, por lo que la representación de la voz sería en multiples frecuencias. Sin embargo, existe una frecuencia que se denomina *"fundamental"* y que es la que concentra la mayor energía de la señal.

Las operaciones matemáticas de la señal de la Eq. 2.1, en el dominio del tiempo, son de una complejidad mayor, por lo que es preferente trabajar con su representación matemática en frecuencia. La Fig. 2.3 muestra la transformación al dominio de las frecuencias de la señal de voz presentada en la Fig. 2.2. Se puede apreciar que en torno a los 300 Hz se encuentra el máximo de energía de la señal. Luego se aprecian otros máximos locales, que representan las frecuencias de las señales menores, conocidos

como *"armónicos"*.

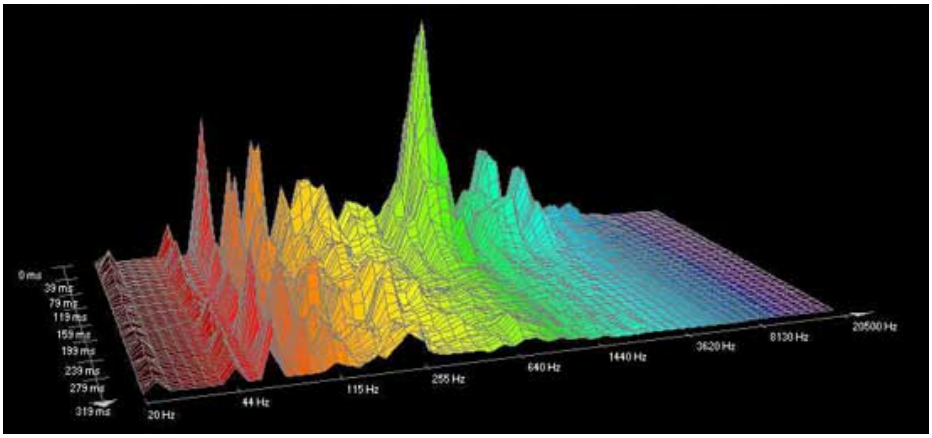

Figura 2.3: Transformación de una señal de voz en frecuencia

De la Fig. 2.3 se desprende, además, que la señal está restringida entre las frecuencias 50 y 3200 Hz, situación que podría variar en la medida que su potencia aumentara, con lo cual se crearía interferencia a las otras señales de las frecuencias vecinas.

La transmisión de una señal por un medio, sufre pérdida de su energía, por lo que se atenúa en su forma y varían algunos de sus parámetros, lo que se conoce como introducción de ruidos. En este caso, lo que está ocurriendo es que cada medio posee una respuesta en frecuencia que le es propia, de forma que la señal transmitida a través de él, sufre distintas atenuaciones en función de la frecuencia. En consecuencia, un determinado medio tiene sólo un rango de frecuencias por las cuales una señal con una frecuencia comprendida dentro de este rango, puede ser transmita y sufrirá la misma atenuación en los armónicos que se encuentren dentro del rango de frecuencias del canal.

Se define como ancho de banda de un canal, al rango de frecuencias donde los armónicos de una señal sufren la misma atenuación durante su transmisión. Entonces, este ancho de banda queda representado por la diferencia entre la frecuencia superior

e inferior que se puede transmitir con atenuación, pero sin distorsión por un medio físico empleado como canal de comunicación.

2.2. Modelamiento de señales

Las señales se pueden representar como como ondas, las cuales se transmiten por un medio material (mecánicas) y las que lo hacen por el vacío (electromagnéticas). Son estas últimas las que al no necesitar un medio de transmisión, alcanzan la velocidad de la luz en su propagación, es decir, 300.000 $\frac{km}{seg}$.

Como ya ha quedado consignado en la sección anterior, señales como la mostrada en la Fig. 2.4 se pueden representar por parámetros tales como amplitud, frecuencia y longitud de onda.

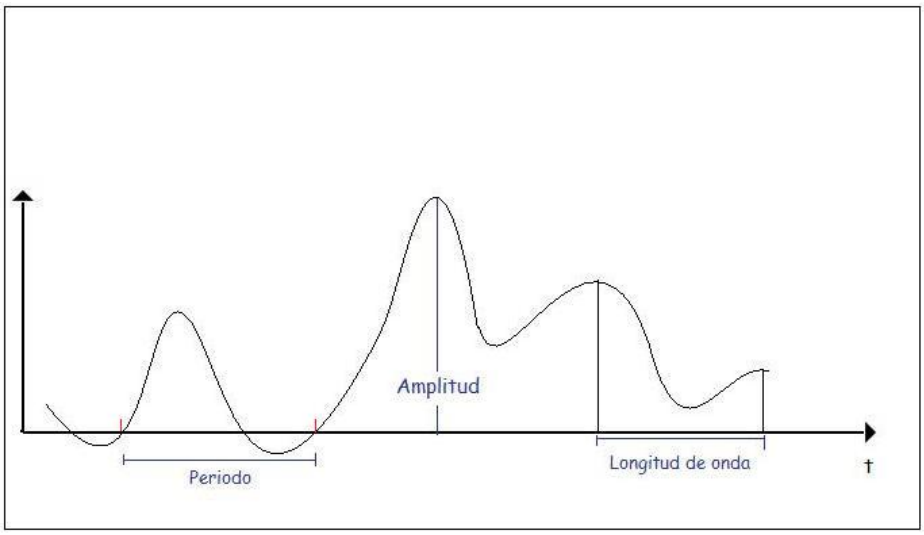

Figura 2.4: Modelamiento de Señales

La **amplitud de onda** corresponde a la altura maxima que puede alcanzar una señal, medida desde el origen hasta el punto mas alto en sentido vertical. La unidad de medida preferente de la amplitud de onda es Volt.

La **frecuencia** es el número de repeticiones de una onda en una unidad de tiempo. El número de ciclos por segundo se mide en Hertz (Hz). Para obtener la frecuencia, se puede dividir una unidad por el período de la onda.

El oído humano es capaz de percibir frecuencias entre 20 y 20.000 Hz, aunque con la edad esta capacidad va disminuyendo. Esta respuesta en frecuencia se conoce como *"audiofrecuencia"*, y es el umbral con que el ser humano puede percibir una onda sonora, es decir, todo aquel estimulo acústico fuera de este rango, no lo capta el oído de una persona, pero si es factible que lo haga una especie inferior como un perro o un gato, cuyos sistemas auditivos, producto de la evolución, pueden captar señales en frecuencias más altas.

A partir del uso de equipos de medición de señales, es posible analizar en qué frecuencia son transmitidas, lo que permite configurar un verdadero mapa por donde se puede apreciar donde viajan las ondas. A este mapa se le denomina *"espectro de electromagnético"* y conceptualmente es infinito.

Por otro lado, a longitud de onda corresponde a la distancia que recorre la onda en el intervalo de tiempo entre dos maximos seguidos, como se aprecia en la Fig. 2.4.

Finalmente, ruido en las telecomunicaciones es todo aquello que modifica el contenido de información de una señal.

2.3. Transmisión de señales

La transmisión de una señal que ya se sabe cuál es su estructura, se realiza provocando variaciones en esta que luego pueden ser identificadas en el receptor. A este proceso se le conoce con el nombre de *"modulación"* y puede ser realizada en: frecuencia, amplitud y fase.

Para que la modulación se lleve a cabo, se necesita una onda portadora y una moduladora. La información es contenida por la onda portadora, la cual es modificada o

modulada según la característica (fase, frecuencia o amplitud) de la señal moduladora para luego transmitirla al receptor.

Cuando se modula en frecuencia, se modifica la frecuencia de la onda portadora (la amplitud se mantiene constante) para alcanzar la de la onda moduladora, resultando lo que se puede observar en la Fig. 2.5.

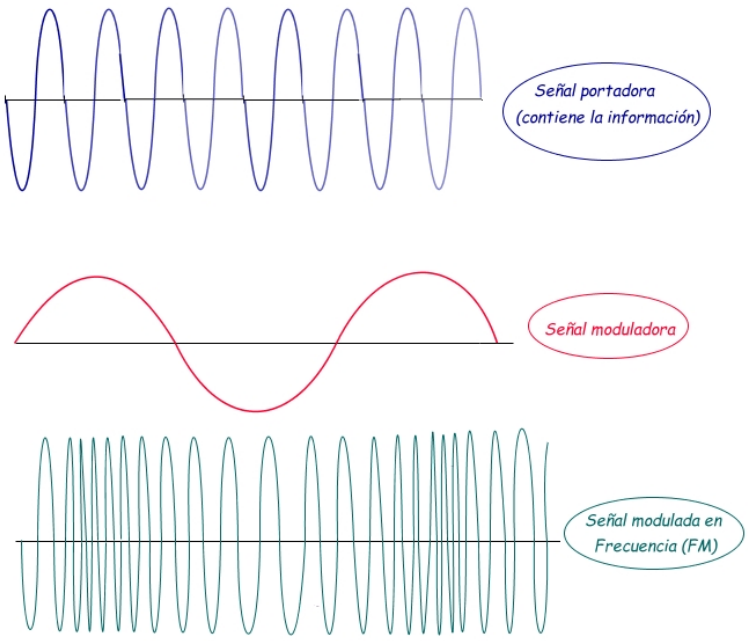

Figura 2.5: Modulación en frecuencia

En cambio si se modula en amplitud, se mantiene constante la frecuencia y se varia la amplitud, como se muestra en la Fig. 2.6.

La modulación menos común es la de fase, la cual requiere de un tratamiento más complejo y por ende del uso de equipos más sofisticados.

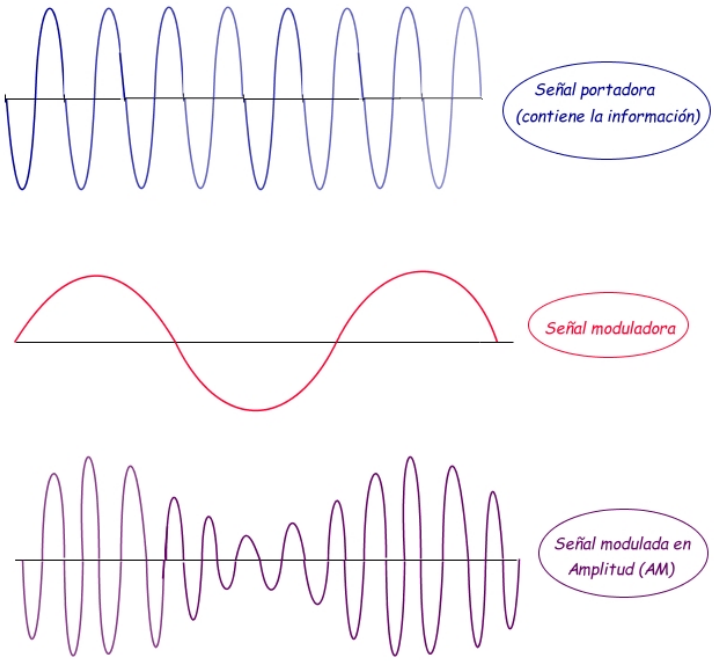

Figura 2.6: Modulación en amplitud

2.4. Espectro radioeléctrico

El espectro radioeléctrico es el espacio que permite la propagación sin guía artificial de ondas electromagnéticas cuyas bandas de frecuencias se fijan por convención [17], la cual es dictada y regulada por la Unión Internacional de Telecomunicaciones (UIT), institución ubicada en Ginebra, Suiza, compuesta por 191 países y más de 700 organizaciones asociadas.

El espectro radioeléctrico posee caracter de recurso natural limitado sobre el cual cada Estado ejerce soberanía, e involucra todo tipo de frecuencias y órbitas asociadas (incluidos los satélites geoestacionarios). Por lo tanto su uso deber ser en forma racional, eficaz y económica.

Del punto de vista jurídico, el espectro radioeléctrico se conceptualiza como un *"bien nacional de uso público"*, los cuales son definidos por el Código Civil en su artículo 589 como *"aquellos cuyo dominio pertenece a la nación toda. Si además su uso pertenece a todos los habitantes de la nación, como el de calles, plazas, puentes y caminos, el mar adyacente y sus playas, se llaman bienes nacionales de uso público o bienes públicos"*.

El derecho civil considera que el uso de los bienes nacionales de uso público *"pertenece a todos los habitantes de la nación"*. En algunos casos, el uso de estos bienes esta sujeto a una regulación que otorga permisos o concesiones por parte de alguna autoridad. En el caso del espectro radioeléctrico, la ley General de Telecomunicaciones, en su artículo 6^o le otorga al Ministerio de Transportes y Telecomunicaciones, a través de la Subsecretaría de Telecomunicaciones (Subtel) la facultad de entregar dichas concesiones y permisos. Dentro del mismo articulo se le entregan atribuciones a la Subtel para que en forma exclusiva realice la *"interpretación técnica de las disposiciones legales y reglamentarias que rigen las telecomunicaciones"*.

El espectro electromagnético está compuesto por longitudes de onda que van desde frecuencias inferiores a 30 kHz a superiores a 30 EHz [1]. Sólo una parte de estas pertenecen al espectro radioeléctrico, la cual comprende frecuencias de onda de 153 KHz a los 300 GHz aproximadamente. Tal como podemos observar en la Fig. 2.7, se incluyen en esta categoría a frecuencias moduladas FM y frecuencias amplificadas AM (radios), televisión, microondas, telefonía inálambrica y radar.

2.5. Análogo a digital

En su definición básica, la señal analógica caracteriza la evolución en el tiempo de un fenómeno medible. Matemáticamente se representa a través de una función continua. Prácticamente todas las señales que el ser humano logra percibir son de

[1] Exahertz, que corresponde a $10^{18} Hz$

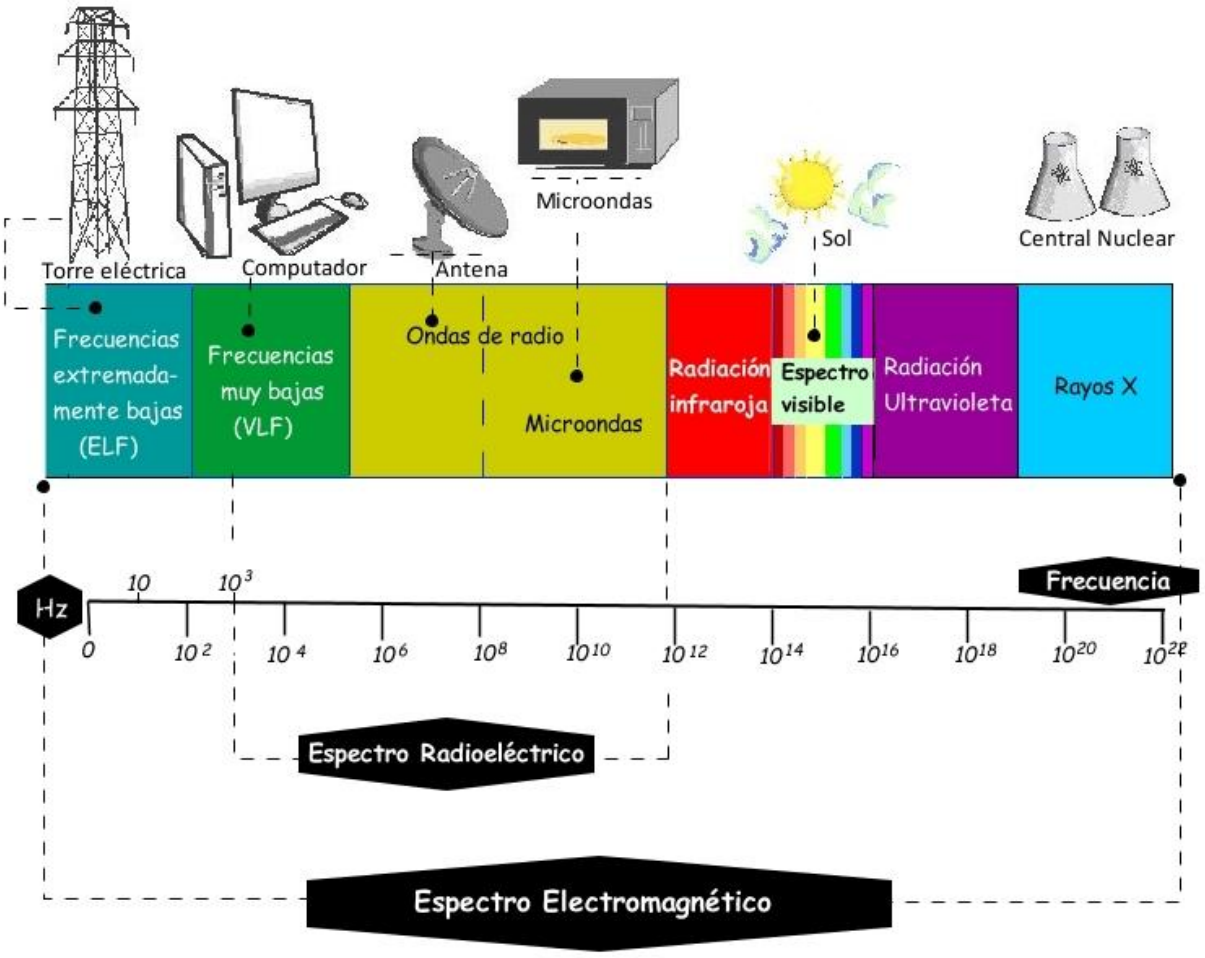

Figura 2.7: Espectro Electromágnetico y Radioeléctrico

naturaleza análoga, donde la continuidad en la forma es la característica fundamental. Por ejemplo, al percibir la luz, aunque esta se descomponga como es el caso del arcoiris, de todas maneras se aprecia una continuidad en la recepción, es decir, no hay sobresaltos.

Este tipo de señales, está muy afecto a ser contaminada por ruidos provenientes de otras señales, lo que las degrada en el tiempo y en la distancia que deben recorrer para llegar a un receptor. Por ejemplo, nótese la recepción que se tiene cuando se habla por un walkie talkie. Este dispositivo no sólo recepciona la señal que ha enviado un emisor, sino que también acopla otras comunicaciones.

En transmisión de datos, el uso de una señal análoga no es eficiente, por el alto ruido que esta posee y lo difícil que en general es su manipulación. En este sentido, la digitalización de la señal, es decir, su transformación a un formato binario, ha sido la solución para su transmisión, minimización de la distorsión y atenuación.

La digitalización de una señal, está basada en la teoría de la información desarrollada, entre otros, por Harry Nyquist en 1928 [38] [2] quien formuló el teorema que es la piedra angular de las telecomunicaciones: El teorema del muestreo, el cual fue demostrado formalmente por Claude E. Shannon en 1949 [44] [3]. El teorema afirma que una señal analógica, puede ser reconstruida, sin error, a partir de muestras tomadas en intervalos de tiempos iguales, siendo el único requisito que el muestreo se debe ser igual o superior al doble del ancho de banda de la señal a digitalizar, como se muestra en la Fig. 2.8.

Figura 2.8: Teorema del muestreo

A modo de ejemplo, si se desea digitalizar una señal de voz humana, cuyo ancho de banda es 4.000 [Hz] aproximadamente, entonces, la tasa de muestreo, también conocida como *"Frecuencia de Nyquist"*, debe ser igual o superior a 8.000 muestras por segundo ($\frac{1}{8000}$).

Mientras más alta sea la tasa de muestreo aplicada, mayor será la información

[2] Certain topics in telegraph transmission theory
[3] Communication in the presence of noise

que se puede almacenar o transmitir, con lo que se puede reconstruir la señal analógica con mayor calidad. Sin embargo, existen señales, como las acústicas, que no requieren de tasas muy altas de muestreo, por cuanto sus usuarios finales son personas, cuyos aparatos auditivos sólo pueden detectar frecuencias de 20 Hz a 20 KHz, con lo que es suficiente un muestreo de al menos 40 KHz para reconstruir la señal de sonido aceptable al oído humano.

Una vez que se han obtenido las muestras, corresponde su codificación para ser transmitida a través de un medio. En el receptor, se realiza el proceso inverso, es decir, se pasa de la representación binaria a la analógica reconstruyendo una señal muy parecida a la original. En la Fig. 2.9 se puede observar el cambio en la señal al pasar de análogo a digital, resultando una función discreta. Las etapas consideradas en este proceso son: Muestreo, Retención, Cuantificación y Codificación.

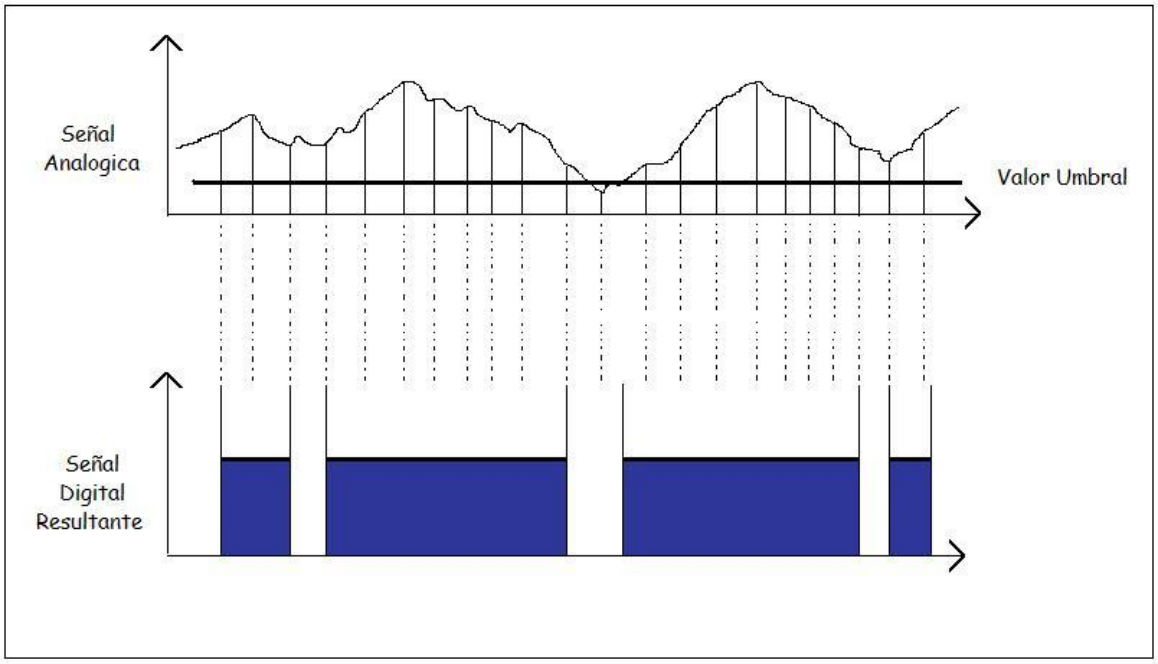

Figura 2.9: Análogo a digital

Por su naturaleza, la señal digital no es continua, por lo que su representación

matemática se hace a partir de elementos discretos.

Una vez que la señal es transferida mediante una red, se convierte la señal digitalizada a analógica para su recepción final. En la Fig. 2.10 se ejemplifica la conversión que experimenta una señal producida por una persona haciendo karaoke, para ser transferida por un sistema MIC (Modulación por Impulsos Codificados), que luego será reconvertida a analógica para ser amplificada.

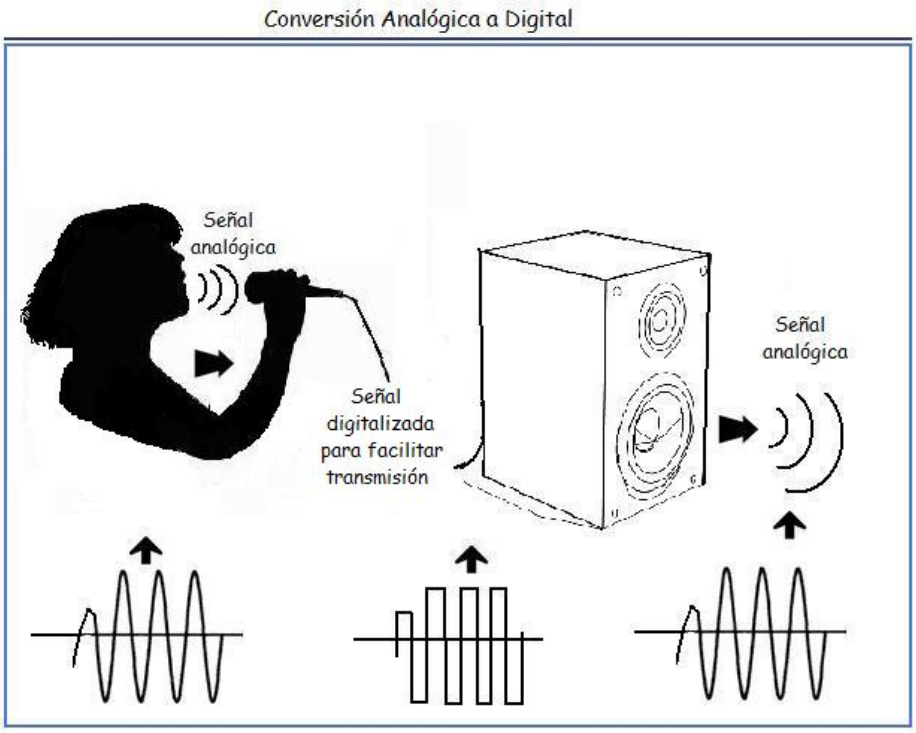

Figura 2.10: Conversión de señal analógica a señal digital

En transmisión de datos, ya no se puede pensar en señales analógicas por cuanto su alto ruido y problemas anexos las hacen poco eficientes, con mucho retardo y una alta probabilidad de pérdidas. Se podría decir que estamos en el ocaso del mundo análogo y el digital ya es una realidad.

2.6. Representación binaria

La transmisión de una señal análoga que ha sido previamente muestreada y codificada, requiere de una representación que permita, con muy poco esfuerzo, la detección fidedigna de lo que se está transmitiendo.

La codificación binaria, es la transformación de un valor numérico en una secuencia estados que indican la ausencia o presencia de una medida, comúnmente asociada con el voltaje. En este sentido, el código "**0**" se asocia al 0[volt] y el "**1**" a los 5[volts], valores que son muy fáciles de monitorear y corregir en un sistema eléctrico. A estos símbolos se le conoce con el nombre de *"bits"*.

De la teoría de la información [44] se sabe que todo número puede ser codificado utilizando una base binaria. Por ejemplo el número 5, queda representado en base binaria en la Eq. 2.2

$$5 = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 \rightarrow \mathbf{101}. \tag{2.2}$$

con lo que lo que habría que transmitir una señal eléctrica que contenga 5, 0 y 5 volts. En este tipo de transmisión, también se agregan *"bits redundantes"* (checksum) que sirven para verificar si la transmisión fue recibida correctamente, situación análoga a la que sucede con el dígito verificador que se utiliza en la cédula de identidad.

El almacenamiento de la información binaria transmitida, se realiza a partir de grupos de bits que son gravados en algún medio, por ejemplo magnético. La tabla 2.6 muestra algunas las distintas unidades de almacenamiento de datos que se utilizan hoy en día.

El nivel de la digitalización de una señal análoga, estará siempre condicionada por la capacidad de almacenamiento y transmisión de los dispositivos electrónicos que estén involucrados. La Fig. 2.11 presenta una señal simple la cual es muestrada a una

Unidad	Tamaño	Almacenamiento aproximado
1 bit		Sólo almacena un estado
1 Byte	8 bits	Una letra
1 KB	1024 Bytes	Un párrafo de 10 líneas
1 MByte	1024 KBytes	Una novela
1 GBytes	1024 MBytes	Una película en baja resolución
1 TB	1024 GBytes	La biblioteca del Congreso de Chile
10 TB		La colección impresa de la biblioteca del congreso de EE. UU.

Tabla 2.1: Algunas medidas de almacenamiento de información

tasa muy baja, por lo que la codificación binaria necesaria no excede los 3 bits.

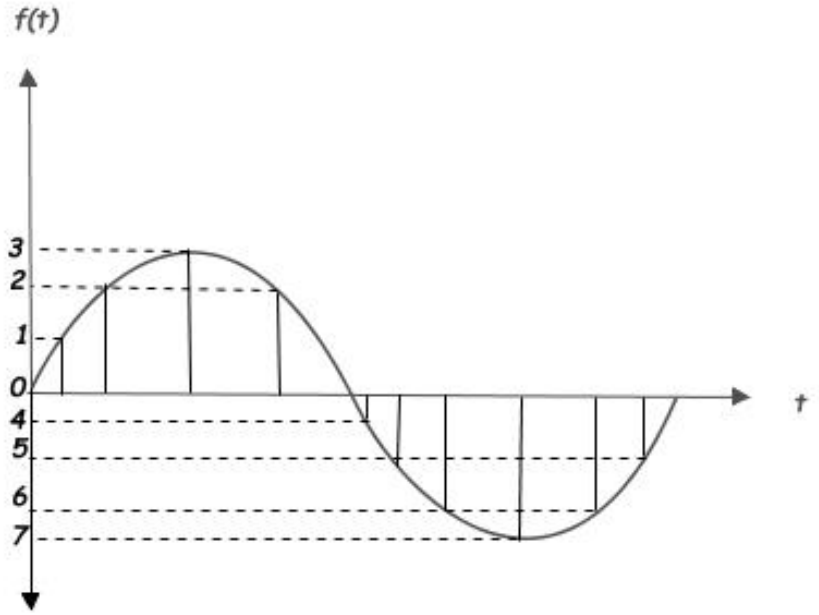

Figura 2.11: Muestreo básico de una señal senoidal

La codificación binaria para las 8 muestras extraídas de la Fig. 2.11 se presenta en la tabla 2.2.

Entonces, la información a transmitir será **000001010011100101110111**, lo

Medida	Representación binaria
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

Tabla 2.2: Ejemplo de representación binaria de una señal muestreada

que tendría su representación eléctrica como una secuencia de cero y cinco volts en un cable. Si el medio fuera fibra óptica, la transmisión se haría a través de un haz de luz, emitido por un diodo láser.

Desde el punto de vista de la transmisión digital de señales, el ancho de banda de un medio de comunicación se mide preferentemente según la cantidad de bits que pueden ser transmitidos en forma efectiva, es decir, que llegan finalmente a su destino y sin errores.

2.7. Naturaleza jurídica del servicio de acceso a Internet

La transmisión de señales a distancia ha sido y seguirá siendo motivo de la aplicación de leyes, normas y regulaciones varias. Lo anterior debido a que independiente del medio, siempre existe la posibilidad de vulnerar los derechos de otros por la operación de la actividad. En efecto, si se desea usar el espectro radioeléctrico, un bien escaso y que posee una alta demanda, se necesita regular. Lo mismo ocurre cuando se establecen canales de comunicación como por ejemplo cables, por cuanto se debe implementar una infraestructura que los soporte, como lo son los clásicos postes, estéticamente indeseables y que junto con el resto de artefactos necesarios para el

transporte de las señales, irrumpen en forma negativa en el medio ambiente.

Dada la importancia que tienen las telecomunicaciones para el desarrollo social y económico de un país, la regulación al respecto siempre ha ido en la línea de asegurar la interconexión, la interoperabilidad y la libertad de comunicación. Esto último está consagrado en el artículo 19 de la Constitución Política de la República de Chile, que en su número 12, asegura a todos los habitantes de esta tierra la *"libertad de emitir opinión y la de informar, sin censura previa, en cualquier forma y por cualquier medio, sin perjuicio de responder de los delitos y abusos que se cometan en el ejercicio de estas libertades, en conformidad a la ley, la que deber ser de quórum calificado"*.

Los conceptos anteriores forman parte del acerbo previo a la expansión de Internet y posteriormente la Web. Sin embargo, se encuentra una similitud muy interesante entre los principios que inspiran la legislación de las telecomunicaciones y los que hacen posible el funcionamiento de Internet y de la Web. En Internet, desde un comienzo existió la idea de *"la conectividad es un fin en si misma"*, es decir, se busca por todos los medios posibles que el dato enviado pueda ser encaminado en la red, lo cual exige que todos los equipos de interconexión aseguren que no habrá discriminación en los datos recibidos/enviados. También los obliga a someterse a estándares que aseguren la interoperabilidad de la gran red. Por último, no existe un medio más libre para la emisión de opiniones que la Web, de hecho su operación tiene su base en la idea fundacional de *"compartir información a escala global"*.

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 [11], 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.

La ley 18.168 define en su artículo 3^o los siguientes servicios de telecomunicaciones:

Servicios públicos de telecomunicaciones. Son aquellos destinados a satisfacer las necesidades de telecomunicaciones de la comunidad en general (telefonía fija, transmisión de datos, etc). Deben estar diseñados para interconectarse con otros servicios públicos.

Servicios intermedios de telecomunicaciones. Son los que satisfacen las necesidades de conmutación y transmisión de concesionarios o permisionarios de telecomunicaciones.

Servicios de telecomunicaciones de libre recepción o de radiodifusión. Son transmisiones destinadas a la recepción libre y directa del público en general. Estos servicios comprenden emisiones sonoras, de televisión o de otro género. Constituyen una subcategoría los servicios de mínima cobertura, que son los constituidos por una estación de radiodifusión cuya potencia no exceda de 1 watt.

Servicios complementarios. Se trata de 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.

Servicios limitados. Son los que satisfacen necesidades especificas de telecomunicaciones de determinadas empresas, entidades o personas. No permiten la interconexión con las redes públicas.

Servicios de aficionados a las radiocomunicaciones. Intercomunicación radial y experimentación técnica y científica sin fines de lucro.

En base a las definiciones anteriores, se puede argumentar que el acceso a Internet es un servicio complementario. Entonces, dado que la Web utiliza como soporte tecnológico y de transmisión de datos a las redes TCP/IP y en particular a Internet, se encuentra que los medios empleados para el envío/recepción de las señales portadoras son redes de telecomunicaciones, entonces, las comunicaciones a través de las cuales operan la Web e Internet responden al concepto de telecomunicación, en consecuencia la instalación, operación y explotación de los servicios que en ellos se preste ubicados en el territorio nacional, quedan regulados por la Subsecretaría de Telecomunicaciones (Subtel) como el organismo técnico, cuya finalidad principal es la aplicación y control de la ley y sus reglamentos. Adicionalmente, Subtel debe velar por que todos los servicios de telecomunicaciones, sistemas e instalaciones que generen ondas electromagnéticas, independiente de su naturaleza, sean instalados y operados de forma tal que no causen lesiones a personas o daños a otros artefactos, ni interferencias perjudiciales a los servicios de telecomunicaciones nacionales o extranjeros o interrupciones en su funcionamiento.

Entonces, si se considera que el acceso a Internet constituye un servicio de telecomunicaciones, 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 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.

Hasta el momento en lo referente a temas relacionados con el desarrollo de Internet, la estrategia ha sido de mínima intervención, con el fin de que el mercado potencial crezca y no se ahogue en regulaciones y normas que puedan estancarlo, asegurando

a los prestadores del servicio que el marco jurídico que enfrentan no sufrirá grandes cambios en el mediano plazo.

Capítulo 3

Redes de computadores

We are caught in an inescapable network of mutuality, tied in a single garment of destiny. Whatever affects one directly, affects all indirectly.
Martin Luther King, Jr

Las redes de computadores, tienen su origen en la necesidad de compartir un recurso limitado, por lo general de un costo alto y que debe ser usado de manera intensa. En un principio, elementos tales como las impresoras, entraban perfectamente dentro de la definición anterior, por lo que su uso era intenso y compartido por todos los usuarios de una red.

Para que dos computadores o dispositivos puedan establecer una comunicación, se hace necesario un medio de transmisión de datos, un protocolo de comunicación y por supuesto las aplicaciones de software, entre otros elementos.

Dependiendo del tipo de medio (cable, espectro electromagnético, fibra óptica, etc.) cada computador necesita de una interfaz de comunicación, comúnmente denominada *"tarjeta de red"*, la cual transforma la información desde la aplicación[1], a un formato transmitible dentro de un medio, por ejemplo volts si se trata de un conductor de electricidad, como puede ser un cable coaxial.

[1]Ejemplos: navegadores, lectores de correos electrónicos, etc.

Hay que tener presente que todo medio de comunicación, independiente de su ancho de banda, siempre estará expuesto a saturación, es decir, que muchas aplicaciones traten de usarlo para la transmisión de datos, por lo que su uso eficiente es siempre una prioridad. Para lograr lo anterior, las redes de computadores han evolucionado en el tiempo, mejorando los medios, las interfaces de comunicación, los protocolos, equipos de interconexión etc. Lo anterior no significa que la conectividad esté siempre garantizada, sino todo lo contrario, sigue siendo prioridad el buen uso del recurso escaso, por lo que el diseño de una red es esencial para su correcto funcionamiento presente y futuro.

3.1. Topologías de redes

A lo largo de la historia de la computación, el desarrollo de dispositivos y redes cada vez más rápidas, obligó a la creación de distintas topologías de red para hacer frente al desafío de transmitir datos en grandes cantidades y en un tiempo muy corto.

Conforme se desarrollaban nuevos medios de transmisión de datos, con un ancho de banda cada vez más amplio, fue necesario perfeccionar la forma en que se interconectaban los computadores asociados a la red, es decir, crear nuevos métodos y dispositivos que permitieran hacer un uso más eficiente del ancho de banda. Nacieron, entonces, elementos como los hub, switch, router, etc. los cuales vinieron a estandarizar la forma en que se construyen las redes, permitiendo asegurar una conectividad continua y sin mayores sobresaltos.

3.1.1. Bus de datos

Tal vez la más antigua y simple de las topologías de redes, el bus de datos representa un concepto muy sencillo, un cable con varios computadores conectados a través de este. La Fig. 3.1.1 muestra dos buses de datos conectados a través de un repetidor.

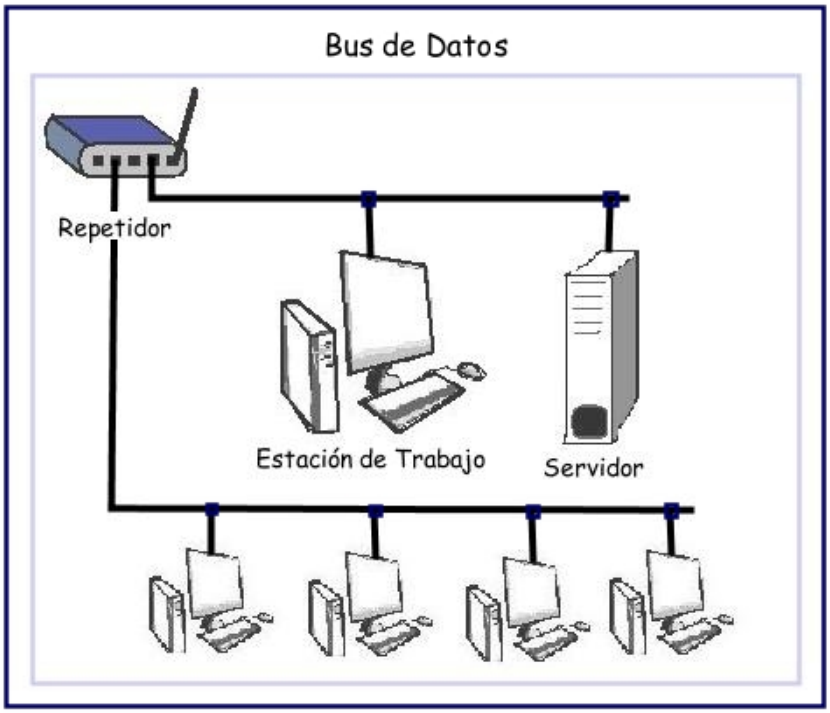

Figura 3.1: Bus de datos

La operación básica de la Fig. 3.1.1, es que cuando un computador decide realizar una transmisión, la envía directamente al bus, el cual por ser estar formado por un medio conductor de electricidad, excita inmediatamente a las interfaces de comunicación de los demás computadores de la red. Lo anterior genera un problema cuando dos computadores al mismo tiempo desean comunicarse, pues se produce una *"colisión"*, lo cual obliga a ambos computadores a esperar un tiempo aleatorio para transmitir nuevamente, con la consiguiente perdida de tiempo.

Esta topología fue muy utilizada al comienzo de la redes de computadores, pero fue discontinuada por problemas de ancho de banda efectivo para la transmisión, retardos por colisiones y poca flexibilidad en su estructura.

3.1.2. Bus estrellado

Como se dijo en la sección anterior, uno de los grandes problemas del bus de datos es su inflexibilidad estructural, la cual causa su çaída.ªnte situaciones simples como que alguien mueva el computador que está conectado al bus y se pierda la señal transmitida. Ante esta situación, se desarrollaron topologías con nuevos dispositivos para la interconexión de computadores, siendo una de estas el bus estrellado, como se muestra en la Fig. 3.2.

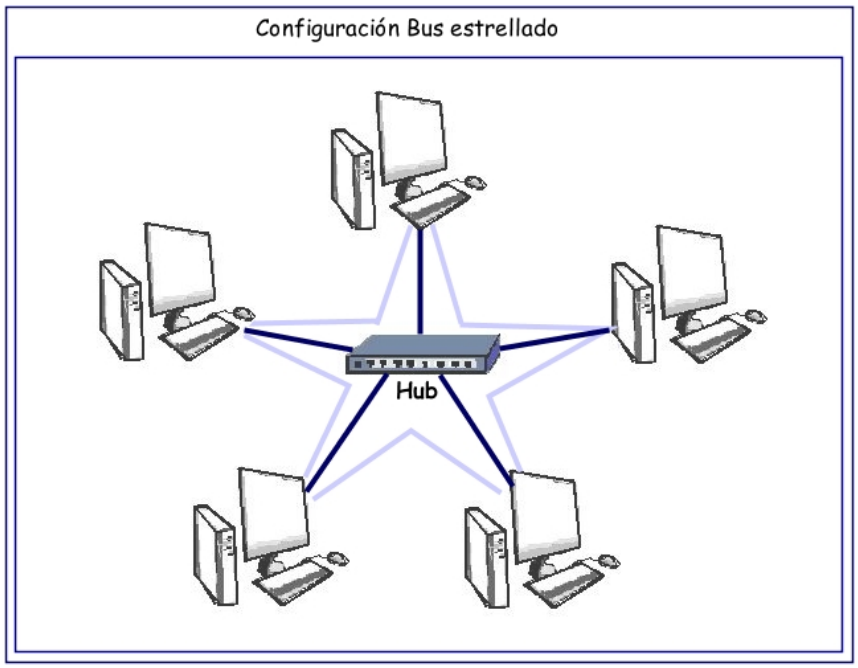

Figura 3.2: Bus estrellado

En la topología presentada en la Fig. 3.2, la conectividad se logra a través de un dispositivo concentrador de datos el cual se conecta con los computadores a través de un cable tipo par trenzado (similar al que se encuentra en las líneas telefónicas). El dispositivo provee de una cantidad limitada de conectores, los cuales permiten la anexión de un computador en forma simple y directa. La forma de transmisión de

los datos se mantiene, es decir, cuando un computador inicia la comunicación, esta se dirige al concentrador, el cual la reenvía a todos los computadores anexados a la red. Nuevamente, el problema de la colisión de datos se hace presente.

3.1.3. Token ring

La colisión de datos estudiada en las secciones precedentes, puede ser muy nefasta en redes de alta demanda. Como medida de solución, surgen las topologías de *"paso de testigo"* o *"token"*, las cuales consisten en que un computador puede transmitir, siempre y cuando tenga el token en su poder, tal como se muestra en la Fig. 3.3.

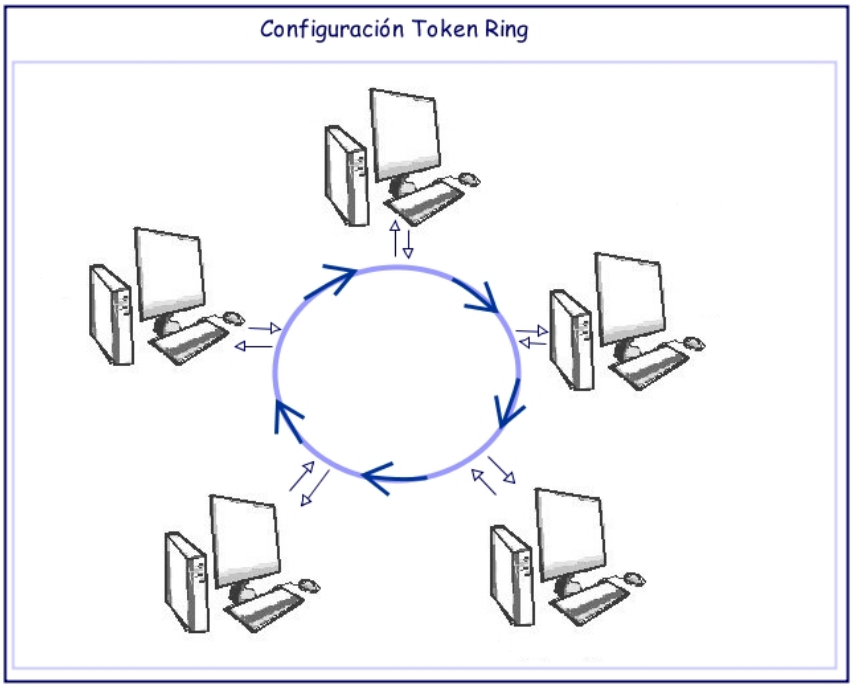

Figura 3.3: Token Ring

Este tipo de topología ha sido muy utilizado en redes de alta demanda, pero con pocos computadores conectados. Su principal problema es el retardo inherente de estar esperando a que el token llegue al computador que necesita transmitir.

3.1.4. Ethernet

Tal vez el estándar más exitoso de la historia en la creación de redes de área local. Ethernet fue propuesto en 1973 por el Dr. Robert M. Metcalfe en el PARC (Palo Alto Research Center) de la compañía Xerox, como una forma de lograr un medio de comunicación entre computadores que fuese mejor que el utilizado a través de vías telefónicas (en esa época muy lentas en transmisión de datos), pero no tan costoso como las redes de alta velocidad existentes.

La Fig. 3.4 muestra la configuración clásica de la red Ethernet, en la cual los computadores están conectados mediante cable coaxial o de par trenzado y compiten por acceso a la red utilizando un modelo denominado CSMA/CD (Carrier Sense Multiple Access with Collision Detection), el cual disminuye el problema de la colisión en la transmisión estudiado en las secciones precedentes. En su versión original, este tipo de redes fue pensada para la transmisión de información a 10Mb/s, pero hoy en día es común ver velocidades que superan los 100 Mb/s.

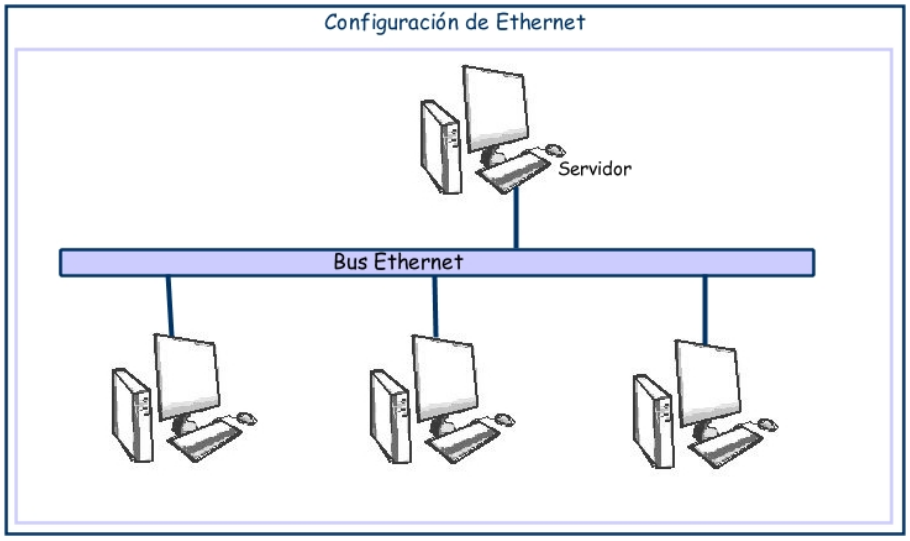

Figura 3.4: Configuración clásica de una red Ethernet

Tal fue el éxito de la red Ethernet, que en 1980 se adoptó como estándar, creándose la norma IEEE 802.3, que define el método de acceso CSMA/CD. Ya para 1982, Ethernet fue gradualmente adoptada por la mayoría de los organismos de estandarización (ISO, ANSI, IEEE, ECMA y NIST)

Un distintivo de las redes Ethernet, es su interfaz de comunicación. Cada tarjeta de red tiene asignado en su hardware un número de seis bytes que no se puede alterar y que es usado para su identificación en forma única, es decir, si un fabricante desea producir tarjetas ethernet, entonces debe solicitar a la IEEE la asignación de un número de 24 bits (3 bytes), con el cual se identifica a la institución que fabricó la tarjeta. Este número se conoce como OUI (Organizationally Unique Identifier) o código del vendedor. Luego cada fabricante agrega a su OUI otros 24 bits, con lo que se alcanzan los 6 bytes que identificarán a la tarjeta.

A estos 6 bytes, por lo general en formato hexadecimal, algo parecido a **00:1f:5b:d5:2e:b4**, se le conoce como la dirección MAC (Media Access Control) y es la que identifica en forma física a la tarjeta. Si el dispositivo conectado a la red Ethernet posee sólo una tarjeta, entonces la dirección MAC servirá también para su identificación en forma única.

La construcción de redes Ethernet es relativamente simple, salvo por el problema la poca flexibilidad que se logra debido a los cables. En efecto, para que un computador se conecte a este tipo de red, se hace necesario que exista un cable o punto de red para conectar el dispositivo.

Como solución a este problema y siempre pensando en caminar hacia las redes ubicuas, es decir, conectividad no importando donde se encuentre el dispositivo, se desarrollaron las redes inalámbricas, donde la tecnología WiFi[2] ha sido hasta el momento la más difundida y aceptada.

Wifi utiliza ondas de radio para la transmisión de datos, siendo su cobertura

[2]Wireless Fidelity o Fidelidad Inalámbrica

cercana a una decena de metros. A partir de esta frontera física, se hacen necesario repetidores u otros equipos de similar tecnología para seguir transmitiendo/recibiendo las señales.

Las redes Wi-Fi poseen varias ventajas frente a las que utilizan cables, por ejemplo la posibilidad de portar equipos sin necesidad de un cable o punto de red, economizan infraestructura de cableado y su mantención, existe una fuerte estandarización lo que permite que un equipo pueda operar casi en cualquier parte del mundo, aunque aun persisten problemas con las normas de seguridad en la transmisión de datos que hace que algunos equipos no puedan conectarse a una de estas redes.

La principal desventaja de estas redes son el ancho de banda efectivo que pueden entregar a sus usuarios (menos de 50Mbps) y los problemas de seguridad, que las hace relativamente fáciles de vulnerar por aplicaciones de detección y análisis de paquetes, las cuales permiten encontrar, por ejemplo, las claves de acceso a la red.

3.1.5. Equipos de interconexion

Como una forma de mejorar continuamente las comunicaciones y teniendo presente la inminente interconexión entre redes con otras redes, se hizo necesario crear una serie de dispositivos que permitieran la comunicación entre dispositivos de forma eficiente y eficaz. La Fig. 3.5 muestra los principales equipos de interconexión y creación de redes.

Cada equipo fue creado para atender a un problema en específico. La Fig. 3.6 presenta al primero de ellos: el HUB, el cual en esencia es un repetidor simple, es decir, todo dato que recibe desde un dispositivo, es retransmitido a todos los demás dispositivos presentes en la red.

La cantidad de computadores que pueden usar el HUB, queda inmediatamente limitado por el número de cables que acepta. Sin embargo, es posible conectar HUBs en cascada para aumentar el número de computadores de la red, pero en desmedro del

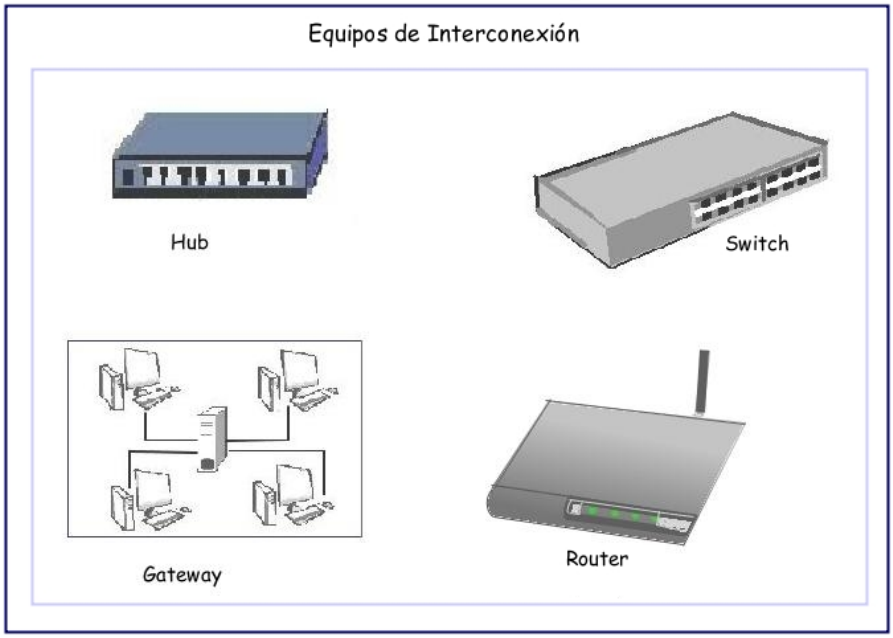

Figura 3.5: Equipos de interconexión usados en la creación de redes

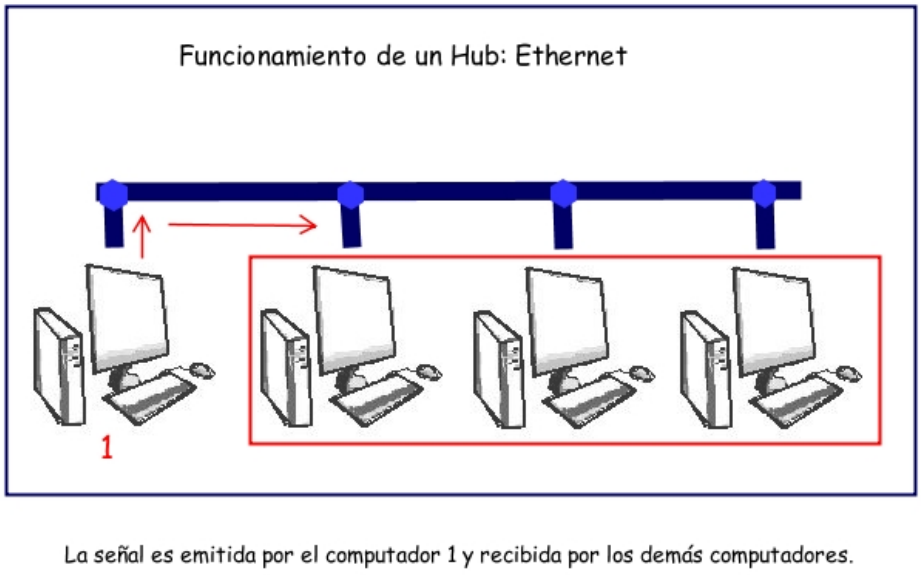

Figura 3.6: Funcionamiento del Hub

ancho de banda efectivo, por cualquier dato que se envíe, se retransmite a los demás computadores.

Evidentemente, el envío de datos que propone el HUB es altamente ineficiente, por lo que el siguiente paso en el desarrollo de los equipos de interconexión fue pensar en cómo enviar los datos sólo al dispositivo de destino al que van dirigidos. La Fig. 3.7 muestra el funcionamiento del SWITCH, dispositivo que posee una aplicación de software que le permite despachar un dato al remitente en específico.

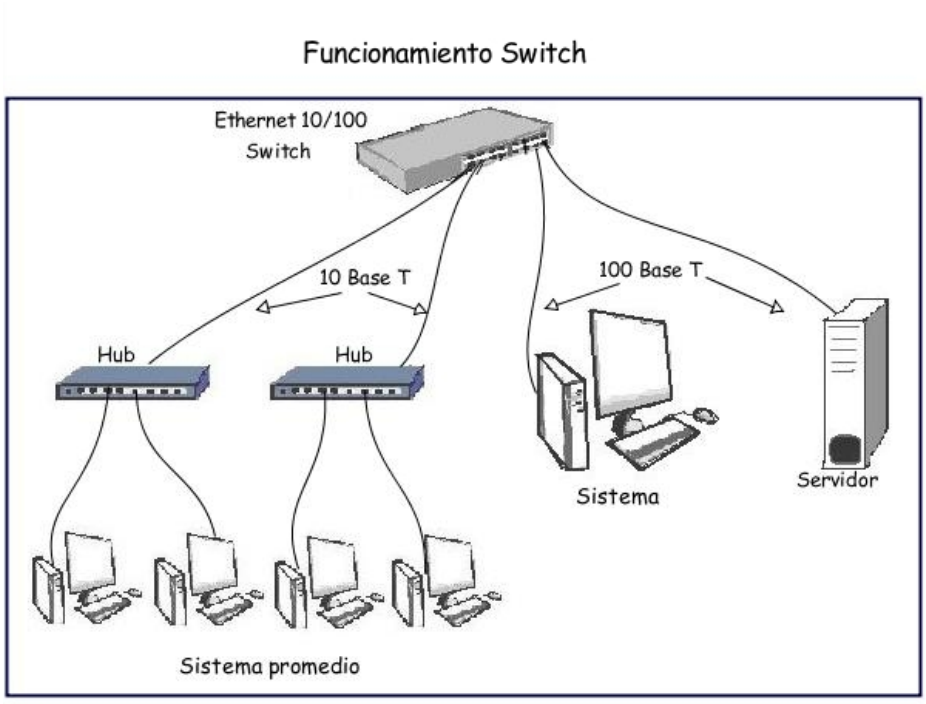

Figura 3.7: Funcionamiento del Switch

Con la incorporación de los SWITCH se abre una nueva ventana de posibilidades para la creación de redes. Como se puede apreciar en la Fig. 3.7, no sólo se pueden conectar computadores a la red, sino que HUBs y sus dispositivos asociados.

Con los equipos anteriores es posible crear redes LAN muy eficaces y eficientes, posibilitando el siguiente paso: la conexión de una red con otra red. En la Fig. 3.8

se muestra la operación del ROUTER, equipo cuya principal característica es la de contar con una aplicación de software que puede *"rutear"* o encaminar los datos que van hacia otras redes de protocolos similares.

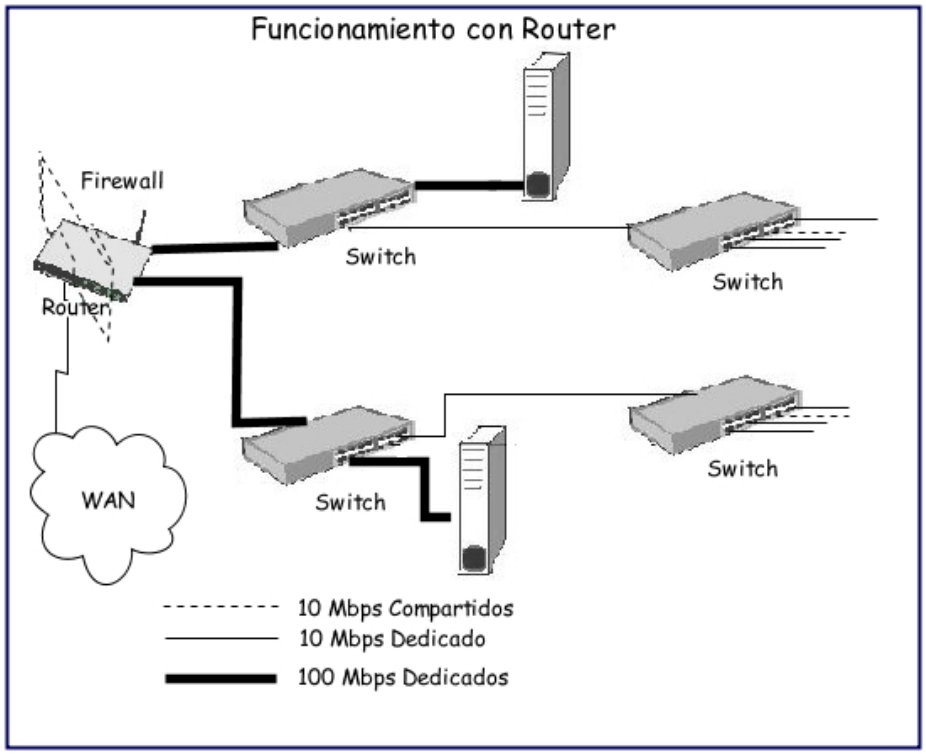

Figura 3.8: Funcionamiento del router

El encaminamiento siempre ha sido un proceso no trivial, sobre todo cuando el destino está geográficamente alejado y se requiere del paso por varios ROUTERs antes de llegar a destino. Este es el caso de la red Internet, la cual puede ser vista como un enjambre de muchos ROUTERs que permiten la comunicación entre computadores a escala global.

Un último equipo, que conceptualmente opera igual que un ROUTER, es el GATEWAY, el cual permite la conectividad entre dos redes que utilizan protocolos distintos en su comunicación interna. En la actualidad, debido al avance de protocolos

como el TCP/IP, prácticamente todas las redes poseen el mismo lenguaje, por lo que las palabras ROUTER y GATEWAY tienden a confundirse y a usarse casi como un sinónimo.

3.2. Modelo de Interconexión de Computadores ISO/OSI

En TICs, hay un principio que se remonta a la época romana *"divide ut regnes"* o *"dividir para reinar"*. En el contexto tecnológico, esta frase quiere decir que un problema muy complejo, se puede dividir en unidades mucho más pequeñas y solucionarlas por separado. Luego la solución para el problema complejo será la unión de las soluciones parciales a los problemas más pequeños.

Con esta idea en mente, el Dr. Andrew S. Tanenbaum en su obra *"Computer Networks"* [47] visualizó que la problemática detrás de comunicar dos computadores a través de sus aplicaciones, se podría abordar a partir de solucionar problemas más pequeños, con lo cual anunció su ya celebre modelo de interconexión de estándares abiertos para computadores, el se compone por siete capas, tal como se muestra en la Fig. 3.9.

Aplicación. En este nivel, se establece la mejor forma de que el usuario tome contacto con las aplicaciones, por ejemplo un lector de e-mails tal como Thunderbird o Eudora. Luego lo que el usuario escribe debe ser codificado antes de pasar al siguiente nivel.

Presentación. Se crean las estructuras o paquetes de datos, con el formato y sintaxis[3] adecuada para ser transmitidos por la red.

Sesión. Establece el mecanismo de inicio y fin de la sesión. Aquí se resuelven problemas tales cómo que sucede cuando se corta la comunicación en forma abrupta,

[3]Los paquetes de datos tienen una estructura sintáctica semiflexible, donde se especifican encabezados, direcciones de origen/destino, los datos mismos, etc.

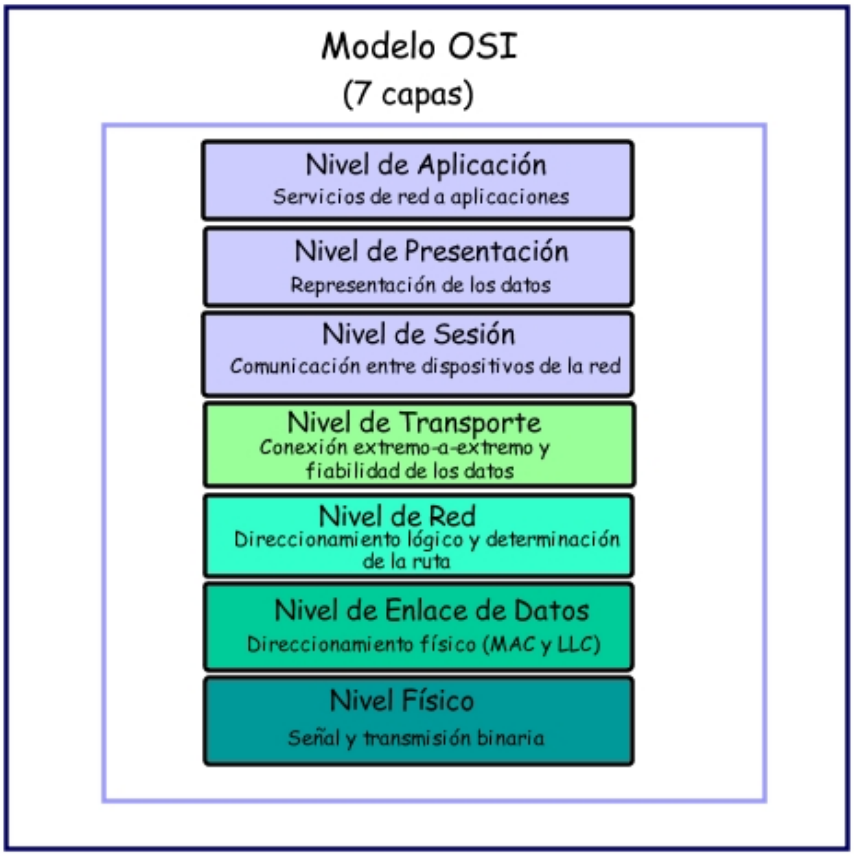

Figura 3.9: Modelo de referencia OSI para interconexión de computadores.

cómo recuperarla, qué hacer con los datos que no llegaron a destino, etrc.

Transporte. Asegura que la llegada de datos del nivel de red encuentra las características de transmisión y calidad de servicio requerido por el nivel 5 (Sesión).

Red. Proporciona el enrutamiento de mensajes, determinando si se les debe enviar al nivel 4 (Transporte) o bien al nivel 2 (Enlace). Es responsabilidad de este nivel establecer, mantener y terminar las conexiones.

Enlace. Recibe los mensajes provenientes de la capa de red, y les asigna las direcciones de origen y destino del nivel de enlace, gestiona la detección y corrección de errores y toma decisiones de reenvío o pausa cuando la red está muy con-

gestionada. Por último, el mensaje es transformado al formato que se requiere en el nivel físico, por ejemplo, si se trata de una red de fibra óptica, entonces el mensaje se transforma en un has de luz.

Físico. Define el medio de comunicación utilizado para la transferencia de datos, por ejemplo fibra óptica, cable, etc.

3.3. Protocolos de comunicación

En el fenómeno de la comunicación, siempre están presentes dos entidades: transmisor y receptor. El primero es quién inicia la comunicación y el segundo quien la recibe. Lo que se transfiere entre ambas partes se denomina **mensaje**.

En la Fig. 3.10 se presenta el proceso mediante el cual se transmite el mensaje. Primero el receptor codifica lo que va a transmitir a través de un medio. El receptor, luego de recibir el mensaje, tiene que descodificarlo para poder entenderlo. En el caso de que el receptor decida responder el mensaje (feedback), lo hace a través de una respuesta que también debe ser codificada y transmitida por el medio, generando un nuevo mensaje, invirtiendo los roles de las entidades participantes. Cuando la entidad, ahora receptora, recibe el mensaje, tiene que descodificarlo para poder entenderlo. De esta forma, se ha completado un ciclo de comunicación. Es importante señalar que no siempre existe feeback en la comunicación (comunicación unidireccional) y que puede suceder que las entidades sean sustituidas durante el proceso de transmisión del mensaje.

Para que exista la comunicación descrita en la Fig. 3.10, es necesario que se establezcan algunas normas que preserven la secuencia de eventos, de lo contrario no habría entendimiento. Se necesita entonces de un *"protocolo que norme la comunicación"*.

En su definición más simple, un protocolo de comunicaciones se define como una

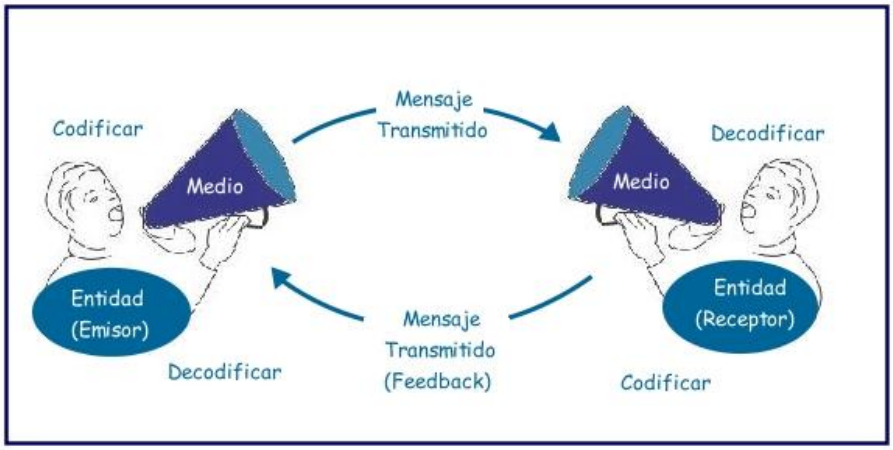

Figura 3.10: Ciclo básico de la comunicación entre dos entidades

convención o estándar que controla o permite la conexión, comunicación, y transferencia de datos entre dos puntos finales. En el caso de los computadores, se trata de puntos que están geográficamente distantes, los cuales se comunican a través de aplicaciones que son ejecutadas en los respectivos sistemas operativos residentes.

Dado que estas aplicaciones están en dispositivos separados, es decir, el funcionamiento de una no condiciona el de la otra, se hace necesario un mecanismo de interacción y coordinación para la sincronización de las tareas que se ejecuten en forma colaborativa. Entonces, el protocolo a utilizar tendrá como objetivo principal el establecer un ambiente de trabajo donde los períodos y secuencia de secuencia de eventos es desconocida, con una alta probabilidad de errores en la transmisión de los datos.

De un punto de vista informático, el término protocolo se usa para describir el intercambio de datos entre aplicaciones, es decir, programas que se ejecutan en un determinado dispositivo de la red. Formalmente, el protocolo especifica la sintaxis de la comunicación entre las aplicaciones, siendo sus funciones más importantes:

- Control de errores. Detectar los errores de transmisión de datos e iniciar acciones

para enmendarlos y/o solicitar retransmisión de estos.

- Control de Flujo y Congestión. La idea central es que la red pueda compartir sus recursos entre un gran número de usuarios, entregando a cada uno de ellos un servicio satisfactorio y libre de errores.

- Estrategias de encaminamiento. Destinadas al uso eficaz y eficiente de los recursos de la red, aumentando la disponibilidad de los servicios asociados al proveer caminos alternativos entre los dispositivos.

En TICs, el objetivo de los protocolos es permitir que se produzca la comunicación entre las aplicaciones que se están ejecutando en dispositivos de hardware, geográficamente distantes. Entonces, la aplicación de origen debe conocer el mecanismo, comúnmente una dirección, a partir de la cual podrá encapsular los datos a transmitir en un mensaje y enviarlo. Cuando esto ocurre, la aplicación de origen entra en un estado de espera de la respuesta que tendrá que darle el receptor del mensaje.

La aplicación receptora, por su parte, si acepta el mensaje, responde a su emisor y por lo general describe detalles de cómo recibió el mensaje, para hacer chequeos y tomar decisiones de reenvío en el caso que sea necesario.

Debido a que no es 100 % seguro de que un mensaje llegue a destino, producto de que las redes pueden tener problemas de las más variadas índoles, lo estándar es que ambas aplicaciones que intervienen en la comunicación posean configurados un tiempo de espera, a partir del cual, si no han recibido noticias, se genera un nuevo mensaje para restablecer la comunicación. Este proceso, comúnmente se realiza unas tres veces, antes de que la aplicación le informe a su usuario de que se perdió la comunicación.

3.4. Medios de transmisión de datos

Desde un comienzo, el desarrollo de redes de computadores ha estado marcado por la necesidad de más ancho de banda, a un menor costo y de alta disponibilidad. Es así que el avance científico y tecnológico ha permitido pasar de tasas de Kbytes a TBytes de transmisión de datos. La tabla 3.1, describe alguno de estos medios de transmisión con sus principales ventajas y desventajas a la hora de ser utilizados en la creación de una red.

En la Fig. 3.11 se pueden observar los distintos medios de transmisión de datos, que se describieron en la tabla 3.1.

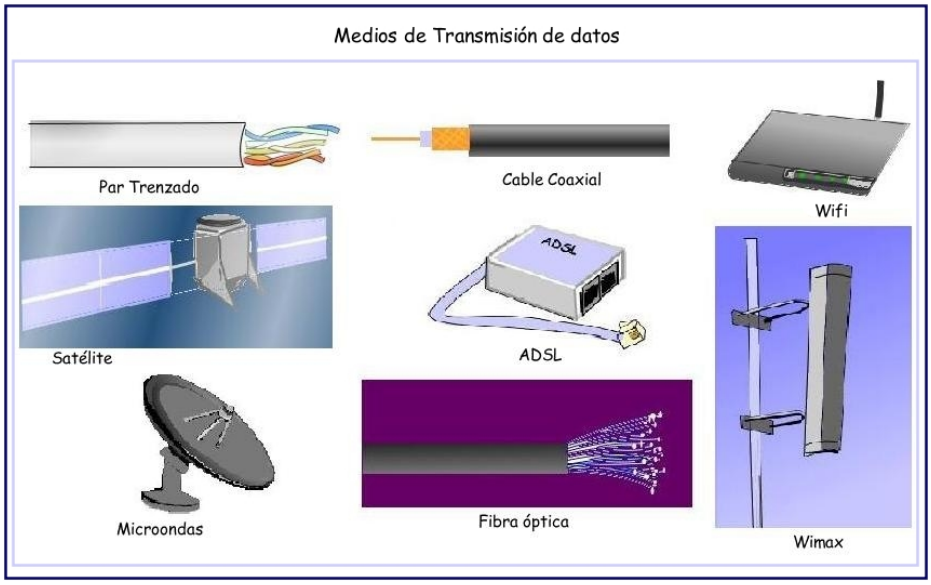

Figura 3.11: Medios de Transmisión de datos.

El uso de los medios de transmisión está fuertemente ligado a las limitaciones topológicas y geográficas del lugar por donde se desea transmitir las señales, consideraciones de seguridad y por supuesto económicas. Por ejemplo, sería ideal si por sobre los campos de hielos patagónicos se pudiera utilizar fibra óptica, por su enorme ancho de banda. Sin embargo, lo accidentado e inaccesible del terreno hacen de esa empresa

Medio	Descripción	Ventajas o Desventajas comparativas
Cable Coaxial	El cable coaxial se compone de una cubierta externa de plástico, una capa de cobre (tejida con finas fibras del material), una capa de aislante y un centro de metal, que por lo general es cobre.	El cable coaxial ha sido reemplazado por la fibra óptica, ya que no es capaz de igualar el ancho de banda que presenta la fibra óptica, además que el cable es necesario aislarlo de esfuerzos de tracción y de presión, por lo que hay que protegerlo con estructuras de acero, mientras que la fibra al ser casi completamente aislada es resistente a las condiciones externas.
Par Trenzado	El par trenzado se compone de un cable que recubre a nivel externo y cables finos aislados, entrelazados a nivel interno.	Una ventaja es la disminución de interferencia debido a su estructura entrelazada, la cual permite una separación entre los cables que lo conforman, formando cada uno una una especie de espiral que impide el acoplamiento entre las señales que se transmiten. Sin embargo, el cable coaxial disminuye más aun la interferencia.
Fibra Óptica	La fibra óptica es un cable delgado y liviano, que se compone de tres capas. La mas externa es un material que protege el interior. La siguiente capa, cuyo material es el vidrio, limita la transmisión de los pulsos de luz que viajan por el centro.	Evita las interferencias electromagnéticas, permitiendo la transmisión de datos entre largas distancias a un precio bajo. Como no emite energía al exterior, se vuelve una forma de transmisión segura (no se puede detectar la transmisión, entonces es imposible interferir el sistema, sin tener . que romper la fibra, lo cual sería detectado)
Satélite	El satélite transmite las señales al funcionar como repetidor.	Evita el alto costo de ubicar repetidores cada cierta cantidad de kilómetros, porque actúa como repetidor de alta capacidad. Sin embargo, su costo es altísimo.
Línea Telefónica	La línea de teléfonos convencional fue utilizada casi desde el comienzo de las redes como	En un principio, el ancho de banda estaba restringido a la transmisión de voz, pero luego con la digitalización

Medio	Descripción	Ventajas o Desventajas comparativas
	un medio barato y eficiente de interconexión, para lo cual se desarrollo un dispositivo que transformaba los datos en señales de audio (Modem)	de las redes telefónicas, se pudo usar casi toda la capacidad de transporte de señales que posee el cable de cobre (uso de tecnología XDSL ver sección 4.6).
Microondas	Las microondas son todas aquellas bandas de frecuencia en el rango de 1 GHz en adelante, el término microondas viene porque la longitud de onda de esta banda es muy pequeña (milimétricas o micrométricas), resultado de dividir la velocidad de la luz (3x108 m/s) entre la frecuencia en Hertz.	Las microondas también fueron desplazadas por la Fibra óptica, pero aun son utilizadas como medio de respaldo.
Wifi	Wireless Fidelity utiliza las ondas de radios en vez de cables. Por lo que, si se busca describir Wifi se podría mencionar los elementos que se utilizan para su funcionamiento como el router y puntos de acceso en la recepción de la señal y tarjetas receptoras (Ej.: el puerto USB)	Al utilizar las ondas de radio para la transmisión de las señales, el exceso de usuarios de Wifi están colapsando el espectro y reduciendo la velocidad de transmisión. Esta desventaja, le está restando terreno a este medio y favoreciendo la masificación otros medios.
Wimax	Worldwide Interoperability for Microwave Access también utiliza las microondas para la recepción de las señales y ondas de radio para la retransmisión de la señal	Al no tener que ubicar instalaciones como las que se necesitan en una red con cables, es posible satisfacer la demanda de lugares alejados y con baja densidad poblacional.

Tabla 3.1: Medios de Transmisión. Elaboración basada en [24]

un sueño casi imposible, por lo que se recurrre a otros medios, como los satélites y al uso de las redes que están por el lado argentino de la Patagónia.

3.5. Las redes de computadores en la legislación

En la sección 2.7 se establecieron las razones para considerar que el acceso a Internet constituye un servicio de telecomunicaciones, por lo cual le corresponde a la Subsecretaría de Telecomunicaciones velar por su correcto funcionamiento. Dada la intima relación que guarda la gran red con las actuales redes de computadores locales, cualquier normativa que regule su acceso afecta directamente a las redes corporativas, por lo que es necesario hacer un análisis respecto de la normativa vigente y futura respecto del desarrollo de Internet.

3.5.1. Normas y reglamentos para el correcto funcionamiento de las redes de computadores

Desde el comienzo de Internet, las direcciones IP públicas han sido entregadas de acuerdo a una distribución geográfica que facilite el envío de los paquetes de datos. Al principio la obtención de direcciones IP públicas para una institución no era un tramite muy engorroso, de hecho era posible obtener un rango de IPs sin que esto fuera un problema mayor. Si embargo, con el crecimiento explosivo de Internet, las direcciones IP se hicieron cada vez más cotizadas y escasas, por lo que los organismos gestionadores de la gran red tuvieron que establecer una normativa para la entrega de números, los que cada vez son escasos en el actual protocolo de comunicaciones IPV4.

Para los proveedores del servicio de acceso a Internet, en la actualidad existen ciertos rangos muy limitados de direcciones IPs públicas que pueden asignarle a sus contratantes. De hecho la práctica más habitual ha sido el uso de IPs privadas, es decir, que sólo tienen sentido dentro de una red LAN y que luego cambian a una IP pública cuando se accede a algún recurso externo.

Las IPs públicas permiten la identificación de un dispositivo en Internet, por lo que pueden ser usadas para probar y luego demostrar que un determinado usuario realizó un acto indebido valiéndose de la conectividad que ofrece Internet. Al respecto, la Subtel en el Decreto 142 de 2005 sobre Interceptación y Grabaciones Telefónicas y de Otras formas de Telecomunicación, señala que *"los proveedores de acceso a Internet deberán mantener, en carácter reservado, a disposición del Ministerio Público y de toda otra institución que se encuentre facultada por ley para requerirlo, un listado actualizado de sus rangos autorizados de direcciones IP y un registro, no inferior a seis meses, de los números IP de las conexiones que realicen sus abonados. Asimismo, deberán otorgar las facilidades necesarias para llevar a cabo las intervenciones que fueren ordenadas, debiendo sujetarse al respecto a lo prescrito en el artículo 2.º del presente reglamento."*

Es interesante notar que esta norma, en el caso de redes LAN que acceden a Internet a través de una única IP pública, tendrían que mantener un registro de todas las conexiones internas, por cuanto todo dispositivo con IP privada dentro de la red que desee acceder a un recurso en Internet, sufrirá una traducción de IP en el router, es decir, se cambiará la IP privada por una pública. Lo anterior implica que desde Internet sólo se ve la IP pública asignada al router, pero nada se puede ver de las IPs privadas asignadas a los dispositivos internos de la red LAN.

Otro elemento importante que ha cubierto, en parte, la resolución exenta 698 de 2000 emitida por Subtel, ha sido la fijación de indicadores calidad de los enlaces de conexión para cursar el tráfico nacional de Internet, con el correspondiente sistema para que dichos indicadores puedan ser conocidos por todos los usuarios. Esta norma establece valores mínimos para la Tasa de pérdida de paquetes, Latencia y Tasa de ocupación de un enlace. Lo anterior repercute directamente en la calidad de la conectividad que se puede garantizar en una red LAN, ya sea institucional o de un usuario doméstico.

Finalmente, la resolución exenta 1.483 de 1999 emitida por Subtel fija el pro-

cedimiento para establecer y aceptar conexiones entre los proveedores del servicio Internet. En su texto señala que *"con el objeto de garantizar el buen funcionamiento y la no discriminación en la calidad del servicio de acceso a Internet prestado a los usuarios, los ISP deberán, previo al inicio de servicio, establecer y aceptar conexiones entre sí para cursar el tráfico nacional de Internet"*. Esta resolución garantiza la conectividad a nivel nacional. Antes de su promulgación, era frecuente que un usuario abonado a un proveedor A, en su navegación hacia una página accesible a través de un proveedor B, tuviese que dar la vuelta al mundo para poder enviar sus datos, pues no había conectividad entre proveedores.

3.5.2. Neutralidad en la Red

Este concepto fue acuñado por el Prof. Tim Wu[4] de la Columbia Law School para definir como red neutral a aquella que *"permite comunicación de punto a punto sin alterar su contenido"*. Wu se baso en la forma de operación de las redes telegráficas del siglo XIX, en las cuales no se discriminaba entre los servicios que utilizaban dichas redes.

En Agosto de 2005, la Federal Communications Commission (FCC) en EEUU suprimió la ley de *"common carrier"* que impedía a las compañías telefónicas norteamericanas controlar el contenido de lo que transitaba por sus redes. El efecto inmediato fue pensar que iba a suceder con las redes de computadores, por lo que el concepto de Neutralidad en la Red comenzó a estar presente en las discusiones referente al futuro de Internet.

De un punto de vista técnico, neutralidad en la red es una forma de diseñar redes, cuyo objetivo es maximizar la transmisión de datos a todos los usuarios, para lo cual la red debe tratar por igual todos los contenidos que por ella circulen. Más aún, no hará excepción entre las plataformas que se interconecten, sólo así la red podrá servir para comunicar todo tipo de dato, independiente del servicio.

[4]http://www.timwu.org/network_neutrality.html

Mucho se ha discutido respecto de lo que en realidad implica el concepto de neutralidad en la red y aunque aun no está del todo claro, al menos hay tres corrientes que intentan definir de qué se trata al menos la idea [61]:

- Absoluta no discriminación.
- Discriminación limitada, sin grados de servicio.
- Discriminación limitada, con diferentes grados de servicio.

Para analizar cada una de estas posturas, conviene recordar las palabras del Dr. Vint Cerf, co-inventor del protocolo IP, quien sostiene que *"Internet fue diseñada con ningún guardián sobre nuevos contenidos o servicios. Una suave pero aplicable regla de neutralidad de red es necesitada para que continúe creciendo".*

Aquellos que están de acuerdo con la primera postura, sostienen que la red funcionará y evolucionará mejor si el acceso y el tráfico están totalmente separados del contenido. De esta forma, las empresas proveedoras del servicio Internet sólo se dedicarán a proporcionar acceso y tráfico, sin que exista discriminación alguna del contenido o procedencia de los datos transmitidos.

Las otras dos posturas ya requieren de algún tipo de regulación que salvaguarde los intereses de los usuarios finales. El problema es que controlar que se cumpla una regulación al respecto, demandará un esfuerzo no menor, de difícil cuantificación a priori y que siempre se verá envuelto en supuestos de conectividad y de calidad del servicio que pueden dar la razón a todas las partes involucradas en un eventual conflicto por la transmisión de datos en las redes de computadores.

Mirando hacia futuro y centrándose en el mayor y más destacado de los servicios sobre las actuales redes de computadores, es decir, la Web, y en palabras de su creador Sir Tim Berners-Lee, la neutralidad es necesaria para la existencia de una sola red, por cuanto *"quienquiera intente cortarla en dos se dará cuenta que el pedazo que le corresponda no es muy atractivo. Es mejor para todos y más eficaz si hay un mercado*

donde obtenemos nuestra conectividad y otro mercado donde conseguimos el contenido. Para tomar mis decisiones utilizo información. No sólo para decidir lo que compro, sino también para decidir a quien voto- Algunas compañías de los EE.UU. quieren cambiar esto. Quieren llegar a la situación en que si yo deseo ver una cadena de TV por internet, esa cadena haya tenido que pagarles para que la señal me llegue bien". [5]

En el contexto nacional, Chile ya ha dado el primer paso para la Ley Internet y Neutralidad de Red para defender derechos de los usuarios. Esta iniciativa ya había sido aprobada por unanimidad en la Cámara de Diputados y en general por el Senado. Ahora se busca acelerar su aprobación final, dándole urgencia al tramite legislativo.

El proyecto de ley [6], mediante la regulación de la prestación del servicio de internet a través de una modificación de la ley de derecho al consumidor, pretende consagrar el principio de la neutralidad en la red, asegurando "*a todos los usuarios el acceso libre de contenidos o ejecutar aplicaciones o utilizar los dispositivos de su elección sin condicionamientos de ningún tipo*"[7].

[5]Texto traducido de la charla que dió Sir Tim Bernes-Lee en la duodécima edición de la Campus Party, que tuvo lugar en Valencia en Julio de 2008

[6]Boletín 4915-19 Consagra el principio de neutralidad en la red para los consumidores y usuarios de Internet. Disponible en: http://sil.senado.cl/pags/index.html

[7]Mensaje Boletín 4915-19

Capítulo 4

Redes TCP/IP

Las lenguas actuales poseen una gran cantidad de palabras que ya son internacionales que son conocidas por todos los pueblos y que son un tesoro para la futura lengua internacional.
Ludovico Zamenhof (Creador del Esperanto)

El paso natural luego de que se estabilizó la tecnología de creación de redes LAN, fue pensar en la posibilidad de interconectarlas para así expandir el horizonte de posibilidades que podría tener un usuario. Sin embargo, lo que pareció algo simple de implementar, tuvo muchos problemas a su comienzo. La falta de un único lenguaje o protocolo que permitiera la interconexión, la alta demanda en transmisión de datos y lo lento de los equipos de la época, hacían de esta empresa un sueño a ratos imposible.

Se necesitó, entonces, de la creación de un protocolo de comunicaciones que pudiera hacer frente a los problemas estructurares de las redes (baja conectividad, perdida de mensajes, poca seguridad en la entrega) y considerada que en cada una de ellas era posible que las comunicaciones tuviesen un lenguaje propio. En ese sentido, el TCP/IP surge como la solución para la conectividad en la red LAN, asegurando la entrega de los mensajes entre sus dispositivos, y más aun, permitiendo la conectividad con otras redes de protocolos distintos, tratando siempre de hacer el mejor esfuerzo para que el mensaje llegue a su destino.

Tal vez por encontrarse en un ambiente académico, las primeras versiones del TCP/IP no consideraron mecanismos de seguridad, tales como la criptografía, firma digital, etc. No fue hasta la versión 6, que está gradualmente adoptada en el mundo, que en forma nativa, se consideró una capa de seguridad en la transmisión de datos. Por lo pronto, la actual versión 4 del TCP/IP necesita de aplicaciones y protocolos adicionales para que la transmisión se realice en forma segura.

4.1. Propósito del protocolo

El primer problema a enfrentar es cómo comunicar las aplicaciones que existen entre dos computadores conectados a la misma red LAN, asegurando que el ciclo de comunicación se realice totalmente (ver sección 3.3). Con esa premisa, el primer protocolo que se desarrolló fue el Network Control Protocol o NCP, el cual se hizo cargo de asegurar una conexión full-duplex, es decir, todo mensaje enviado, recibe de vuelta una comunicación que dice el estado de su recepción.

NCP funcionaba bien como de comunicación en redes LAN, pero presentaba deficiencias de seguridad, falta de control de errores, de capacidad para direccionar los mensajes a otras redes, por lo que pronto fue necesaria una mejora drástica, la cual llegaría en 1974, cuando los Drs. Vint Cerf y Robert Kahn presentan a la comunidad científica, el protocolo TCP/IP [1], el cual subsanó gran parte de los problemas presentados en NCP, para lo cual establecieron una serie de requerimientos a satisfacer [28]:

1. Algoritmos para evitar la pérdida de paquetes en base a la invalidación de las comunicaciones y la reiniciación de las mismas para la retransmisión exitosa desde el emisor.

2. Provisión de pipelining (*"tuberías"*) host a host de tal forma que se pudieran enrutar múltiples paquetes desde el origen al destino a discreción de los hosts

[1] Transmission Control Protocol/Internet Protocol

participantes, siempre que las redes intermedias lo permitieran.

3. Funciones de pasarela para permitir redirigir los paquetes adecuadamente. Esto incluía la interpretación de las cabeceras IP para enrutado, manejo de interfaces y división de paquetes en trozos más pequeños, si fuera necesario.

4. La necesidad de controles (checksums) extremo a extremo, reensamblaje de paquetes a partir de fragmentos, y detección de duplicados, si los hubiere.

5. Necesidad de direccionamiento global.

6. Técnicas para el control del flujo host a host.

7. Interacción con varios sistemas operativos.

8. Implementación eficiente y rendimiento de la red, aunque en principio éstas eran consideraciones secundarias.

En TCP/IP, siempre se trata de hacer un uso eficiente del ancho de banda efectivo del enlace que posee el dispositivo transmisor, para lo cual todo archivo que se desee enviar, primero es transformado en una secuencia de paquetes o datagramas, que luego serán reagrupados por la aplicación receptora. Cada paquete posee una estructura rígida, que comienza con las direcciones IP de origen/destino, un campo de control y detección de errores, con lo que se resuelven los puntos 1 y 4.

También por operación del protocolo IP, los puntos 2, 3 y 5 encuentran una solución adecuada. Este protocolo es el encargado de asignar las direcciones de origen/destino que permite el enrutamiento de los datagramas a escala mundial, imponiendo condiciones de uso eficiente del ancho de banda, lo cual se traduce en la división del mensaje a transmitir, en datagramas pequeños, los cuales incluso se pueden dividir nuevamente, dependiendo de la condición del enlace que enfrente un envío. Cada datagrama posee una marca que le indica a la aplicación receptora cuál es la secuencia con que debe reagrupar todos los paquetes recibidos.

En este esquema de trabajo, son los equipos de interconexión, tales como switch, routers y gateway, los que determinan el encaminamiento de cada uno de los datagramas enviados/recibidos.

El protocolo TCP/IP, sigue los postulados en el modelo ISO/OSI de interconexión, pero fusiona alguna de sus capas. La Fig. 4.1 muestra la operación del TCP/IP en el modelo de referencia mencionado.

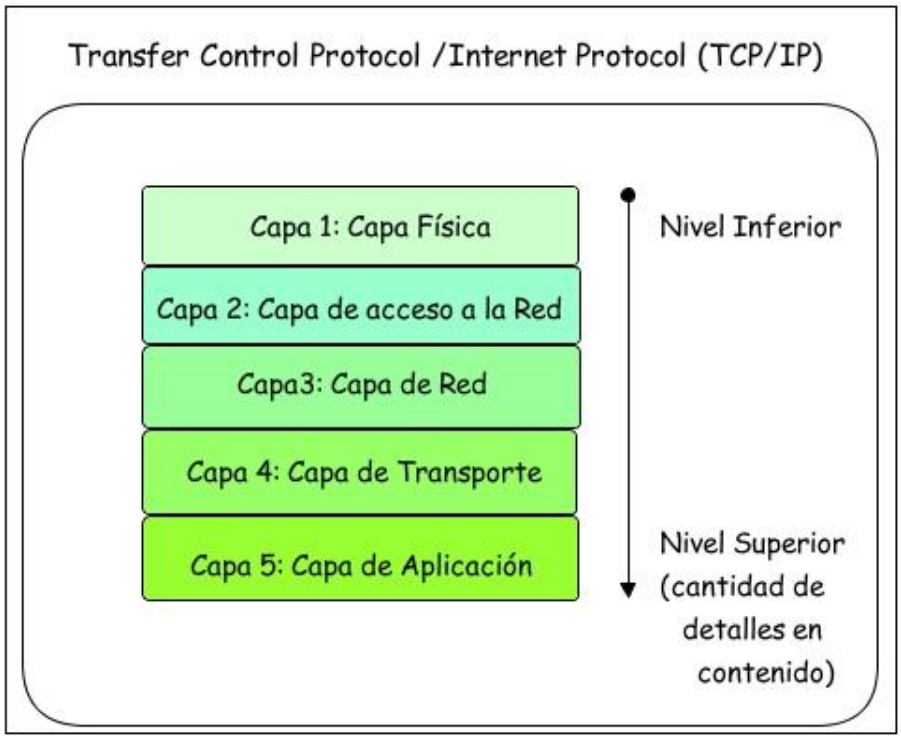

Figura 4.1: Protocolo TCP/IP. Elaboración propia.

A nivel de la **capa física** se encuentran medios de transmisión, tales como líneas de telefonía, Satélite, Fibra óptica, etc. (para un mayor detalle, ver sección 3.11)

En la **capa de acceso a la red**, se establece la comunicación entre los dispositivos involucrados en una comunicación, a través de una tarjeta de red. La más frecuentemente usada es la ethernet, estudiada en la sección 3.4.

La siguiente es la **capa de red o capa de Internet**, la cual proporciona al datagrama las direcciones IP de origen/destino para que pueda ser ruteado desde la red local hacia una red distante y finalmente al dispositivo de destino. Por operación del TCP/IP, cada dispositivo asociado a la red posee una IP que le es propia y que está directamente relacionada con la tarjeta de red, es decir, si un dispositivo posee varias de estas unidades, es muy posible que tenga la misma cantidad de IPs asignadas.

Aquí surge una pregunta que es obvia ¿Para qué se necesita de una dirección IP si con la dirección MAC de la tarjeta se identifica a un dispositivo en la red?. Lo anterior es verdadero, salvo por el hecho de que no todas las redes locales son Ethernet y es posible que tengan protocolos distintos. Entonces, el TCP/IP debe ser capaz de enfrentar estos detalles y establecer la comunicación, con lo que la consigna es **no importa que protocolo se use en la red local, si desea interconexión con otra red, entonces use TCP/IP**.

La **capa de transporte** contiene los protocolos de TCP y UDP[2] . El primero se denomina orientado a la conexión, es decir, por cada grupo de datagramas que se envíen, se espera que el receptor envíe un mensaje de conformidad (en jerga técnica, un *ack*) para poder enviar el grupo siguiente. En una red a escala global como Internet, con una alta tasa de perdida de paquetes, el mecanismo de respuesta es muy útil. Sin embargo, en redes LAN el protocolo TCP tiende a ser muy lento, ya que demasiados tiempos de espera por respuestas, no se justifican. En este sentido, el UDP es un protocolo que no espera respuesta, lo cual lo hace menos efectivo en redes de alta tasa de perdida de paquetes, pero muy rápido.

Finalmente, la **capa de aplicación** es llamada de nivel superior, por la cantidad de detalles que incluye con respecto a las otras. Aquí se se encuentran protocolos como el http, SMTP[3], POP3[4]. Algunas aplicaciones frecuentemente usadas para la

[2]User Datagram Protocol
[3]Simple Mail Transfer Protocol
[4]Post Oficce Protocol 3

conexión remota son: FTP[5], TELNET y SSH [6].

4.2. Direcciones IPV4 e IPV6

Desde que se implementó el protocolo, se han usado las IP como dirección a la cual remitir la información que se ha solicitado. En su versión 4, el protocolo definió el uso de 4 números de un byte cada uno para las IP. A esta versión se le conoce como IPV4 y se consideró que con esa cantidad de números era posible identificar cualquier dispositivo en Internet, por cuanto la cantidad de combinaciones que ofrecían los 4 número era de 2^{32} bites o 4 mil millones de direcciones aproximadamente, es decir, suficiente para todo el mundo.

En IPV4, la dirección se estructura en cuatro números separados por un punto. Cada número tiene asignado un byte para su almacenamiento, por lo que el rango de posibilidades va de 0 a 255, con o que una IP adopta una forma como 146.92.83.1, la cual se conoce como *"Common Internet Address Notation"*.

Con el correr de los años y el crecimiento exponencial de Internet, las direcciones IP comenzaron a sufrir una merma importante, debido a su alta demanda durante su proceso de asignación. En ese tiempo, se comenzó a pensar en una drástica modificación del protocolo, pero la expansión de Internet y lo compleja que ya re la red, hacía contraproducente cualquier cambio, a menos que fuese gradual, pero tardaría muchos años.

La otra solución vino de la mano de preguntarse ¿cuándo realmente se requiere del uso de una IP única a escala global?. Simple: cuando se sale de la red local y se entra en Internet. Entonces, era posible que los dispositivos de interconexión manejen una dirección IP real o pública, pero en la red interna sólo se asignen IPs privadas, es decir, no existen en Internet.

[5]File Transfer Protocol. Se utiliza para descargar archivos de la red, por ser un medio rápido y eficiente en esta materia.

[6]Secure Shell

En la Fig. 4.2 se puede observar la interacción de distintos equipos que tienen IP para su identificación. El router, por ejemplo posee dos IP, una privada para la red interna y otra pública para salir a Internet.

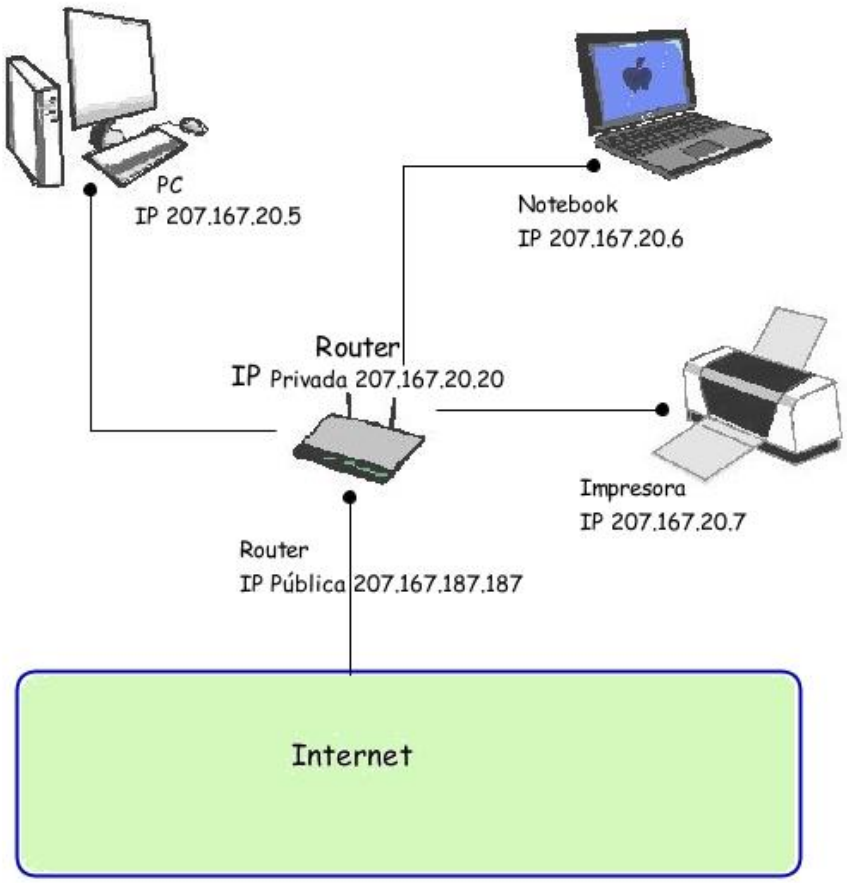

Figura 4.2: Interacción de equipos de interconexión e Internet.

La estructura de un datagrama en IPV4 posee una estructura rígida en su encabezado, como se puede apreciar en la Fig.4.3. La primera fila del datagrama contiene la versión del protocolo (en este caso IPV4), el IHL[7](cuyo valor mínimo es 5, correspondiente a 160 bits o 20 bytes), el tipo de servicio y el largo total del datagrama en

[7]IP Header Length o Longitud del encabezado de la dirección IP, el cual por ser de tipo IPV4 se tiene en unidades de 32 bits

bytes. El tipo de servicio se refiere a dos cosas principalmente: la prioridad y condiciones de la forma de transmisión del mensaje. La prioridad se mide en una escala que va de 0 a 7, con 7 la máxima prioridad, mientras que la transmisión puede tener sólo atrasos cortos o bit D[8] para envíos rápidos, transmisiones de alto rendimiento o bit T[9] y bit R[10] para la minimización de errores en el envío, como perdidas, duplicaciones o daños.

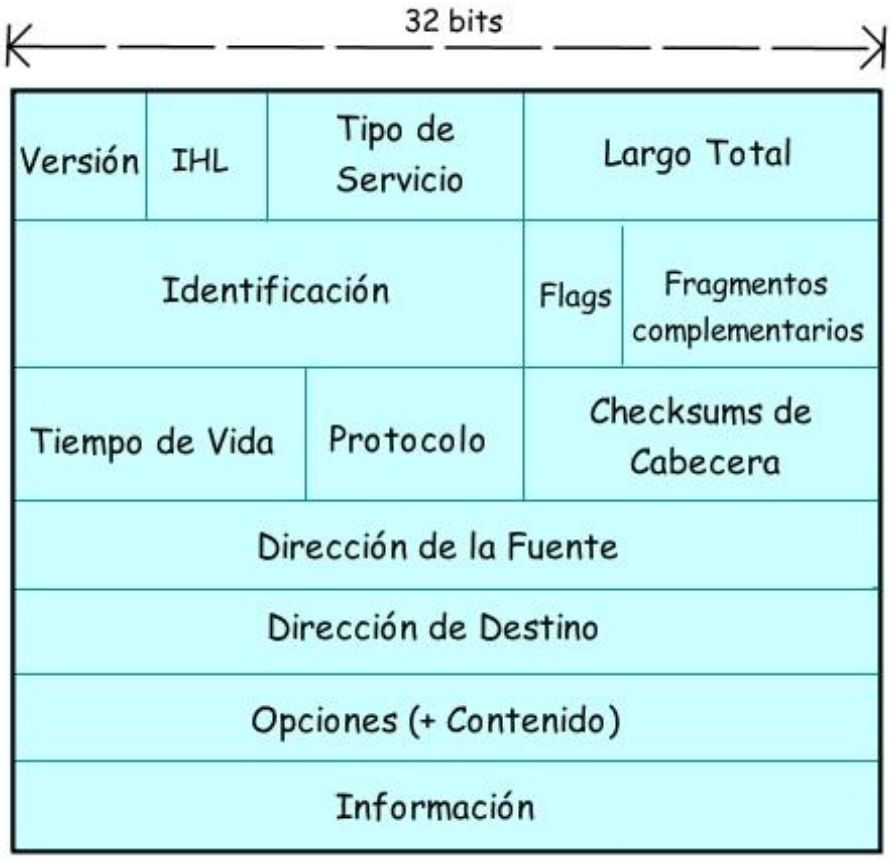

Figura 4.3: Datagrama en versión IPV4.

La segunda fila, contiene la identificación. El flag indica si fue fragmentado con la

[8]Se refiere a bit D por delay
[9]T por Throughput, en el sentido de enviar una gran cantidad en el menor tiempo posible
[10]El bit R, se denomina así por *"Reliability"*

sigla MF[11] y si es el último fragmento o no se puede fragmentar con la sigla NF[12] y en el último campo de esa fila, los fragmentos complementarios, que explican en qué posición debe ubicarse el datagrama para que la información se pueda reestructurar una vez que llegue a destino.

En la fila siguiente aparece el campo de *"tiempo de vida"*[13] el cual es un contador que va decreciendo cada vez que el paquete pasa por un router. El contador comienza en 255, por lo que si llega a cero, la única posibilidad es que el paquete se haya perdido o este en un loop dando vueltas por Internet (las mayores distancias entre dos dispositivos, contada en número de routers no sobre pasa 50). Luego aparece el campo checksum de cabecera, el cual indica si los datos han sido dañados o contienen errores.

Las dos filas siguientes son la IP de origen y destino respectivamente. Cabe destacar que ambas ocupan 32 bits para ser codificadas (4 bytes de 8 bits cada uno) y lo hacen en forma fija, es decir, si se deseara ampliar la cantidad de bits presentes en las IP, el cambio es no menor, por cuanto habría que reconfigurar todos los equipos de interconexión que, por software, reconocen en ese sector del datagrama la posición de las IPs.

Por construcción, los números presentes en una IP sirven para identificar al dispositivo en una red y a la red dentro de Internet. Durante la etapa larvaria de Internet, se asignaron dos números a la parte red y los otros dos a la parte host, lo cual rapidamente demostró no ser eficiente. Como medida paliativa, se crearon las clases en IP, tal como lo muestra la Fig. 4.4.

Para que el datagrama pueda ser encaminado, el router debe conocer el tipo de red al que se dirige, por lo que los primeros bits de una dirección son usados para obtener dicha información Así, las de clase A se identifican por tener un cero al comienzo, mientras que las de clase B, un 10.

[11]MF viene de *"More Fragments o Más Fragmentos"*
[12]Significa "No Fragmentar"
[13]TTL o Time To Live, indica cuando el paquete ya debe ser retirado de su circulación

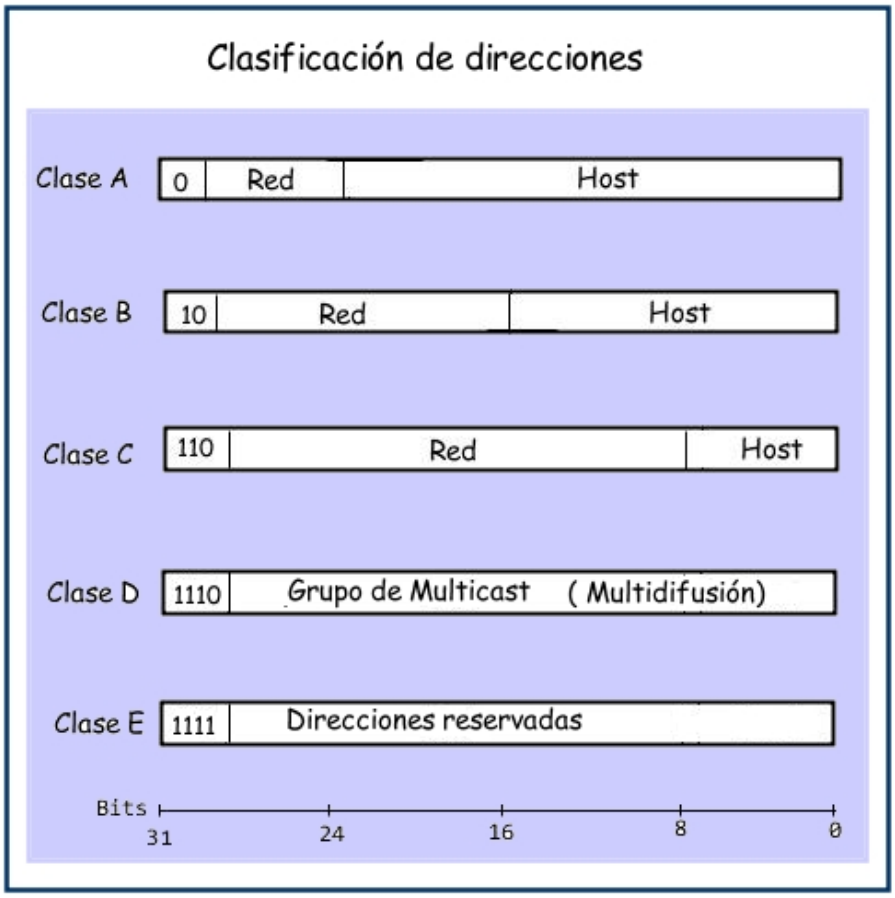

Figura 4.4: Tipos de datagramas en IPV4

Las de clase A son las más grandes, en cantidad de dispositivos que pueden albergar, y sólo se han entregado de forma muy justificada a grandes organizaciones, por cuanto se está hablando de 2^{24} o 16.777.216 direcciones, ya que el primer byte se usa para codificar la red y los otros tres para los dispositivos dentro de ella, tal como se aprecia en la Fig. 4.4. El espacio de direcciones clase A va desde 1.0.0.0 hasta 126.0.0.0.

Las de clase B tienen capacidad para 2^{16}, por cuanto los dos primeros bytes se usan para codificar la red, siendo el rango de direcciones desde la 128.1.0.0 hasta 191.254.0.0. Por lo general este tipo de redes fueron entregadas a grandes compañías o universidades.

En tanto que las de clase C, tienen capacidad de 2^8 direcciones para los dispositivos, ya que los tres primeros bytes se usan para codificar la red. De esta forma, el rango de IPs va desde 192.1.1.0 hasta la 223.254.254.0 y por lo que se otorgan a compañias medianas.

Las de clase D y E son particulares. Las de D, ya no están dentro de la denominación de redes sino de grupos de direcciones. En la clase D, cuando se transmite un datagrama, es recibido por todos los host o estaciones que comparten la dirección multicast. La aplicación que se puede obtener con esta clase es, por ejemplo, una video conferencia o televisión por IP. Las IP de estos grupos pueden ir de 224.1.1.1 a 239.254.254.254. Las de clase E, en cambio, se encuentran reservadas y van desde 240.1.1.1 hasta 254.254.254.254.

Aunque los esfuerzos por asignar eficientemente las IPs han sido heroicos, siempre se supo que era una solución temporal, por cuanto lo que realmente se debe hacer es un cambio drástico en el protocolo, lo que fue realizado en el IPV6, el cual más que un reemplazo es una evolución natural del IPV2, por ejemplo se mantiene la idea de dividir la información en datagramas, que viajan independientes unos de otros, en tanto que, soluciona la falta de direcciones IP al proporcionar un campo de 128 bits, es decir, 2^{128} combinaciones. Además, existen aspectos del diseño que permiten mejorar el rendimiento del IPV6 entre las redes: la cabecera de la IP es mas simplificada, reducción del espacio y espacio fijo, no se permite la fragmentación intermedia, sólo la fragmentación que se realiza en el origen, generando un aceleramiento del procesamiento de los datagramas. Se puede observar en la Fig. 4.5, que los nuevos datagramas poseen 8 campos, en vez de los 14 que componían los datagramas IPV4.

En el campo *"Clase de tráfico"* se indica la prioridad del datagrama. En *"Límite de saltos"*sirve para reemplazar al campo *"Tiempo de vida"* fijando un límite a la cantidad de de saltos, numero que decrece en una unidad a medida que pasa de un punto a otro, eliminándose cuando llega a cero y evitando, de esta manera, que queden datagramas dando vueltas por la red indefinidamente.

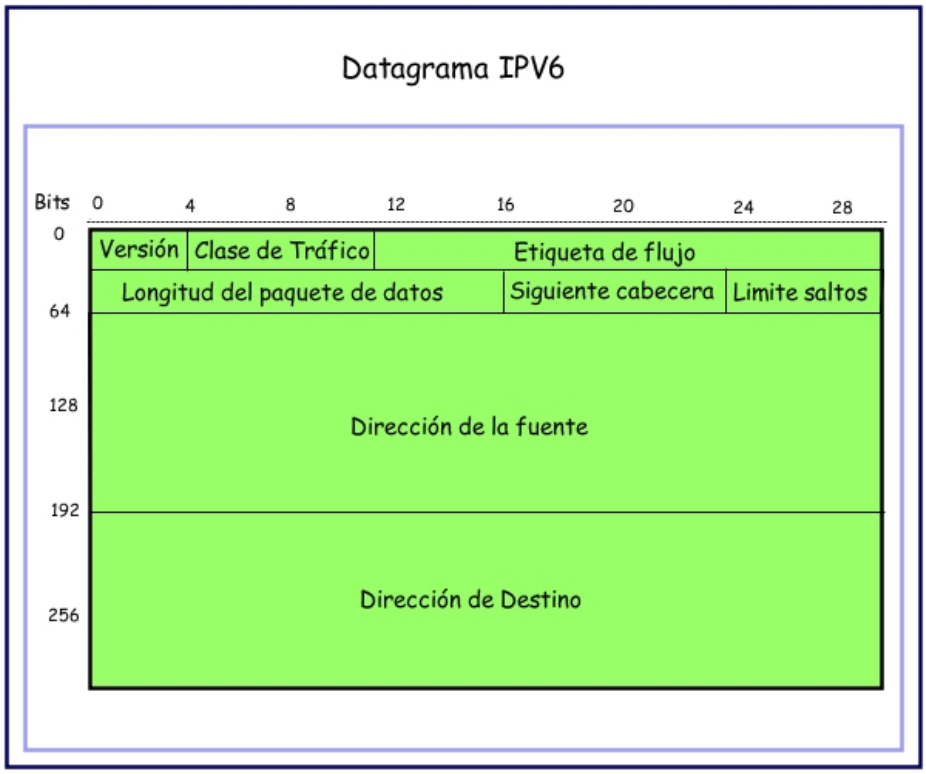

Figura 4.5: Datagrama de dirección IP en versión 6.

La dirección en IPV6, se expresará de distinta forma a la usada en IPV4. Ahora la IP estará compuesta de ocho grupos de cuatro caracteres cada uno, separados por dos puntos, y compuestas por un prefijo de 64 bits y un identificador de interfaz también de 64 bits, mientras que los grupos se expresaran en notación en hexadecimal. Una dirección válida será, por ejemplo, 2001:DB8:0:2F3B:2AA:FF:FE28:9C5A. En cuanto a la ubicación de un servicio por parte de los usuarios, se mantendrá el concepto de DNS.

Se añadirán dos mecanismos de seguridad: cabecera de autentificación y cabecera de encapsulamiento seguro. El primero provee de autenticación e integridad a los datagramas, ya que se asegura que un paquete este viniendo realmente de un emisor indicado en la dirección de origen del datagrama. El segundo provee mecanismos de integridad y confidencialidad a través de campos donde se envían las llaves de encriptación para permitir una conexión segurizada.

En resumen, se puede decir que las ventajas del direccionamiento IPV6 son [24]:

- Mayor número de direccionamientos (2^{128} direcciones).
- El ruteo es jerárquico y por esto es más eficiente que el anterior.
- Mayor seguridad al contar con la autentificación y encriptación de los datos.
- Adaptación de nuevos protocolos (se basa en el protocolo de Internet).
- Mayor espacio para llevar información.

4.3. 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[14] , 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 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

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

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 qué 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.

4.4. Ruteo de datagramas

El TCP/IP es un protocolo que fue diseñado para que todo lo que se desee transmitir pueda llegar a destino. En ese sentido, funciona casi como el juramento de los antiguos carteros (ni el viento, ni la nieve, ni .... etc.) tratando de hacer el mejor esfuerzo para lograr su cometido. Sin embargo, esta idea utópica también falla y las razones son muy numerosas. Sin embargo, son tantos los beneficios que los errores parecen despreciables.

En TCP/IP se manejan tres principios:

- La conectividad es un fin en si misma. Un datagrama siempre será ruteado, independiente del router que se encargue de su envío parcial mientras navega hacia su destino.
- Los datagramas contienen la información necesaria y suficiente como para que puedan ser ruteados en Internet.
- La inteligencia está en las puntas. Son las aplicaciones de envío y recepción de los datagramas las que toman la decisión de si aceptan/rechazan una transmisión, reagrupación de los datagramas, solicitud de reenvío de datos, etc.

Toda transmisión en TCP/IP, requiere que ya sea de forma manual o automática, por ejemplo usando un DHCP[15], los dispositivos que intervienen en la comunicación tienen que tener asignada una dirección IP válida, ya sea esta una privada, en el caso de una red local, o pública si se desea ser visto desde cualquier lugar de Internet.

Hay que recordar que en una comunicación entre dispositivos, lo que en realidad sucede es que algunas de las aplicaciones de software son las que están enviando/recibiendo información. En ese sentido, el siguiente paso es que ambas aplicaciones puntas acuerden cómo se realizará el envío recepción de los datagramas.

Debido a que lo más probable es que el archivo a enviar exceda la capacidad de portabilidad de los datagramas, la aplicación punta emisora resolverá transformarlo en un conjunto de paquetes, cada uno de ellos con una marca de secuencia. A este proceso se le conoce comúnmente como *"paquetización"* y es el paso previo antes del envío a través de la red local y luego por Internet.

El proceso continua con la identificación de la ruta que deben seguir los paquetes. En ese sentido, si el dispositivo receptor se encuentra dentro de la misma red que el emisor, entonces, se realiza un envío directo, comúnmente usando protocolo UDP. Si este no es el caso, el dispositivo emisor contacta al router de la red y le traspasa el paquete para que lo envíe a la red que corresponde, basándose en la IP de destino.

La Fig. 4.6 muestra el proceso por el que pasa un mensaje que se quiere enviar a una dirección específica, donde el protocolo TCP lo separa en partes más pequeñas y le añade una cabecera con información relevante como el orden en que deben unirse las partes del mensaje. Otro dato relevante es la suma de comprobación (checksum), que viene a ser el total de datos que contiene ese paquete, y que sirve para finalmente verificar que se tengan todos los datos iniciales o si se perdió información durante la transferencia.

Una vez que los datagramas han abandonado el router de la red de origen, comien-

[15]Dynamic Host Configuration Protocol

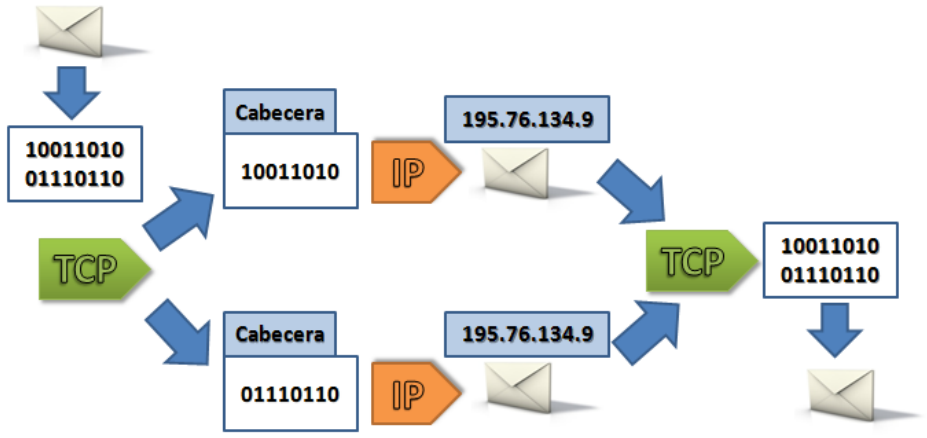

Figura 4.6: Funcionamiento del protocolo TCP/IP

zan su viaje en Internet, hasta llegar al router de la red de destino, el cual revisará si existe la IP del dispositivo receptor, en cuyo caso enviará finalmente el paquete.

Finalmente, será la aplicación receptora la que comenzará a reagrupar los datagramas y a enviar peticiones de reenvío, en caso de que falte alguno o varios de ellos, o simplemente existan errores en los datos recibidos

En la descripción anterior, existen muchos supuestos para que realmente funcione el TCP/IP. Uno de ellos es que los routers *"sepan"* hacia donde enviar los datagramas. En parte este conocimiento viene dado por la configuración de las llamadas tablas de ruta del router, verdaderos mapas camineros que le indican al dispositivo de interconexión hacia dónde enviar el datagrama. Evidentemente el proceso no está exento de errores, retardos y problemas varios. Sin embargo, su mejora continua y bajos costos han hecho de Internet el medio por excelencia para lograr una comunicación a escala global.

4.5. Registro de nombres de dominio

Las direcciones IP son difíciles de recordar, incluso para los usuarios más experimentados, amén de que en un contexto comercial serían poco prácticas, sobre todo en una campaña publicitaria. Ante estos problemas, y existiendo la necesidad de identificar a un dispositivo en las redes locales, se recurrió recurrió a un método simple de asociación de un nombre a la dirección IP, para lo cual, cada computador en la red mantenía un archivo con el listado de los nombres de los demás computadores y la IP asociada[16].

El método anterior, no estaba exento de problemas, siendo el más común, la actualización de la tabla de nombres cuando se incorporaba un nuevo computador, lo cual implicaba intervenir cada una de las tablas existentes en cada uno de los computadores de la red, complicando el tan ansiado *"plug and play"*, es decir, conectar un dispositivo y usarlo de inmediato.

Si lo anterior ya era un problema grande, el concebir la actualización de las tablas de nombre a escala global, era simplemente impracticable. Surge como alternativa solución, muy eficiente por lo demás, la creación de una gran base de datos distribuida por toda la Internet, la cual se basa en la creación de dominios de nombres, que a su vez permiten delegar subdominios a quienes se inscriban *"debajo"* de ellos, manteniendo una estructura jerárquica. De esta forma nacen los dominios primarios, tales como **.mil, gov, .cl, ar.** etc. y debajo de ellos se van creando y delegando sub-dominios, como por ejemplo **uchile.cl**. Esta delegación implica que **.cl** conozca la IP de **uchile**, pero no sepa nada acerca de una ip de algún subdominio debajo de **uchile**, por ejemplo, **derecho.uchile.cl**.

Algo que debe quedar en claro, es que las aplicaciones en TCP/IP NO *"entienden"* de nombres, es decir, requieren de que se produzca su transformación en una IP valida para poder enviar los paquetes de datos entre dos computadores. Al proceso anterior

[16]En el S.O. Unix, este archivo se encuentra en /etc/hosts

se le conoce como *"resolución del nombre"*.

Evidentemente, la mantención de esta base de datos distribuida se realiza a través de aplicaciones informáticas que automatizan el proceso de resolución del nombre, conocidas como DNS (Domain Name System). Estas aplicaciones se encargan de resolver el nombre de un computador en la red local y de articular todas las consultas necesarias hacia otros DNS pertenecientes a otras redes, con el fin de obtener la IP asociada a un determinado nombre [60].

Los nombres de dominio se registran y asignan por el NIC (Network Information Center) de cada país. Por su asociación directa con la marca institucional, es común que las empresas inscriban su nombre como dominio dentro de un NIC y de esta forma puedan ser fácilmente reconocidas en el mercado digital. Lo anterior plantea un problema: ¿qué sucede si dos o más empresas desean inscribir un mismo nombre de dominio debajo de un NIC?. Al respecto, NIC Chile, organismo perteneciente a la Universidad de Chile y a cargo del dominio .cl, ha generado una normativa interna, la cual considera un sistema de arbitraje cuando un dominio se haya en conflicto por poseer dos o más interesados.

El funcionamiento de los DNS se estructura esquemáticamente como un árbol en cuyas ramas se delegan los subdominios, como se muestra en la Fig. 4.7. Entonces, cuando una aplicación desea resolver un nombre, primero consulta al DNS de la red local. Si este desconoce el nombre a resolver, contactará al DNS primario que tenga configurado, el cual generalmente es del país donde físicamente reside el computador de la aplicación que está solicitando la resolución. Si hasta ese punto aun se desconoce la IP asociada, entonces se investiga qué NIC posee el dominio primario asociado al nombre y se continua la consulta a ese DNS, el cual dará la dirección del subdominio asociado, para que la resolución siga un camino de preguntas hasta encontrar el DNS que conoce la IP asociada al nombre que se intenta resolver.

Los NIC primarios, fueron asignados a instituciones sin fines de lucro, por lo general universidades e institutos de investigación, por la IANA (Internet Assigned

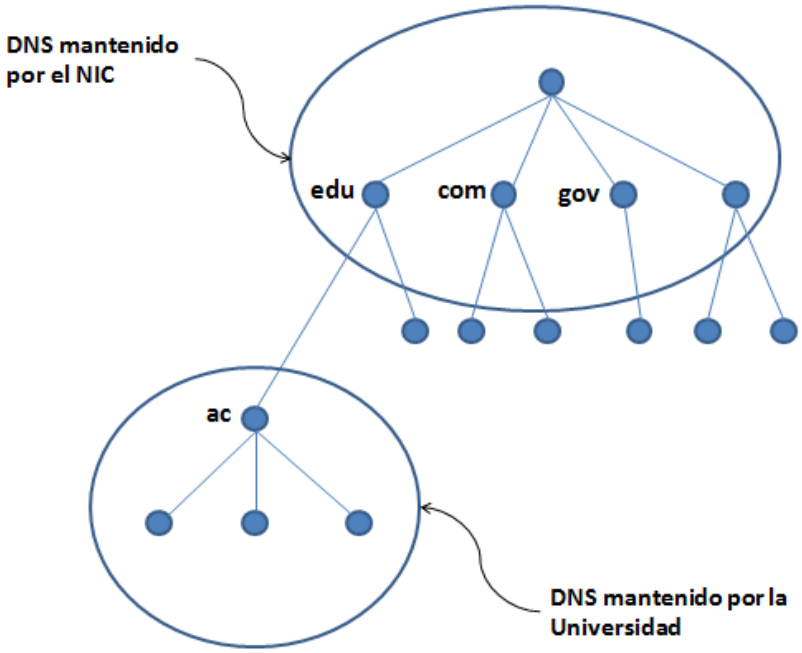

Figura 4.7: Jerarquía de un DNS

Numers Authority, actualmente ICANN) con el propósito de llevar el registro de los nombres de dominio. En el caso chileno, después de un comienzo muy tímido y austero, NIC Chile impulsó una campaña muy asertiva para que empresas e incluso personas naturales inscribieran dominios bajo el .cl. Es así que la cantidad de dominios inscritos ha crecido exponencialmente, como se muestra en la Fig.4.8.

El funcionamiento del DNS se puede ilustrar mejor con el siguiente ejemplo: *El usuario derinf@derecho.uchile.cl desea enviar un archivo de 6Mb vía correo electrónico a la cuenta lorena@law.inf.int.sk. Sabiendo que el ancho de banda efectivo de la conexión del usuario emisor es inferior a 500 Kb. ¿cómo se realizaría el ruteo del mensaje para que el archivo llegue correctamente a la cuenta de destino?.*

Para responder esta pregunta, hay que recordar el el funcionamiento del protocolo TCP/IP. En primer lugar, dado que el mensaje a enviar contiene un archivo muy pesado (6Mb), se debe paquetizar, es decir, transformar en un conjunto de datagramas de un tamaño mucho menor.

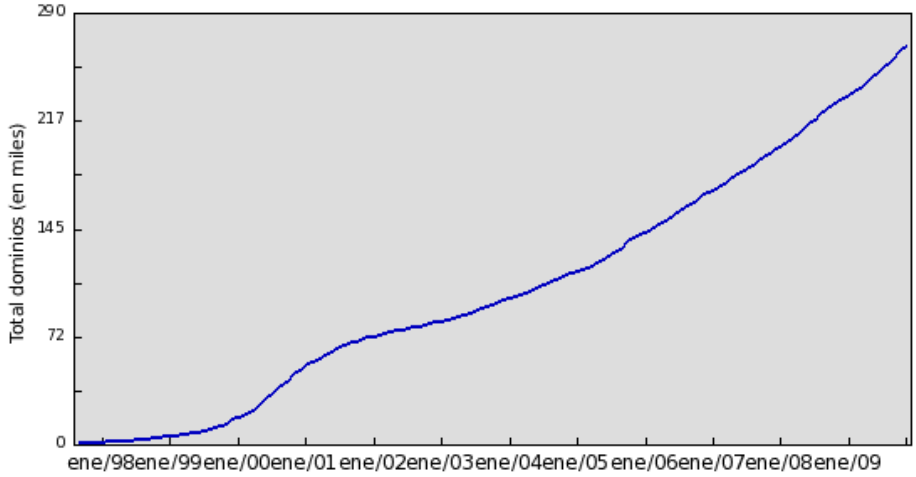

Figura 4.8: Total de nombres en dominio .cl
Fuente: http://www.nic.cl/stat/inscritos.html

Cada datagrama se creará con un identificador que permita luego su concatenación en el lado receptor. Luego se especifica en el encabezado la aplicación transmisora, que en este caso es el SMTP (Simple Mail Transfer Protocol) [46] del correo electrónico. Continua con el nivel de transporte y finalmente con las IP de origen y destino. Es en este punto donde se hace necesario resolver el nombre para **derecho.uchile.cl** y **law.inf.int.sk**. El primer nombre se resuelve de forma casi inmediata, pues es parte del dominio de la red interna de la institución donde se encuentra la aplicación transmisora. El segundo nombre es un poco más complejo de resolver, por cuanto intervienen varios DNS.

El SMTP transmisor, preguntará al DNS de la red local si conoce el nombre law.inf.int.sk por cuanto existe la posibilidad de que otra aplicación de la misma red haya hecho esa pregunta con anterioridad, en cuyo caso ya estaría resuelto en nombre y almacenado en la base de datos de este DNS. Como lo más probable es que el proceso anterior no haya sido aplicado, el DNS le contestará al SMTP con la IP del primario a nivel país, en este caso NIC Chile, el cual posee a IP de todos los primarios del mundo, en particular la del NIC de **.sk**.

La resolución del nombre continúa con la consulta a NIC de .sk respecto de la IP de **.int**, para luego hacer la misma pregunta al DNS de **.int** respecto de la IP de **.inf**. Una vez resuelta esa parte del nombre, se le consulta al DNS de **.inf** por el computador de nombre **law**, con lo cual se ha resuelto todo el nombre para **law.inf.int.sk**.

Resueltas las IP de origen y destino, se comienza con el envío de los datagramas, tal como lo muestra la Fig. 4.9 [46]. Cada vez que se quiere enviar un correo electrónico, la aplicación cliente (por ejemplo Outlook Express) interactúa con el servidor de correo (SMTP) para que gestione el envío, indicándole la dirección de destino y de origen del correo, así como el contenido del email. En el caso de que la dirección de destino se encuentre en el mismo dominio que la dirección de origen, sólo se utiliza un protocolo llamado POP3 quien recibe el mensaje y lo coloca en la bandeja de entrada del usuario destino. En caso contrario, el SMTP debe comunicarse con un servidor de dominio o DNS correspondiente de manera que le indique la dirección IP del servidor SMTP de destino. Cabe recordar que será la aplicación de destino, en este caso el SMTP receptor, el que se encargará de concatenar todos los datagramas, solicitar los faltantes e incluso de rechazar el archivo.

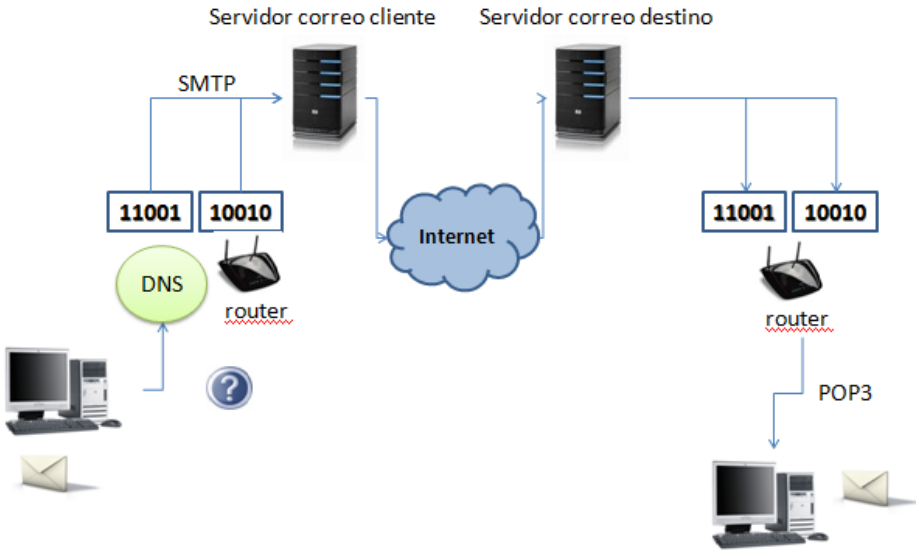

Figura 4.9: Envio de un mail vía SMTP

4.6. Conectándose a Internet

En los comienzos de Internet, el acceso se lograba a través del pago de un enlace exclusivo a la gran red, el cual era compartido por todos los dispositivos pertenecientes a una institución, lo que hacia prácticamente inviable que un usuario común lograra algún tipo de conectividad de desde su casa.

Con el nacimiento de los Internet Service Provider (ISP), comienza la verdadera gran explosión de Internet. Miles de usuarios lograban conectividad a precios razonables, o al menos mucho más bajos que los que se cobran por el uso de enlaces dedicados.

En esta sección, se revisarán algunas de las formas más utilizadas por los usuarios no corporativos para acceder a Internet, por cuanto se trata del público más numeroso y es justamente el que posee menos medios para defenderse contra la vulneración de su privacidad y la restricción a su navegación en la Web.

4.6.1. Dial Up

Una de las primeras formas para todo público de acceso a Internet, se realizó utilizando la capacidad instalada de las redes de telefonía, tal como se muestra en la Fig. 4.10, donde el computador, utilizando un dispositivo llamado modem (modulador/demodulador) realiza una llamada convencional a un número provisto por el ISP. A continuación, el modem comienza la transformación de los datos a transmitir en pulsos audibles, por cuanto se está utilizando el el rango de frecuencias que se utilizan en las conversaciones telefónicas, lográndose velocidades de transmisión cercanas a los 56Kbps.

Aparte de la baja velocidad en transmisión de datos, el esquema Dial Up tiene el problema de que deja tomada la línea telefónica mientras dura la conexión a Internet, por lo que rápidamente su uso fue desplazado por nuevas formas de acceso, más rápidas

y que permiten utilizar la línea de teléfono al mismpo tiempo que Internet.

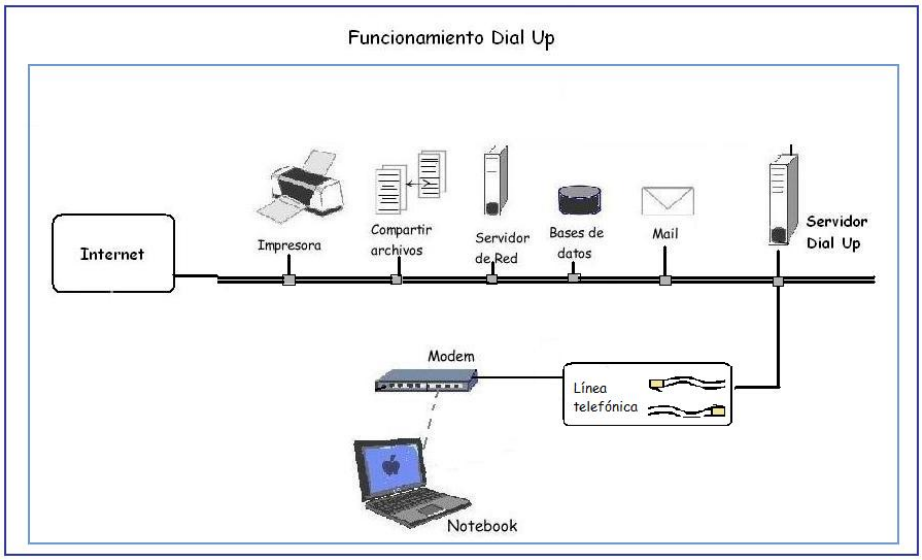

Figura 4.10: Funcionamiento del Dial Up.

4.6.2. ADSL

El ADSL o Línea de Suscripción Digital Asimétrica es una forma de conexión de alta velocidad que utiliza la línea telefónica convencional para la transmisión de datos, utilizando el ancho de banda que no ocupa comunicación telefónica convencional. En efecto, por construcción, una línea telefónica sólo utiliza parte del ancho de banda para la transmisión de la voz. El dispositivo ADSL, se encara de la utilización de la otra parte del ancho de banda, lo cual permite una conexión permanente, con velocidades de 1,5 a 6 Mbps de bajada (recibiendo datos) y 16 a 576 Kbytes/seg de subida (enviando datos). Por construcción este esquema permite simultáneamente los servicios de telefonía e Internet.

En la Fig. 4.11 es posible observar el funcionamiento del ADSL. Con esta configuración se crean tres canales. El primero de alta velocidad desde la red hacia el abonado.

El segundo, es de tipo duplex, con información que viaja en ambas direcciones y por último, el circuito telefónico convencional.

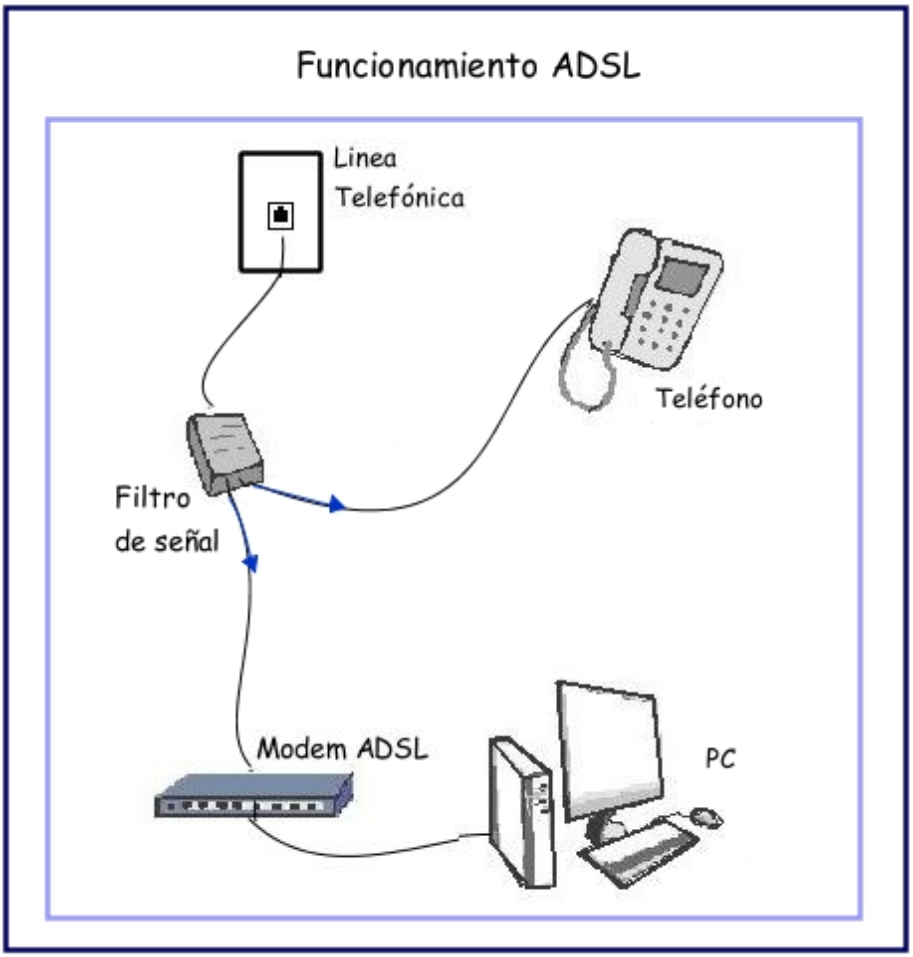

Figura 4.11: Funcionamiento de ADSL.

4.6.3. Conexiones inalámbricas

Este tipo de conexiones se ha hecho cada vez más populares entre los usuarios, sobre todo por sus características básicas: ubicuidad, precios razonables, ancho de banda adecuado, existencia de una red fija para la transmisión de datos y mejora continua de la tecnología.

Para que un dispositivo logre conexión a redes con tecnologías inalámbricas, es necesario que posea una tarjeta de red que permita la transformación de los datos en señales electromagnéticas.

Para efectos de estudiar la conexión del usuario no corporativo a Internet, sólo interesan aquellas tecnologías inalámbricas clasificadas como de última milla, de entre las que se destacan:

Wimax. Es la sigla de Worldwide Interoperability for Microwave Access (interoperabilidad mundial para acceso por microondas). Esta tecnología realiza la transmisión de datos usando ondas de radio. Su principal ventaja es que permite dar acceso a un canal de datos a sectores donde la densidad de usuarios es muy baja, lo que haría inviable, del punto de vista económico, el uso de otros medios de transmisión como cables o fibra óptica. Para garantizar el estándar y la interoperabilidad, se creó Wimax Forum, organismo encargado de la certificación de los equipos que proveen distintos fabricantes, los cuales están calibrados para funcionar entre las frecuencias de 2,5 y 3,5 Ghz.

Internet móvil. Aquí se agrupan todas las tecnologías que han permitido la conexión inalámbrica de un dispositivo en movimiento, el cual por excelencia ha sido el teléfono celular. De esta forma, se denota a la primera generación de teléfonos y redes celulares (1G) , donde cada aparato pesaba casi un Kilogramo o mas de peso y parecían un ladrillo en su forma. Esta tecnología tuvo su apogeo en la década de los 80's.

La evolución de la tecnología permitió digitalizar al 100 % las redes celulares análogas, con lo cual se pudieron introducir servicios de datos, como la mensajería entre teléfonos moviles, también conocido como SMS y los correos electrónicos. Con lo anterior ya se entró en las redes de segunda generación (2G).

El tercer paso en esta saga, lo constituye el lograr mayor velocidad en la transmisión de los dato que no son voz. Es así como las empresas de telefonía celular, reservan en su ancho de banda consecionado, un trozo de aproximadamente

5Mbps para la transmisión de datos y el resto para la voz. Nacen las redes de tercera generación (3G) orientadas a proveer servicios de correo electrónico, navegación en la Web, GPS, etc., en el fondo todo lo que el TCP/IP pueda brindar, En gran problema en 3G es el escaso ancho de banda efectivo para los usuarios, lo cual obliga a las empresas proveedoras del servicio a realizar fuertes inversiones en antenas celulares para que no colapse el canal de datos.

Por último, y no quiere decir que aquí se detendrá el desarrollo tecnológico, la cuarta generación (4G), promete incrementar significativamente la tasa de transmisión de datos. Aunque está solo en etapa de experimentación, se espera que sea 100 % basada en IP y que alcance velocidades de transmisión de 100 Mbps a Gbps. En nuestro país, específicamente en la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, se acaban de realizar las primeras pruebas de conectividad, con tasas de transmisión superiores a los 50Mbps, lo que se conoce como hiperbanda ancha móvil [63].

WiFi. Aunque se trata de una tecnología inalámbrica para la creación de redes LAN (ver detalles de su operación ver sección 3.1.4) , existen casos donde esta tecnología puede ser considerada de última milla, por ejemplo cuando se está en un lugar donde se cobra por el uso de la red inalámbrica.

Por las características de la Internet Inalámbrica, esto es cobertura a sectores de difícil acceso y conectividad prácticamente directa, sólo se necesita la tarjeta de red adecuada, muchos países han desarrollado políticas de Estado para asegurar que esta forma de acceso a la red sea totalmente gratuita, por el alto valor social que posee. En efecto, en Corea del Sur por ejemplo existen dos ciudades con conectividad 100 % gratis utilizando alguna de las tecnologías móviles inalámbricas. Aunque aun es prematuro hablar del impacto económico y social, ya se han realizado algunos estudios que muestran que la tecnología ha sido correctamente asimilada por los usuarios y que el bajo costo o incluso su gratuidad ha devengado en una mayor conectividad, potenciando los negocios electrónicos y los nuevos emprendimientos [7].

4.6.4. Los Internet Service Providers (ISP)

El servicio de conectividad que brindan los ISP, posee un costo que depende, entre otros factores, del ancho de banda solicitado por el contratante. En Chile principalmente este servicio es otorgado por empresas ligadas históricamente a la telefonía y las telecomunicaciones, tales como Movistar, VTR, Entel PCS o Telmex. Hay diversas formas para conectarse a este servicio, ya sea mediante ADSL (Asymmetrical Digital Subscriber Line), fibra óptica, vía satélite, dial-up, banda ancha, cable módem o vía inalámbrica (wireless). Los principales ISP se concentran en EE.UU, de ellos podemos apreciar su participación en el mercado en la Fig. 4.12 .

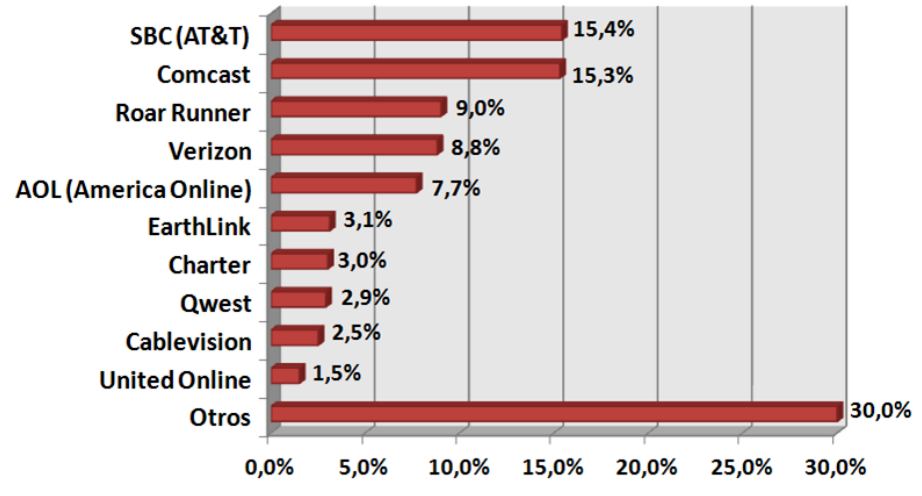

Figura 4.12: Proveedores de Internet en USA
Fuente: Alex Goldman, ISP Planet: Top 23 US ISPs 2008

El costo de conectividad que cobran los ISP en Chile, es alto, comparado con servicios similares que se brindan en otras latitudes[17]. La razón, en parte, viene dada por la baja competencia que se observa en el mercado, lo cual es un incentivo directo a la generación prácticas de índole monopólico.

En la Fig. 4.13 se pueden apreciar algunas de las tarifas que cobran los ISP, para

[17]Ver estudio realizado por www.oecd.org

distintos rangos de ancho de banda. A primera vista, se puede notar que se utilizan formas de conexión como el ADSL, Modem y Móvil.

Proveedor de Internet	Tipo de Tecnología	Velocidad de Bajada (Kbps)	Precio Mensual Plan ($)
Movistar	Móvil	700	9.990
VTR	MÓDEM	1024	11.990
Cmet	MÓDEM	128	14.500
Telefónica del Sur	ADSL	1044	16.990
Entel Pcs	Móvil	200	19.990
Entelphone	ADSL	64	21.476
Telefónica CTC	ADSL	600	23.390
GTD Manquehue	ADSL	1024	23.990
Telmex	MÓDEM	2048	25.500

Figura 4.13: Servicios de conexión que ofrecen distintos ISP en Chile.
Fuente: Ofertas de Acceso Residencial a Internet de Subtel (Marzo 2009)

4.7. Institucionalidad jurídica de Internet y construcción de protocolos

Internet nace en un contexto científico, no comercial, donde la intención primaria es el compartir conocimiento, a través del envío de documentos y en realidad cualquier tipo de comunicación. Por lo tanto, más que un dueño, la Net posee organismos gestores, controladores y fiscalizadores. En concreto, es el Gobierno de Estados Unidos quien posee una organización privada, sin fines de lucro, llamada ICANN (Internet Corporation for Assigned Names and Numbers) y establecida el 18 de Septiembre de 1998, quien es la responsable de asignar las direcciones de protocolo IP con lo cual se permite identificar a un computador en la red de Internet. Antiguamente se llamaba ISOC (Internet Society) y está compuesta de tres organizaciones dedicadas al desarrollo y publicación de los estándares que debe cumplir Internet. A continuación se nombran estas organizaciones:

1. **NIC** (Network Information Center): Organismo encargado de la asignación de nombres dominios de Internet a personas naturales o empresas de tal forma que puedan establecer su presencia en la red. Para cada dominio primario (.cl, .mil, .gov, etc.) existe un NIC responsable por la mantención de los dominios que se soliciten bajo cada primario. Por ejemplo en nuestro país, NIC Chile, institución perteneciente a la Universidad de Chile, es el encargado de la mantención del .cl.

   El buen funcionamiento del sistema dominios e IPs es estratégico para el país. Aun más, de su correcta operación dependerá el desarrollo futuro de la sociedad de la información y de la economía digital. Lo anterior llevó a que el año 2003, a través del decreto número 5, la Subtel creara el Consejo Nacional de Nombres de Dominio y Números IP, con el objetivo de formular recomendaciones sobre las políticas aplicables para la mejora continua de de la red Internet en Chile, en materia de nombres de dominio y de números IP. El consejo está integrado por la Subtel, Subsecretaría de Economía y NIC Chile, junto a diversas organizaciones de la comunidad internetnacional, correspondiéndole a NIC Chile actuar como Secretaría Ejecutiva

2. **ISOC** (Internet Society):

   Es una organización internacional sin fines de lucro, fundada en 1992 cuyo objetivo es proveer la adecuada dirección relativa a los estándares en desarrollo y usados en Internet, para lo cual ha declarado como misión el *"asegurar el desarrollo abierto, evolución y uso de Internet para el beneficio de las personas al rededor del mundo"* [45]. Está conformada por tres organizaciones:

   a) **IAB** (Internet Architecture Board): Consejo de Arquitectura de Internet cuyo objetivo es reglamentar las decisiones sobre estándares que regirán a Internet. Esta organización determina qué necesidades deben ser cubiertas a mediano y largo plazo. También toma decisiones sobre la orientación en tecnología que debe tomar a NET y es la encargada de aprobar/rechazar los estándares y recomendaciones que se proponen a través de la redacción

de los documentos denominados RFC.

b) **IETF** (Internet Engineering Task Force): Grupo de Tareas de Ingeniería de Internet. Se trata de una organización de profesionales técnicos especialistas en obra de ingeniería en telecomunicaciones. Se preocupan principalmente de la mejora continua de los protocolos de comunicación, en algunos casos de darlos de baja, etc. [20].

c) **IRTF** (Internet Research Task Force): Grupo de Tareas de Investigación en Internet. Es una organización hermana a IETF, cuya mision es *"promover investigación de frontera para potenciar la evolución futura de Internet, a través de la creación de grupos altamente especializados de investigadores trabajando en áreas relacionadas con los protocolos, aplicaciones, arquitectura y tecnología relacionada con la Internet"* [20].

3. **IANA** (Internet Assigned Numers Authority): Es la Agencia de Asignación de Números de Internet, que antiguamente se conocía como el registro central de los protocolos de Internet, como puertos, números de protocolo y empresa, opciones y códigos. Esta agencia fue sustituída por el ICANN en el año 1998, la cual es encargada de sus procesos actualmente.

Desde su creación, Internet ha funcionado como una red que no restringe el tráfico generado por sus usuarios sino todo lo contrario, uno de sus principios fundacionales es justamente que la conectividad es un fin en si misma. De esta forma, los organismos de gestión de la red se han preocupado de que su crecimiento apunte hacia brindar acceso universal. Sin embargo, este principio se contrapone muchas veces con los intereses comerciales de los ISP (Internet Service Provider), quienes tienen todo el poder para bloquear, filtrar o priorizar las comunicaciones de sus usuarios.

El principio de *"Neutralidad de Red"* [61], propone que la conectividad debería ser regulada por ley, para prevenir actitudes que vayan en contra de del acceso universal a Internet, por parte de los ISP, los cuales a su vez plantean que cualquier tipo de

regulación, sólo vendría a entorpecer el desarrollo de la red, por lo que es contra producente.

Capítulo 5

Internet y la Web

The challenge is to manage the Web in an open way-not too much bureaucracy, not subject to political or commercial pressures.
Tim Berners-Lee (Inventor de la Web)

Durante años Internet se mantuvo como un desarrollo tecnológico utilizado por unos cuantos privilegiados que conocían complejas instrucciones que permitían el envío y recepción de archivos. Incluso se llegó a pensar que su futuro era incierto, por cuanto la cantidad de usuarios potenciales era muy acotada y no tenía un uso comercial aparente, pasando a ser un invento que sólo le competía al mundo científico, aislándola de las necesidades de las masas.

Con la invención de la Web, se abrieron posibilidades para ampliar el uso la Internet insospechadas para la época. Este nuevo invento, tampoco estuvo exento de críticas, sobre todo del punto de vista de la utilidad que tendría para el usuario común.

El advenimiento de los primeros buscadores de información en la Web, ocasionó un quiebre en la forma en que se puede acceder al conocimiento universal, y en particular a cualquier dato que sea publicado en un sitio.

Desde un punto de vista informático, lo que hace complejo el análisis de la Web, es justamente aquello que la hace tan atractiva: su amplia variabilidad y diversidad

de datos. Se podría decir que en la Web se encuentran todos los tipos de datos de la historia de la computación, por lo tanto, toda herramienta de análisis que se desarrolle, enfrenta un desafío del punto de vista de los formatos, y por supuesto del volumen de los datos.

5.1. Origenes de la NET

Desde sus orígenes, Internet ha marcado un hito tanto en el mundo de la informática como en el de las comunicaciones, convirtiéndose en un medio de difusión a escala global y de paso en un mecanismo para el traspaso de información que potencia también para la colaboración e interacción entre las personas, independiente de su posición geográfica.

Es claro que hoy el mundo depende de Internet casi en la misma forma que lo hace de la telefonía celular, que en los últimos años se ha masificado debido a los avances tecnológicos. La integración que la red ha tenido en la vida cotidiana es tal que si se apagará por tres días, no dejaría indiferente a ningún ser humano, como se aprecia en la Fig. 5.1.

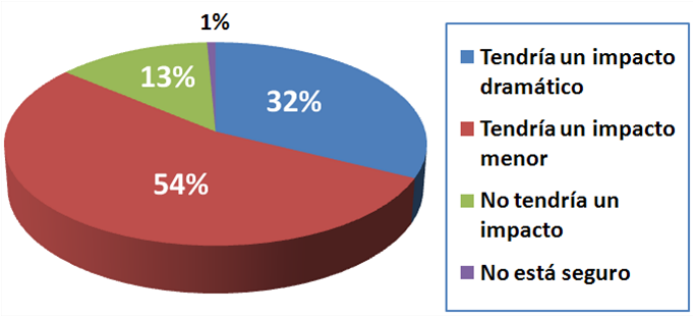

Figura 5.1: Impacto de la falta de Internet si se detuviera 3 días
Fuente: Zogby International (Junio 2009)

Uno de los principios que funda el desarrollo de las redes de computadores, es la

de compartir recursos que son limitados y escasos. Por ejemplo, hace 20 años atrás una impresora por computador era algo poco práctico, dado que los usuarios no estaban todo el tiempo imprimiendo y era mucho mas razonable el compartir este recurso escaso, entre varios usuarios, a través de una red.

Con el avance tecnológico, las economías de escala y la siempre escasa capacidad de procesamiento, las redes de computadores permitieron el procesamiento distribuido de la información, es decir, que cada computador asociado a la red se hiciera cargo de parte de los datos a procesar y en su conjunto, formaran una sola unidad, lo que acuñó el concepto de .[el] computador es la red. Es en este punto surge la idea de compartir los recursos de manera eficiente, a través del trabajo en red o *Networking* [28], lo cual se puede ver en la Fig. 5.2.

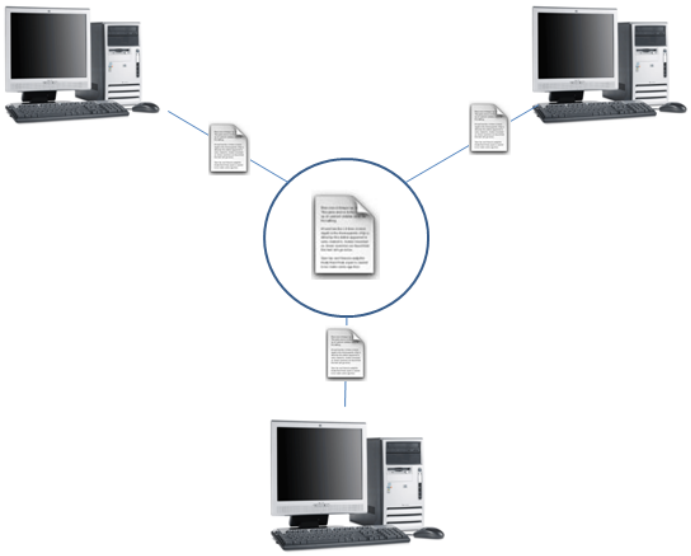

Figura 5.2: Red de computadores

Para que se produzca la comunicación entre computadores y aplicaciones de una misma red, es necesario que se defina un mismo protocolo o lenguaje de *"entendimiento"*. Varios fueron los protocolos de comunicación creados para estos efectos, lo cual planteó un nuevo desafío: ¿cómo conectar computadores de protocolos distintos?

o más aún ¿cómo interconectar redes de protocolos distintos?

Durante la primera década del 60, científicos de la University of California in Los Angeles (UCLA), comenzaron el desarrollo de una nueva familia de protocolos de comunicación destinados a conectar redes heterogéneas. La idea de base era preservar el protocolo imperante en una red LAN, pero cuando se accediera a la red común, se produjera una especie de traducción transparente a un lenguaje común. Nace el concepto de Interconexión de Redes.

Durante esa misma época, la desaparecida Unión Soviética y los Estados Unidos, estaban en una carrera por la supremacía tecnológica en todos los ámbitos, dentro del contexto de lo que era la Guerra Fría. Fue entonces que el país del norte desarrolla en 1969 el *"Advanced Research Projects Agency"* o ARPA con el objetivo de alcanzar los mayores avances tecnológicos. Con el tiempo, se le agregaría un elemento adicional *"Defense"* con lo cual nace DARPA, institución que operó hasta finales de la década de los 80 y que dentro de sus primeros años de funcionamiento enfrentó un desafío bien particular: interconectar sus unidades de defensa, las cuales por construcción poseían protocolos de comunicación distintos.

El problema a enfrentar es qué sucedía si ante un ataque que involucrara armas de destrucción masiva, desaparecía uno o mas centros de comando y control, junto con toda la información que poseían o estaban procesando[1]. Se hacía urgente contar con una red que permitiera a estos centros transferir toda o parcialmente la información que poseían de una unidad a otra, asegurando que tan preciado bien no se perdiera si ocurría una emergencia.

La investigación ligada a los mecanismos de interconexión de redes de protocolos distintos desarrollado en la UCLA, pareció ser la solución para la defensa de EE.UU. El paso natural fue desarrollar una red que interconectara los distintas unidades de DARPA, dando origen, en el año 1966, a ARPANET. Esta red utilizó una técnica de

[1]En lenguaje bélico, cuando uno de estos centros está en problemas, la unidad de ataque/defensa está inutilizada

transmisión muy novedosa para la época: el uso de paquetes de datos. Pero la idea no quedó ahí, muy pronto se trató de unir tanto las redes militares, comerciales y científicas, las cuales son bases para lo que conocemos el día de hoy como Internet.

El número de dispositivos o nodos conectados a ARPANET, creció rápidamente. Ya para el año 1973, se contaba con más de 100 en EE.UU. y la primera conexión fuera del país, con el Colegio Universitario de Londres (Inglaterra).

De ahí en adelante, varias redes fueron surgiendo, respondiendo a la necesidad creciente de mantener comunicación en diversos ámbitos, como BITNET o NSFNET. NSFNET[2] nació para conectar a distintas universidades, bibliotecas y centros de investigación de todo el mundo. El motivo principal era facilitar la conexión evitando los estrictos protocolos de ARPANET. En conclusión, NSFNET y ARPANET, en conjunto originaron fueron las antecesoras de INTERNET, cuyos desarrollos transitaron por caminos un tanto separados. En efecto, ARPANET, por su parte, al marcar el hito de crear la primera red de computadores y NSFNET, al evolucionar para convertirse en INTERNET, termino que comienza a ser utilizado a partir del año 1982 a partir del concepto de Inter-Red (**INTER**connection **NET**work) o *"red de redes"*, las cuales conectan dos o más redes entre sí, de manera que puedan compartir sus recursos.

A partir de la idea original de ARPANET, es decir, *"una sola red"*, Internet plantea múltiples redes independientes que se comunican a través de paquetes y protocolos, que son quienes indican los pasos a seguir de manera que la información se envíe y reciba de la misma manera, y por lo tanto permiten la comunicación entre redes heterogéneas [28].

En Chile se realizó la primera interconexión a nivel nacional, usando el protocolo TCP/IP, en el año 1986, fecha en la cual se realiza el envío del primer correo electrónico entre el Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile al Departamento de Ingeniería Informática (DIINF) de la Universidad de Santiago, cuyo mensaje decía *"Si este mail te llega, abramos una botella de cham-*

[2]National Science Fundation Network

paña" [40]. Desde un comienzo, la adopción de Internet en Chile estuvo enmarcada dentro de un contexto netamente académico, cuyo objetivo era establecer un sistema de comunicaciones expedito entre las universidades chilenas.

A partir de 1987, ya se puede decir que Chile estuvo plenamente conectado a Internet, gracias a un enlace satelital provisto por NASA para la red BITNET y que fue reutilizado. El nodo central de Chile estuvo ubicado en el Departamento de Ciencias de la Computación de la Faculrad de Ciencias Físicas y Matemáticas de la Universidad de Chile. El resto de la historia continua con la expansión de Internet a nivel nacional, con la anexión de más nodos universitarios y posteriormente con empresas que luego serían los primeros Internet Service Providers (ISP).

Desde sus orígenes, tuvieron que pasar varios años para que Internet fuese del uso del común de los usuarios. No fue hasta que aparece la Web, a comienzos de la década del 90, que realmente se masifica su uso, siendo adaptada rápidamente por los usuarios, mucho más que otros medios de comunicación como la televisión o la radio, lo cual se puede apreciar en la Fig. 5.3, donde se ve que por ejemplo la radio tardó casi 70 años en alcanzar un nivel de mercado de 250 millones de usuarios, mientras que el mismo nivel fue alcanzado por Internet en un lapso de sólo 6 años. El rápido desarrollo de la tecnología nos lleva a pensar en mejores versiones de Internet, a partir de nuevas técnicas de conexión que acelerarán la capacidad de transferencia de información, con miras a la educación e investigación académica.

5.2. 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

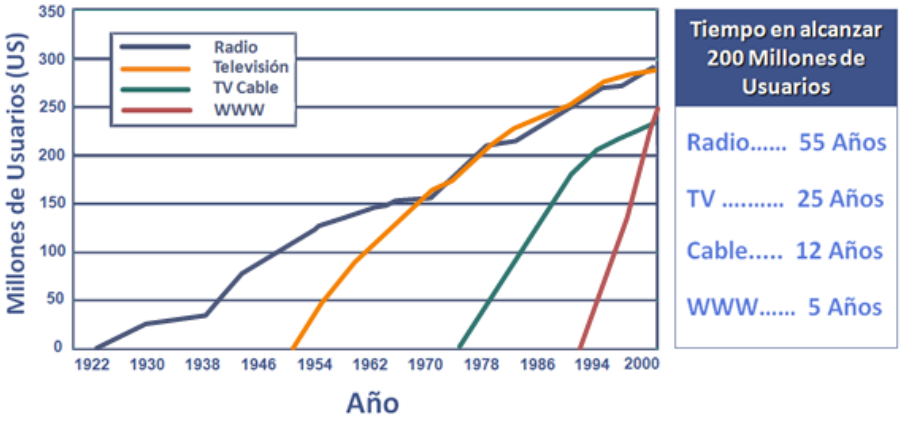

Figura 5.3: Curvas de Adopción de Medios
Fuente: Morgan Stanley, Technology Research, 2009

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"* [4].

Durante varias décadas, el uso de Internet como medio para la transferencia de información entre dos personas, fue considerado un arte mayor, pues era necesario conocer complejas instrucciones para ordenarle al computador una acción de envío o recepción de datos. Lo anterior se complicaba aun más si se considera que de una red a otra, existen computadores, lenguajes y software que son diferentes, es decir la idea de compartir información, quedaba truncada debido a no existía una forma simple de realizar una búsqueda en una jungla tan variopinta.

En el año 1984, Tim Berners-Lee investigador del CERN [3] y consideró que los problemas mencionados anteriormente eran los mismos que ocurrían en contexto de los laboratorios del centro. Por construcción, el CERN dispone de un anillo acelerador

[3]Conseil Europeen pour la Recherche Nucleaire u Organización Europea para la Investigación Nuclear

de partículas de 21 Km de perímetro y a cada cierta cantidad de kilómetros, un laboratorio de investigación. Para poder compartir archivos entre los laboratorios, la solución más simple era enviar la copia física por valija. Evidentemente, este no era un método muy eficiente, sobre todo si la gran mayoría de los documentos estaban en formato digital e iban en aumento.

Consiente de esa realidad, Bernes-Lee propuso en el año 1989 la creación de una *gran base de datos de hipertextos con enlaces* [57] o hipervínculos, a través de los cuales se puede acceder a otros objetos, en el mismo computador, o *"navegar"* hacia otra pagina en otro dispositivo distante geográficamente. Estos enlaces o links pueden ser palabras o imágenes, y permiten redirigir una página web a otra sin necesidad de tener que necesariamente recordar la dirección (web address). Los links por lo tanto, crean una telaraña o web de conexiones por la cual los usuarios pueden navegar y buscar la información en los diferentes servidores de la Internet.

El siguiente paso en la evolución de la Web, fue la incorporación de nuevos servicios en TCP/IP, por ejemplo el conocido correo electrónico o e-mail, el cual en su versión pre-web no permitía, en forma simple, el envío de documentos adjuntos. El crecimiento de la Web no ha parado y su evolución ha traído y traerá cambios que aun no se pueden dimensionar.

5.3. 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. 5.4 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 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 [8].

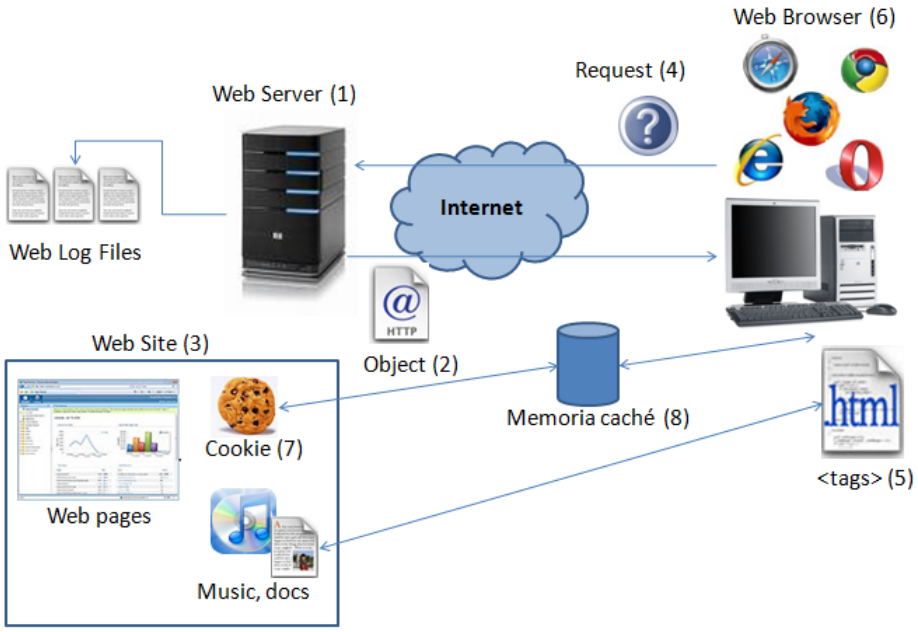

Figura 5.4: Modelo básico de operación de la Web

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.

La interacción anterior, ha quedado registrada en archivos conocidos como web log files [8], 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. 5.5 muestra un ejemplo del contenido y estructura 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 (Mozilla Firefox, Explorer, Opera, etc.).

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 5.5: Estructura de un web log file

La estructura estándar del archivo web log queda definida entonces por los siguientes puntos [57]:

1. **IP**: Es la dirección de Host de Internet, es decir, el identificador del computador desde donde el usuario está accediendo a Internet.

2. **ID**: Es la información de Identidad facilitada por el usuario.

3. **Access**: También llamado Authuser, se utiliza cuando se activa el protocolo SSL (Security Socket Layer). El usuario puede utilizar este campo para recibir o enviar información confidencial.

4. **Time**: Indica la fecha (DD/MM/AAAA) y la hora (HH:MM:SS) en que se un objeto web fue solicitado por el navegador al servidor web.

5. **Request**: Representa el objeto solicitado por el navegador, especificando el método (GET, POST), la URL o dirección del sitio web, y el protocolo utilizado.

6. **Status**: Es un número entero que indica el estado de la petición solicitada al servidor, por ejemplo un status típico es el mensaje de error 404 o Not Found, el cual indica que el cliente fue capaz de comunicarse con el servidor, sin embargo no pudo encontrar la información que se solicitó.

7. **Bytes**: Es el número de Bytes o tamaño que retorna la petición realizada por el usuario.

8. **Referer**: Es un texto enviado por el computador del cliente y que indica la fuente original de una solicitud o enlace.

9. **Agent**: En este punto se indica tanto el nombre y versión del sistema operativo, software y navegador utilizados.

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, 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.

El contenido del sitio web es de vital importancia para la continuidad de la empresa en el mundo digital. Por lo tanto, del usando los objetos correctos con la información necesaria para su descripción, se asegura la atención de los usuarios por un sitio particular. Este es el mecanismo por el cual un sitio se hace más popular en la Web y debe estar en constante actualización, pues las necesidades de los usuarios cambian.

De todos los tipos de objetos presentes en una página web, especial atención reciben los textos libres, ya que son estos los que preferentemente son almacenados e indexados por los motores de búsqueda como Google, Yahoo! o Bing, a los que la mayoría de los usuarios accede para realizar una búsqueda de información.

A partir del análisis de los datos de navegación y preferencias que se generar debido a la visita de los usuarios a un sitio web, es posible establecer lineas para mejorar su estructura y contenido, tendientes a definir qué información es interesante

para el usuario, cuál es la estructura de hipervínculos correcta para el sitio, etc. [57].

La estructura de hipervínculos de la Web, va generando grupos de páginas que comparten información común. A estos conglomerados, se les conoce como *"comunidades en la Web"* [21] y su estudio permite realizar mejor las búsquedas de información y establecer cuáles son los contenidos que son más atractivos para los usuarios. Lo anterior de inmediato categoriza las páginas web según su contenido como aquellas que poseen información importante para la comunidad (Auths), lo cual queda de manifiesto por que son apuntadas por otras páginas, y aquellas que concentran hipervínculos hacia otras páginas (Hubs). Esta categorización se aprecia en la Fig. 5.6, según su estructura de links [57].

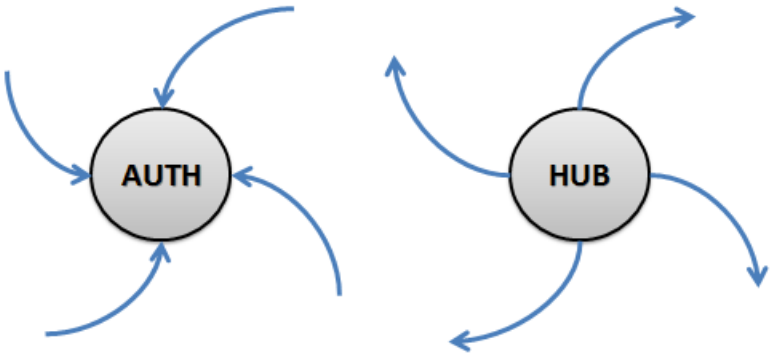

Figura 5.6: Sitios Web según estructura de links

5.4. 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 [46]:.

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), como se aprecia en la Fig. 5.7. En esencia el router recibe un paquete de datos y analiza la IP de destino, para luego decidir a qué router debe encaminarlo, para lo cual cuenta con tablas de ruta que son un verdadero mapa camionero en el cyberespacio. Al router se le pueden activar reglas de filtros de paquetes, ya sea por dirección IP o por servicio al cual se desea acceder. También es factible solicitar un monitoreo total de los paquetes que por el son encaminados, es decir, almacenar toda la información de navegación de los usuarios.

- 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. La idea fundamental es que si un usuario desea visitar una página que ya se encuentra en el proxy, no sea necesario ir a buscarla a la Web, sino que se reutiliza la existente. Lo anterior permite una mejora importante en el uso del ancho de banda desde la red local.

  Por construcción el Proxy almacena toda la relación de navegación en la Web de los usuarios de una red local. Entonces, se trata de un cuello de botella natural al cual incluso se le puede solicitar que cierta páginas no sean accesibles.

- Firewall. El el dispositivo de seguridad por excelencia de una red local. Se trata

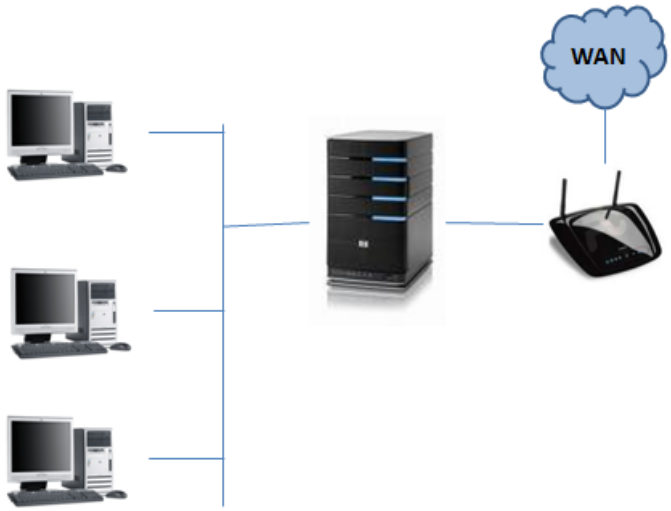

Figura 5.7: Conexión de red LAN con WAN a través de router

de un conjunto de aplicaciones de seguridad que se ejecutan preferentemente en un computador dedicado a estos menesteres, que por arquitectura aísla a la red local de la Internet. Al igual que el proxy, es también un cuello de botella natural, pero esta vez mantiene toda la relación entre los usuarios y la Internet, es decir, no sólo puede monitorear la navegación en la Web, sino que además cualquier comunicación que se produzca desde y hacia la red local.

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 re-

corren la Web de forma automática y sistemática, almacenando información de los sitios web mediante el uso de sus hipervínculos. Son utilizados frecuentemente por los buscadores web como Google o Yahoo! para recopilar y mantener actualizado el registro de páginas web e indexarlas en sus servidores, de manera que las búsquedas sean más rápidas para el usuario.

- 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.

  La utilización de este tipo de métodos de recolección de datos originados en la Web, son muy aborrecidos por los usuarios, quienes instalan a su vez otras aplicaciones destinadas a limpiar el computador de estos molestos espías.

Los datos que se originen en la Web, son muy valiosos para realizar análisis del comportamiento de usuario. Estos registros sirven, por ejemplo, para identificar las características de un sitio o las páginas que atraen al usuario, y por tanto son de gran ayuda para los negocios, en particular los que se desarrollan vía electrónica (E-commerce o E-business). La Web se ha transformado en un canal de comunicación, una manera de mostrar los productos o servicios que se ofrecen a los clientes, con cambios constantes, por lo que se necesita una adaptación rápida de manera de no sólo atraer a nuevos clientes, sino que también retener a los existentes y anticiparse ante posibles fugas que pueden ser previstas [57].

El gran desafío que enfrentará toda empresa que quiera sobrevivir en el mercado digital, es cómo se hará cargo de las necesidades de información de sus usuarios, acerca de los productos y servicios que ofrecen de cara a la explosión en conectividad que vendrá cuando el acceso universal a las telecomunicaciones sea una realidad en todo el orbe.

El acceso universal a las telecomunicaciones es un viejo anhelo que se definió en los países desarrollados en un principio como *"Servicio Universal"* y que sostenía la idea de llegar a cada persona con una línea telefónica. Claramente, aun no se logra concretar este objetivo, por cuanto en muchos países aun no se resuelven problemas más urgentes, como falta de agua y alimentos, mucho menos los relacionados con el acceso a las ventajas de las Tecnologías de la Información y las comunicaciones (TIC).

Dentro de los objetivos que ha planteado la ONU para el siglo XXI, se hace alusión explícita a la necesidad de que todo ser humano aproveche las ventajas de las TIC, reconociéndose esta meta como un derecho humano, sin el cual las personas quedarán cada vez más rezagadas en su desarrollo.

Dado el avance en TICs, ya no es posible pensar que el acceso universal sea exclusivo del ámbito de las telecomunicaciones, sino más bien de cualquier medio que permita lograr una comunicación a distancia. En ese sentido, Internet y la Web se sitúan en un escaño trascendental para el desarrollo de la humanidad y en virtud de la convergencia de los servicios en telecomunicaciones hacia redes TCP/IP, no se puede concebir una idea de acceso universal que no considere brindar a todas las personas el derecho a acceder a la gran red.

El panorama futuro es un incremento exponencial en la conectividad que se logrará producto de la expansión del acceso universal a las TICs, con lo que cualquier institución que tenga presencia en la Web se verá sobrepasada por la cantidad de visitas y búsqueda de información, lo que pondrá suma urgencia a la investigación en el procesamiento masivo de los web data que se originen para mejorar los contenidos y facilitar la navegación de los usuarios en la Web.

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.

5.5. Posibilidades de acceso a datos personales en la Web

Por construcción del protocolo TCP/IP, el acceso a Internet, independiente del medio utilizado, requiere que los datos a transmitir sean encapsulados y rotulados con una dirección IP de origen y otra de destino. Suponiendo que el computador que originó la transmisión cuenta con sólo una interfaz de comunicaciones (tarjeta de red), entonces se tendrá un mecanismo simple de identificación aproximado del usuario que comenzó con el envío de los datos. Razonamiento similar se aplica con el computador y usuario receptor.

Los problemas de identificación comienzan cuando los computadores que intervienen en la comunicación, cuentan con más de una interfaz, lo que implica que la IP no es única. También se puede dar el caso de que se produzca un *"ocultamiento"* del

computador, ya sea por motivos de seguridad (mostrar menos computadores a la red), para lo cual se utiliza un firewall, o de uso eficiente de la direcciones IP, típicamente un software NAT[4] asociado a un router, el cual permite la conexión de varios computadores en la red interna, asignándoles IP privadas (sólo operan en la red interna), las cuales son traducidas a la IP pública del router una vez que se produce una petición hacia Internet.

Analicemos el caso sencillo de una conexión ADSL casera, lo más probable es que se cuente con un router el cual se conecta con el módem que accede a la red del ISP[5]. Esta conexión requiere de la autenticación del contratante del servicio, para lo cual se le asigna una cuenta de usuario y clave secreta. Entonces, todo el tráfico que este usuario genere, puede ser registrado por el ISP y analizar directamente el comportamiento en la red de la persona totalmente identificada.

El esquema anterior se rompe cuando a través del mismo router, varios usuarios acceden a Internet. Por lo explicado anteriormente, el router considera un servicio NAT para permitir las conexiones simultáneas, con lo cual varios usuarios acceden a Internet, usando la misma IP entregada por el ISP. En este caso, ya no podría asegurarse que el contratante generó un determinado tráfico. Tan sólo se puede asegurar que desde un lugar físico, un dispositivo ADSL produció un tráfico durante un período.

Los sistemas de ocultamiendo de dispositivos expuestos, permiten monitorear todo el tráfico que encaminan. De esta forma, es posible saber exactamente qué contenidos fueron visitados desde un dispositivo conectado a la red interna. Si este dispositivo es de la exclusividad de uso de una persona, como sucede generalmente, es posible hacer un análisis profundo de su comportamiento de navegación en la red. En este sentido, todos estos sistemas actúan como un *"gran hermano"*, omnipresente, siempre alerta y revisándolo todo.

En estricto rigor, los datos originados en la Web corresponden solamente a los

[4]Network Address Translation o Traducción de Dirección de Red
[5]Internet Service Provider o Proveedor del Servicio Internet

de estructura, contenido y registro de visitas (web log) en un sitio web. Sin embargo, algunos autores también consideran en esta clasificación a aquellos datos relacionados con el perfil del usuario, es decir, nombre, correo electrónico, sexo, edad, lugar donde vive, etc. Estos datos son consignados por la mayoría de los sistemas de comercio electrónico, por cuanto son necesarios para la realización de una transacción comercial.

Salvo algunos casos excepcionales, donde alguien publica expresamente información sobre una persona, ya sea a través de una página, blog, archivo, etc., no es posible acceder a datos personales sobre un individuo determinado en forma directa. Se requiere, entonces, de una etapa de procesamiento y cruce de datos para lograr la identificación fidedigna o parcial un usuario web.

En la actualidad, la tendencia es que la publicación de datos personales en la Web, sea realizada de común acuerdo con el titular de los datos, por cuando la posibilidad de que cualquier persona pueda visualizarlos en muy corto tiempo, es altísima. En efecto, herramientas poderosas de recuperación de información tales como los motores de búsqueda, pueden en cosa de segundos mostrar quien consulte, datos referentes a una determinada persona, siendo sólo necesario saber su nombre y armarse de paciencia en el caso de que existieran coincidencias con otros usuarios de la Web.

Asumiendo que los datos personales pertenecientes a bases de datos privadas, que posee los resguardos de seguridad necesarios y suficientes, no pueden ser accedidos desde Internet, entonces, habría que analizar qué tan factible resultaría capturar o extrapolar los datos personales de los usuarios, utilizando herramientas tecnológicas.

La evidencia científica muestra que no existe certeza de hasta que punto la tecnología puede lograr la identificación de un patrón, basándose en unos cuantos datos. Por citar un ejemplo, las actuales bases de datos de ADN que se crean para identificar en forma rápida y expedita a quienes cometen determinado tipo de delitos [13, 51]. En ese sentido, se supone que la información almacenada es sólo para identificación, no así para establecer un perfil de enfermedades potenciales o existentes. Sin embargo, ¿cómo se puede tener certeza de que ese trozo de ADN almacenado, junto con técnicas

avanzadas de minería de datos no pueden extrapolar justamente aquello que se desea proteger?. Adicionalmente, sabiendo que el ADN entre personas de una misma familia es muy parecido, ¿qué sucede con los familiares no delincuentes de los cuales se puede saber información sensible vía extrapolación de una muestra que no les pertenece?

De un punto de vista científico, el desarrollo de herramientas de web mining ha estado siempre marcado por un afán de curiosidad propia de la investigación, siempre con una idea altruista de mejorar los contenidos y estructura de los sitios web, para que los usuarios logren principalmente sus objetivos de búsqueda de información en el menor tiempo posible. De esta forma, el desarrollo actual y futuro de servicios basados en la Web, usará toda la información y conocimiento que se obtenga a partir del comportamiento de navegación y preferencias de los usuarios para personalizar su experiencia en el sitio que visitan. Situaciones tales cómo la identificación aproximada o total de una persona son y serán cada vez más habituales. Luego la presentación de contenidos personalizados será un estándar de la industria, lo cual plantea un interrogante, ¿hasta que punto un sistema informático puede decidir lo que debe ver o no un usuario?. Con estas herramientas, ¿se restringe o facilita la búsqueda de información?. Si no quiero que mis datos personales aparezcan en la Web, ¿cuál es la verdadera posibilidad de que esto no ocurra?

Esta claro que el pensamiento científico sobre las herramientas de web mining, difiere muchísimo de lo que las empresas con fines de lucro ostentan como objetivo en su uso. Entonces, la idea de "*minar la web para servir a los usuarios*", cambia diametralmente a "*minar la Web para servirse de los usuarios*".

Capítulo 6

Minería de datos de la Web

El análisis de la Web es más que una secuencia de clics,
es más que una conversión de tasas, es más que sólo números.
A. Kaushik

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 [32, 57], 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 procesa-

miento de datos, cuya operación es al menos cuestionable, desde el punto de vista de la privacidad de los usuarios de un sitio web determinado [48, 49, 51]. 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? [53].

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 [3]. 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 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.

6.1. Limpieza y preprocesamiento de los web data

Los datos originados en la Web o web data, corresponden esencialmente a tres fuentes [8]:

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.

Algunos autores argumentan que también se debieran considerar los datos de perfil de usuario como parte de los web data [8, 19]. 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 [19, 55].

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 [8]:

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 [8, 19, 57]:

1. **Estrategia Proactiva**: Consiste en utilizar herramientas invasivas para identificar a los usuarios, generalmente se trata de cookies o spybots, 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 [51, 62]. 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 [3]. 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 [22, 42, 57]. 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 [19].

6.2. 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. En este sentido, se podría decir que es la aplicación de teoría del data mining al caso particular que revisten los web data [30].

Web Mining ha permitido estudiar la creciente cantidad de datos disponibles, donde la estadística clásica y la revisión manual ya resultan ineficientes [4]. 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 [57]:

1. Patrones de Secuencia. Utilizando estadística, más un análisis de la secuencia de páginas visitadas por los usuarios, se intenta extraer cuáles son los patrones de navegación presentes en un sitio, es decir, rutas probables, secuencia de páginas más visitadas, probabilidad de visitas, etc.

2. Reglas de asociación. La idea central es encontrar correlaciones entre conjuntos de datos bajo una probabilidad de ocurrencia (confianza) para una proporción

del total de datos (soporte). Por ejemplo pan y queso (soporte = 5 %, confianza = 42 %) quiere decir que 42 % de las personas que compraron pan también compraron queso y que esta combinación se dió en 5 % de las transacciones. Una extensión a ésta es la asociación multidimensional, dónde se asocia más de un atributo a otro (pan, mantequilla y queso).

3. 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 durante el proceso.

   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.

4. 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 clúster 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 clúster 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 clúster, 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 clúster 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.

Ya se han nombrado las técnicas más utilizadas de Data Mining con las cuales se puede extraer información de los datos generados en la Web. Para aplicar estas técnicas se cuenta con diversas herramientas, que dependiendo de la situación y condiciones o restricciones convendrá utilizar una u otra. A continuación se muestran las herramientas de Web Mining más utilizadas [57] :

1. **Redes Neuronales Artificiales (Artificial Neural Networks (ANN))**: Son un conjunto de elementos relacionados en un proceso, cuyo funcionamiento se basa en las interacciones que llevan a cabo las neuronas del cerebro humano. La idea principal es que existen elementos o unidades (nodos o neuronas) que almacenan la información a través del aprendizaje directo, que puede ser supervisado o no supervisado, para lo cual se necesita un conjunto de entrenamiento calificado que permita ir actualizando la información con la que cuentan las neuronas, y con esto permitir generalizar los resultados posibles a partir de la experiencia.

   Esta herramienta no conviene utilizarla en el caso que existan muchas capas

ocultas en la red, muchas neuronas o instancias de entrenamiento, ya que debido a esto último puede producirse sobre aprendizaje (overfitting), perdiendo la habilidad de generalización.

2. **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.

3. **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.

4. **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.

5. **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.

6. **Algoritmos Inspirados en la Vida**. Se trata de una nueva familia de algorit-

mos cuya operación está basada en cómo ciertas especies, bacterias y la misma evolución con cambios genéticos, tratan de sobrevivir y perpetuarse en la vida. Entre estos algoritmos se encuentran los basados en el comportamiento de las hormigas[1], abejas, atunes, bacterias y genéticos. En el ámbito de la Web, lo que se trata de modelar es el comportamiento de los usuarios, utilizando los parámetros de los algoritmos antes mencionados.

6.3. 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 [55]. 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.

[1]Ant Colony Optimization

6.3.1. 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. Al respecto, la Tabla 6.1 muestra un resumen de las técnicas más utilizadas para sesionizar registros de log.

Método	**Descripción**	**Grado de Invasión de Privacidad**	**Ventajas**	**Desventajas**
Dirección IP + Agente	Se asume que a cada dirección IP le corresponde un único Agente, y que esto representa a un usuario.	Bajo	Siempre disponible, no es necesario tecnología adicional.	No garantiza que el usuario es único, falla en el caso de IP dinámicas.
Combinar sesiones según ID	Utiliza páginas generadas dinámicamente para asociar el ID con cada hipervínculo.	Bajo - Medio	Siempre disponible, independiente de la Dirección IP.	No puede reconocer usuarios repetidos. Sobrecarga adicional con páginas dinámicas.
Registro	Utilizar registro de usuarios en los sitios web.	Medio	Se puede hacer seguimiento a un usuario, no sólo a un navegador.	Muchos usuarios no se registran, y previo al registro no se obtienen datos.
Cookie	Grabar el ID en el computador del cliente.	Medio-Alto	Se puede hacer seguimiento de visitas reiterativas desde un mismo navegador.	El usuario puede decidir no hacer uso de éstas o eliminarlas.
Agentes Software	Programas que se cargan en el navegador y que envía datos del usuario constantemente.	Alto	Se utilizan datos precisos para un único sitio web.	Hay posibilidades de que sean rechazados por los usuarios.

Tabla 6.1: Mecanismos para identificación de sesiones

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 [8, 19]. En tal sentido, las técnicas más comúnmente aceptadas son las dos primeras de la Tabla 6.1. 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 [33], 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 [42, 58].

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 afines. Desde un punto netamente informático, este cómo se transforma en reglas $\mathbf{if - then - else}$, que junto con los patrones, configuran el conocimiento extraído desde los web data [56].

6.3.2. 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 [54].

Con los supuestos anteriores, es posible reducir una cantidad considerable de textos en la Web a sus componentes más fundamentales. Luego es posible iniciar acciones que permitan la extracción y conocimiento desde todo tipo de documento. En ese sentido, se han hecho trabajos destinados a la detección de las preferencias de las personas, a partir de los textos que han quedado consignados respecto de un tema dado, por ejemplo a través de un foro o en un blog. A este tipo de procesamiento de datos se le denomina Web Opinion Mining [23, 58], y está atrayendo la atención de muchas instituciones, que ven los blogs, foros y temas relacionados, una verdadera mina de oro para analizar al detalle que le agrada o desagrada a sus clientes fijos y potenciales.

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.

6.3.3. 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? [10].

- 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, ya ha sido registrada en la jurisprudencia nacional[2]. Sin embargo, hay que dejar constancia de que el análisis de los contenidos que se publican para detectar situaciones contrarias a la ley, no es algo trivial, siendo en la mayoría de los casos los propios afectados quienes alertan a los administradores del sitio del potencial ilícito que se está cometiendo [9].

6.3.4. 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 [57]. 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.

Este nuevo santo grial del ciberespacio, no es algo trivial de implementar y se puede lograr a partir del uso del conocimiento extraído de los web data, el cual permite aplicar técnicas de clasificación de los usuarios por similaridades con los grupos identificados [55, 56]. En síntesis, cuando un usuario visita el sitio, inmediatamente se procede a analizar su navegación para luego identificar a que grupo de usuarios

[2]Ver fallo del 6 de diciembre de 1999, de la Ilustrísima Corte de Apelaciones de Concepción, rol Nº 243-99

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 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 [48, 49], 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 [8]. 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?) [29, 35, 59], con lo cual se estaría vulnerando abiertamente la privacidad del visitante [3].

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 [34, 15].

6.4. Privacidad y libertad en la navegación desde la perspectiva del web mining

El desarrollo científico y tecnológico entorno al web mining, desde sus inicios ha estado marcado por un afán de aprendizaje respecto de entender cuáles son las motivaciones, preferencias y necesidades de información que poseen los usuarios cuando visitan un sitio web. Todo este nuevo conocimiento es usado fundamentalmente para mejorar la experiencia del usuario en el sitio. Ahora bien, es claro que detrás de esta idea altruista, es posible obtener segundas derivadas que se orienten a producir beneficios pecuniarios a los dueños de un sitio, sobre todo si se trata de una plataforma de comercio electrónico.

Para entender algunas de las motivaciones de mineros de datos de la Web, Wel y Royakkers [62] realizaron una encuesta a más de 100 profesionales y científicos que investigan, desarrollan o usan técnicas de web mining y las respuestas fueron:

1. El web mining no atenta contra la privacidad de los usuarios.

2. Existen leyes para proteger la vida privada de las personas y también mecanismos para garantizar la privacidad de los usuarios en la Web, sólo basta con leer la política de uso del sitio que se visita.

3. Una gran mayoría de los usuarios en la Web ponen a disposición de quien quiera verlo sus propios datos personales, sin que esto les moleste.

4. La mayoría de los web data no hacen referencia directa a datos personales y son usados para la creación de perfiles de usuarios anónimos.

5. El web mining ayuda a disminuir campañas de marketing no solicitadas al lograr un mayor entendimiento de las preferencias de los usuarios de los sitios web.

6. La personalización de la web apunta a brindar un servicio individualizado de los usuarios de un sitio.

Es interesante notar que ninguno de los encuestados ve en el desarrollo de su actividad problemas que limiten con la ética, el uso abusivo de la tecnología, el atentar contra la privacidad y la libertad de navegación de las personas. Lo anterior muestra un vacío que debe ser llenado con educación y difusión sobre las implicancias negativas del procesamiento de web data, afín de que se definan un conjunto de buenas prácticas que preserven las garantías fundamentales que pueden estar siendo afectadas.

Cada una de las respuestas entregadas, resultan ser muy interesantes para analizar hasta dónde el web mining puede atentar contra la privacidad y la libertad de navegación de los usuarios, ya que se está recurriendo a la fuente misma que origina el fenómeno, es decir, los mineros de la Web y sus clientes directos. A continuación, un breve análisis y reflexión de cada una de las respuestas:

Respuesta 1. Desde mucho antes de la invención del web mining, se han utilizado herramientas para la creación de perfiles de clientes, con el objetivo de lograr una mayor comprensión de sus hábitos de consumo. Tradicionalmente, estas herramientas se han orientado a la creación de perfiles basados en grupos de clientes con características similares, utilizando cantidades pequeñas de datos, pues no existía la capacidad de procesamiento para grandes volúmenes.

Con la creación de las herramientas de data mining, las que posteriormente han sido adaptadas para los web data y el avance en hardware para el procesamiento masivo de datos, la creación de perfiles de clientes ha llegado a llegado a un refinamiento tal que ya se puede hablar de que se puede individualizar el comportamiento de una persona. De hecho los postulados del marketing uno a uno, no hacen más que exacerbar el uso del data mining para llegar a saber exactamente que desea una persona, con nombre y apellido.

La posibilidad que brinda la Web de poder acceder a datos que separados no dicen nada, pero que reunidos permiten con un esfuerzo razonable saber preferencias de contenido y navegación de una persona determinada, hacen que el web mining se transforme en una herramienta que a todas luces puede vulnerar la privacidad de los usuarios.

Respuesta 2. Si hay algo que un usuario web no hace, es justamente leer las políticas de uso de los sitios. Es más, está demostrado empíricamente que los usuarios realizan una vista rápida de las páginas cuando visitan un sitio web [36]. Si a lo anterior se le agrega que toda política de uso viene en letra pequeña y contenida sumergida en un mar de otros contenidos del sitio visitado, se puede argumentar que con esta línea de pensamiento, el usuario está totalmente indefenso mientras navega en la Web.

Adicionalmente, el argumento de la existencia de leyes que salvaguardan la privacidad del usuario en la Web en su concepción está equivocado. En efecto, las leyes, por el principio de neutralidad tecnológica, no están enfocadas a regular una tecnología, que se sabe va a variar rápidamente.

Respuesta 3. Si los usuarios que ponen a disposición de todo el mundo sus datos personales supieran hasta qué punto las empresas pueden llegar a penetrar en su intimidad, de seguro que lo pensarían varias veces antes de incurrir en la mencionada práctica. En efecto, existe mucho desconocimiento respecto de lo vulnerable que es la privacidad de las personas cuando se ven expuestas a medios de recuperación de información portentosos como los que se encuentran en la

Web. Antiguamente, si alguien ventilaba su vida privada, lo usual era que esta información quedara en algún medio escrito o programa televisivo que luego de la emisión era difícil de volver a consultar. Pero ahora, es necesaria una consulta a un buscador para encontrar información relativa a una persona que es utilizada para formarse juicios respecto de sus hábitos, fortalezas y debilidades. Incluso más, se ha acuñado el término *"Googlear"* dentro de la jerga de los cibernautas, para la acción de buscar información acerca de una persona usando Google.

Respuesta 4. Si bien es cierto que muchos de los web data por si solos no hacen referencia directa a las personas, su combinación con otros datos o el simple hecho de crear perfiles de usuarios basados en web data anónimos puede usarse para encasillar el comportamiento de un usuario en particular e indagar respecto de información personal [6].

El supuesto de que a partir de datos anónimos no se puede vulnerar la privacidad de los usuarios es errado. Con las combinaciones correctas, la agregación de información adecuada y el proceso de web mining necesario, es posible llegar a un nivel de detalle tal que se pueden vulnerar los derechos fundamentales de los usuarios en la Web [41, 43].

Respuesta 5. Las siempre molestas campañas de marketing masivas, adolecen del problema de la falta de focalización, es decir, ante la imposibilidad de precisar las reales necesidades de los potenciales clientes, optan por el envío a gran escala de catálogos, avisos, etc., pues está demostrado empíricamente que existe un alto porcentaje de personas a las cuales se les puede motivar a adquirir un determinado producto o servicio. En este sentido, el web mining puede contribuir eficazmente a mejorar la forma en que se dirige la publicidad.

Numerosos son las empresas que ya cuentan con sistemas que ofrecen sus productos y servicios basados en las preferencias de sus clientes y extrapolando el comportamiento de aquellos que aun no lo son. Si bien es cierto que estas ofertas en muchos casos ayudan a los clientes y visitantes tomar una decisión de compra, también existe el efecto perverso de la manipulación bajo presión. En

efecto, considérese una fecha importante como puede ser el día de la madre. En los días y horas previas, muchas personas se harán la misma pregunta *"¿qué le regalo a mi madre?"*, interrogante que es muy bien utilizada por los sitios de venta en línea de las grandes tiendas comerciales para ofrecer el regalo *"bueno, bonito y barato"*, el cual será despachado en el menor tiempo posible y que luego no satisface ni a la madre ni al hijo(a) que hizo la compra.

Respuesta 6. La personalización no se percibe como algo malo o fuera de la ley, sino como algo muy beneficioso para los usuarios, que siempre se pierden en sus búsquedas de información en los sitios web. Sin embargo, no se ha pensado en las restricciones a la libertad de navegación que puede tener un usuario. Es más, ¿qué sucede si el sistema que entrega las recomendaciones para personalizar la experiencia del usuario en el sitio entrega resultados erróneos?. Simplemente el usuario no tendrá opción de ver y juzgar por si mismo si los contenidos del sitio le son o no de interés.

La experiencia práctica ha demostrado que todo sistema basado en puntos para dar recomendaciones, tiene su talón de Aquiles en el hecho de que se están comparando situaciones particulares con hechos generales. Por ejemplo, un sistema de scoring asigna puntos a las personas en base a datos recolectados por hechos pasados correspondientes a otras personas, aplicando un principio básico de comportamiento por similaridad, es decir, *"si el individuo en análisis se comporta como el grupo B, entonces la recomendación sería aplicar la regla que corresponde a ese grupo"*. De inmediato aparecen otros efectos, siendo uno de los más nocivos y odiados la discriminación sin un fundamento profundo, salvo lo que arroja el procesamiento de los datos de otras personas para extrapolar el comportamiento de un individuo en particular.

Aunque suene un tanto retórico y casi palabras de buena crianza, los científicos y profesionales relacionados con web mining, visualizan en el uso de estas herramientas una forma altruísta de ayudar a los usuarios en su relación con la Web, por lo tanto es

difícil que puedan ver los efectos perversos que el uso de estas técnicas pueden llegar a tener en la sociedad de la información.

Es necesario sensibilizar a estos grupos de investigadores en torno a la vulneración de la privacidad, la restricción a la libertad de navegación y otros problemas colaterales que pueden causarse por un mal uso o abuso de las técnicas, métodos y algoritmos comprendidos dentro del web mining. Lo anterior debe ser hecho siempre desde la óptica de no frenar el avance científico en torno a la Web, por cuanto su desarrollo es claro ha beneficiado a millones y ha sido en gran parte por que no ha existido una regulación que restrinja el estudio de este fenómeno.

Es de suma importancia que se creen puentes interdisciplinarios donde los investigadores compartan con abogados, legisladores y por que no, público en general para que se puedan obtener los mayores beneficios de la Web, salvaguardando los derechos fundamentales de las personas.

Capítulo 7

Aspectos jurídicos del tratamiento de web data

Cum finis est licitus, etiam media sunt licita.
Hermann Busenbaum

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, tanto en lo que se refiere a la privacidad del usuario como a la legalidad, desde el punto de vista de la criminalidad informática, de los procedimientos involucrados en la obtención de los datos a minar.

El resultado práctico de un proceso de web mining, por lo general se relaciona con la mejora de la estructura y contenido del sitio de donde se extrajeron los web data, lo cual se logra principalmente a través de recomendaciones dirigidas hacia el administrador del sitio y a los usuarios.

Asumiendo que el administrador o web master es un experto en estas artes, las recomendaciones bien pueden hacerles sentido o no, en cuyo caso las desecha. Pero ¿qué sucede con el usuario común? La recomendación que recibe por lo general es en

línea y dice relación con cuales contenidos debe ver y cuales no, la navegación más óptima para que encuentre información y en el fondo, toda sugerencia que su perfil ya analizado permita extrapolar en comparación con otros usuarios que visitaron el sitio.

7.1. Marco legal para el análisis de la vida privada en el Web Mining

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 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."* Esta definición es plenamente coincidente con la adoptada en Chile a través de la ley 19.628 como se verá más adelante.

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. 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 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 [30].

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, por que ve 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 esta tesis 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 conciernen. 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 [58].

Si se analizan los web data utilizados para la extracción de patrones de navegación y preferencias de lo usuarios, estos se pueden agrupar en [27]:

- 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. Estas reglas se derivan directamente de la Directiva 95/46/CE 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* [27, 57, 56].

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 se entra 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 formular reglas de fidelización, promociones, etc.

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, en la medida que sea imposible vincular los web data a una persona determinada o determinable de acuerdo al estado de la ciencia. En consecuencia, la cualificación del web data como dato personal dependerá de la robuztez y capacidad de los sistemas informáticos, siendo imposible sostener a priori que por el hecho de ser un web data no se está ante 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 [27]:

- 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 explicito 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* [14]. 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.

En base a lo expuesto por la ley 19.628, los datos de carácter personal pueden ser clasificados como:

- Dato estadístico. Dato que "*en su origen, o como consecuencia de su tratamiento, no puede ser asociado a un titular identificado o identificable*".
- Datos de carácter personal. "*Cualquier información concerniente a personas naturales, identificadas o identificables*".
- Dato sensible. "*Aquellos datos personales que se refieren a las características físicas o morales de las personas o a hechos o circunstancias de su vida privada o intimidad, tales como los hábitos personales, el origen racial, las ideologías y opiniones políticas, las creencias o convicciones religiosas, los estados de salud físicos o psíquicos y la vida sexual*"

A primera vista, se podría pensar que los web data no contemplan datos sensibles. Sin embargo, con el advenimiento de las aplicaciones Web 2.0, las cuales permiten que los mismos usuarios sean quienes consignan datos de su interés, es posible que ciertos datos sensibles queden expuestos a través de los blogs, wikies, foros y páginas personales. En tal sentido, se pueden aplicar técnicas de web mining para saber o extrapolar la religión, tendencia política o incluso la vida sexual de una persona, si es que esta o un tercero ha consignado datos en la Web que permitan llegar a esas conclusiones. Por ejemplo, es bien sabido que en aplicaciones como Facebook, es factible que el usuario declare abiertamente su tendencia religiosa, gustos, intereses personales, fotos, etc. Por una parte es cierto que el usuario es dueño de hacer lo que le plazca con sus datos, pero por otro lado, ¿estará consciente de que si no aplica al menos reglas básicas de privacidad de sus datos, queda totalmente expuesto a la Web?

7.2. 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 [26, 57]. 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 [56, 58].

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, entrometerse 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 [27]. 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? [19, 25, 27], ¿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 sí 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 [39].

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.

7.2.1. Comentarios finales

Por su naturaleza, los web data abarcan todos los tipos de datos de la historia de la computación, lo que hace posible que dentro de su definición más amplia, se acojan los denominados datos personales y los datos sensibles, según la definición entregada por la ley 19.628. Este hecho, hace que el tratamiento de los web data deba ser realizado conforme a la regulación vigente.

En la mencionada ley, 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. En este caso, se produciría la anonimización de los web data, por lo que ontológicamente no podrían llegar a ser datos personales, quedando por tanto fuera de la aplicación de la ley 19.628.

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 necesaria 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.

Como en otros ámbitos de la vida, la vulneración de la privacidad y afectación de la libertad sólo debe producirse en casos donde se requiera una intervención por parte de un organismo del Estado para la persecución de un delito. Por ejemplo, la Ley 20.000 que sanciona el tráfico ilícito de estupefacientes y sustancias sicotrópicas, en su artículo 25 consigna que *"el Ministerio Público podrá autorizar a funcionarios policiales para que se desempeñen como agentes encubiertos o agentes reveladores y, a propuesta de dichos funcionarios, para que determinados informantes de esos Servicios actúen en alguna de las dos calidades anteriores"*. En este mismo artículo se define al agente encubierto como *"funcionario policial que oculta su identidad oficial y se involucra o introduce en las organizaciones delictuales o en meras asociaciones o agrupaciones con propósitos delictivos, con el objetivo de identificar a los participantes, reunir información y recoger antecedentes necesarios para la investigación"* y al agente revelador como *"funcionario policial que simula ser comprador o adquirente, para sí o para terceros, de sustancias estupefacientes o sicotrópicas, con el propósito de lograr la manifestación o incautación de la droga"*. Claramente este artículo define una forma directa de de invasión de la privacidad de individuos y organizaciones, incluso se puede vulnerar el derecho a la intimidad y la inviolabilidad del hogar, coartando libertades esenciales, al amparo de la ley si el caso que se analiza es un delito relacionado con el narcotráfico.

Haciendo un paralelo con el análisis anterior, las técnicas de web mining permiten notables análisis de los web data para la detección de delitos, tales como la

pornografía infantil. Sin embargo, su aplicación requiere realizar ciertos procedimientos que podrían ser considerados delitos fragantes. En efecto, para descubrir una red de pedófilos en la Web, primero habrá que realizar un trabajo de recuperación y almacenamiento de los web data que puedan ser considerados pornografía infantil, con lo cual se deberá crear un repositorio con este tipo de material, vulnerando lo descrito en el artículo primero número 21 de la ley 19.927 que modifica el Código Penal, el Código de Procedimiento Penal y el Código Procesal Penal en materia de delitos de pornografía infantil [12], agregando en el Código Penal el Artículo 374 bis que en su inciso segundo deja consignado que *"el que maliciosamente adquiera o almacene material pornográfico, cualquiera sea su soporte, en cuya elaboración hayan sido utilizados menores de dieciocho años, será castigado con presidio menor en su grado medio"*.

Para poder aplicar técnicas de web mining en la situación descrita y conforme a derecho, habrá que analizar si se cumplen las condiciones descritas en el número 18 del artículo 1 de la ley 19.927, el cual agrega al Código Penal el Artículo 369 ter que consigna que *"cuando existieren sospechas fundadas de que una persona o una organización delictiva hubiere cometido o preparado la comisión de alguno de los delitos previstos en los artículos 366 quinquies, 367, 367 bis, 367 ter, 374 bis, inciso primero, y 374ter, y la investigación lo hiciere imprescindible, el tribunal, a petición del Ministerio Público, podrá autorizar la interceptación o grabación de las telecomunicaciones de esa persona o de quienes integraren dicha organización, la fotografía, filmación u otros medios de reproducción de imágenes conducentes al esclarecimiento de los hechos y la grabación de comunicaciones"*.

Siendo los web data originados ya sea por el acceso de los usuarios a Internet o por la creación de páginas web que en su transmisión requieran el uso de la gran red, se puede argumentar que al ser el acceso a Internet un servicio de telecomunicaciones, y en virtud del artículo citado anteriormente, la aplicación del web mining para la investigación de organizaciones vinculadas con la pornografía infantil se ajusta a derecho.

También en el citado artículo se consigna que *"igualmente, bajo los mismos supuestos previstos en el inciso precedente, podrá el tribunal, a petición del Ministerio Público, autorizar la intervención de agentes encubiertos"*. En el caso del web mining, estos agentes tendrían su realización práctica en software, controlado por los organismos policiales pertinentes, que simulen actividades relacionadas con la producción y almacenaje de pornografía infantil, con el fin de capturar información valiosa para la detección de las mencionadas comunidades delictivas.

El desafío ahora es establecer los mecanismos entre los organismos policiales y los científicos dedicados al web mining a fin de que sea una realidad la creación de sistemas informáticos complejos que permitan la detección de los hechos delictivos desarrollados anteriormente. En otras partes del orbe ya se han hecho trabajos muy importantes al respecto y aunque Chile ha sido siempre un colaborador importante en la captura y destrucción de este tipo de organizaciones, siempre a asumido una actitud un tanto reactiva, dependiendo de los datos que aportan las policías de países más desarrollados.

7.3. Regulación del derecho al honor, la honra y web mining

Toda publicación de información en la Web, no es del todo confiable. Lo anterior se fundamenta en el hecho de que no existe un ente regulador o certificador de que lo expuesto en una página es una verdad científica o una opinión fundamentada. Dicho de otra forma, se confía en que el publicante ha hecho un trabajo serio de análisis y validación de lo que publica, lo cual es sólo un supuesto motivado por la buena fe entre las personas.

Cuando la Web estaba en sus primeras etapas de desarrollo, sólo los administradores de los sitios web tenían la posibilidad de agregar y/o modificar contenidos. Dado que en la mayoría de los casos los sitios pertenecían a instituciones que debían

velar por que los contenidos de sus portales estuvieran de acuerdo a ciertos códigos y marcos legales, lo publicado era considerado como correcto y validado.

A medida que la Web se hace más popular y comienzan a proliferar las páginas personales, el escenario va cambiando y los contenidos ya no se rigen por algún marco, sino por lo que el usuario desea en ese momento publicar. Con el advenimiento de los servicios Web 2.0, los usuarios definitivamente adquieren un protagonismo central. Ahora son ellos quienes a través de blogs, páginas personales, wikis, foros, etc., van consignando sus propias opiniones y documentos. En este punto, la Web se vuelve poco certera en la veracidad de sus contenidos. De hecho en ciencias no se considera correcto citar los contenidos expuestos en Wikipedia.org, por su alto grado de imprecisiones debido a que son los mismos usuarios quienes han ido consignando ideas y redactado parte de sus artículos.

Sin embargo, la Web en general goza de una buena reputación al momento de obtener información a través del uso de los buscadores como Goole, Bing o Yahoo!. En efecto, se le considera una fuente casi inagotable de conocimiento, donde se puede consultar prácticamente por cualquier tema, en particular sobre referencias de una persona determinada. En este sentido, todos los contenidos que un usuario consigne en la Web quedan a disposición de quien quiera consultarlos.

Cuando se trata de información relativa a una persona, ya sea consignada por el titular de los datos o por un tercero, se plantean varias interrogantes respecto de la veracidad de los datos, del posible menoscabo de un individuo aludido y también de algún tipo de discriminación. De esta forma, la información que un usuario pone a disposición del mundo a través de la Web, puede luego ser usada en su contra, por ejemplo realizando prácticas discriminatorias. Una encuesta de la prestigiosa consultora Harris Interactive[1] aplicada el año 2009 a 2.600 directivos de empresa, mostró que en el 45 % de los empleos ofrecidos en EE.UU, se recurrió al uso de los buscadores para revisar información adicional que pudiese haber de los postulantes en la Web.

[1] http://www.harrisinteractive.com/.

Del porcentaje anterior, un 35 % de los los candidatos fueron rechazados por que se encontraron contenidos que no eran del gusto de las instituciones contratantes, fundamentalmente debido que se había información provocativa o inapropiada (fotos de alguna especie), comentarios hostiles sobre empresas en las que el candidato había trabajado anteriormente, contenidos relacionados con alcohol o drogas, etc.

Lo anterior es un claro ejemplo de discriminación [1] basada en la información que los mismos usuarios consignan en la Web, lo cual incluso puede ser muy inexacto, pues basta que otro usuario exponga juicios de valor sobre una persona determinada para que luego el contratante se forme una idea negativa del postulante a un empleo.

Sin lugar a dudas, la utilización de técnicas de web mining para la extracción de información por sobre datos que pueden ser imprecisos, solo generará resultados imprecisos, por lo que no se debe considerar como una certeza lo que se extraiga desde fuentes poco confiables. Sin embargo, puede ser usada como información complementaria para la toma de decisiones.

7.4. Web Mining y legislación sobre delitos informáticos

Haciendo presente la existencia de variadas definiciones sobre que se entiende por delito informático, entenderemos por tal, siguiendo a García Noguera, a *"todo ilícito penal llevado a cabo a través de medios informáticos y que está íntimamente ligado a los bienes jurídicos relacionados con las tecnologías de la información o que tiene como fin estos bienes"* [37]. Es frecuente a su vez encontrar, en la doctrina especializada, la distinción entre delitos informáticos propiamente tal y los delitos computacionales, primando en estos últimos la utilización de medios tecnológicos para la comisión de delitos cuya naturaleza no difiere mayormente de los consagrados tradicionalmente en el código penal [5].

El delito informático, de acuerdo a la generalidad de los autoridades, puede desa-

rrollarse a través de diversas formas comisivas, pudiéndose mencionar: el sabotaje informático, el fraude informático, la piratería computacional, el espionaje informático y el acceso no autorizado o hacking directo [31]. Del listado precedente, la apropiación de datos y el espionaje informáticos son las modalidades que tienen mayor relevancia en nuestro estudio, en vista de las diferentes conductas que mediante la aplicación de las técnicas de web mining pueden llevarse a cabo por el profesional informático.

A fin de determinar la licitud del uso del web mining en el plano nacional como internacional, precisaremos como actividad sujeta a análisis al tratamiento de datos contenidos en web logs y la utilización de otro mecanismos de identificación de sesiones (véase Tabla 6.1).

7.4.0.1. Ámbito Nacional.

Esta materia se encuentra tratada mediante la Ley 19.223 que tipifica figuras penales relativas a la informática. A través de su breve articulado se detallan diversas conductas típicas que abarcan desde la destrucción de los sistemas informáticos (tanto de los elementos físicos que lo componen como de sus elementos lógicos) hasta diversas formas de uso ilícito de la información contenida en dichos sistemas (alteración, apropiación,difusión).

Siendo la aplicación del web mining una actividad centrada en la obtención de información para su posterior tratamiento, el análisis debe centrarse en el tipo penal que castiga la posibilidad de adquisición ilícita de dicha información. En este sentido, el artículo 2 de la citada ley expresa que "*el que con el ánimo de apoderarse, usar o conocer indebidamente de la información contenida en un sistema de tratamiento de la misma, lo intercepte, interfiera o acceda a él, será castigado con presidio menor en su grado mínimo a medio.*"

De la lectura del artículo podemos apreciar que, para estar en presencia de la caracterización típica de una conducta bajo este delito, el acceso o interceptación de un sistema informático debe realizarse con la intención (dolosa) de apropiarse de

determinada información.

Siguiendo dicha lógica, cuando el profesional informático se analice los web logs generados en un servidor, autorizado por el dueño de dicho sistema, no estaría cometiendo ninguna actividad típica que merezca reproche penal. En el caso de usarse técnicas mayormente invasivas de obtención de información, como lo son el uso de web bots y cookies, el análisis deberá centrarse en determinar en el ingreso malicioso (esencialmente subrepticio y no informado) a los sistemas informáticos del usuario de una determinada pagina web, con el propósito de apropiarse de una información alojada en dicho sistema (p. ej, extracción de datos presentes en el disco duro del usuario).

La necesidad de que se manifieste una intención maliciosa de apropiarse de la información del sistema limita drásticamente las posibilidades de que una conducta habitual de minería de datos sobre los web datas pueda calificarse como típica, antijurídica y culpable.

Cabe destacar que, reconociéndose la falta del tratamiento penal de la interceptación per se de sistemas informáticos en nuestra exigua legislación sobre la materia, se han presentado dos proyectos de ley que dentro de lo relevante para este trabajo, penalizan de manera directa el acceso no autorizado a un determinado sistema, ya sea mediante la modificación del articulado actual de la ley 19.233[2], o mediante la reforma del Código Penal a fin de incluir nuevas modalidades delictivas basadas en el uso de medios tecnológicos informáticos[3].

En el escenario de aprobarse alguno de los dos proyectos, pudiesen ser objeto de revisión típica solo los mecanismos de identificación de sesión que intercepten o accedan subrepticiamente al sistema informático del usuario de forma dolosa.

[2]Boletín N°2974-19 .Modifica ley Nº 19.223, de 1993 , que tipifica figuras penales relativas a la informática. Disponible en: http://sil.congreso.cl/cgi-bin/sil_proyectos.pl?2974-19

[3]Boletín Nº 3083-07. Modifica el Código Penal con el objeto de recepcionar, en los tipos penales tradicionales, nuevas formas delictivas surgidas a partir del desarrollo de la informática. Disponible en: http://sil.congreso.cl/pags/index.html

7.4.0.2. Ámbito Internacional.

El mayor instrumento de carácter internacional que se refiere a los delitos informáticos es la Convención sobre el Cyber Crimen[4] del Consejo Europeo de 2001, abierta para firma el 2001, entrando en vigencia el 1 de Septiembre de 2004. A la fecha dicho tratado se encuentra firmado y ratificado por 29 miembros del Consejo Europeo y por Estados Unidos en calidad de parte no miembro.

El tratado define una serie de conductas ilícitas que debiesen ser objeto de regulación criminal por cada Estado parte del instrumento.

Entre estas conductas se mencionan: el acceso ilegal, la interceptación ilegal, la interferencia de sistema o denegación de servicios , el fraude informático y el mal uso de dispositivos electrónicos.

En lo referido a las posibilidades de acceso ilegal a un sistema, mediante una nota explicativa del convenio, se plantea el caso específico del uso de cookies y bots y señala que:

> "La aplicación de herramientas técnicas específicas pudiese resultar en un acceso *(no autorizado)* de los enunciados en el artículo 2, tales como el acceso a una página web, de manera directa o mediante hipervínculos, incluyendo deep-links o aplicaciones *cookies o bots* dispuestos para localizar y obtener información al servicio de la comunicación. La aplicación de este tipo de herramientas no se considerará per se como un " acceso sin autorización o derecho"(...) El uso de aplicaciones estándar dispuestas para ser utilizadas en los más comunes protocolos de comunicación y programas *no puede considerarse por si mismo un acceso sin autorización*, en particular cuando puede entenderse el consentimiento del propietario del sistema accesado en el uso de dicha aplicación . Ej, en el caso de las

[4]Convention on Cybercrime / Council of Europe. Disponible en http://conventions.coe.int/Treaty/en/Treaties/Html/185.htm

cookies al no negar su primera instalación o no removerla"[5].

Tomando en consideración lo anterior, es posible concluir que en el estado actual del desarrollo de los mecanismos de armonización de legislación criminal internacional no se considera de derecho propio al uso de tecnologías de recolección de la información como una forma delectiva de acceso ilegal, y por ende a grandes rasgos, los mecanismos de obtención de datos para el minado de datos en la web no constituyen una conducta que merezca por si sola sanción legal.

7.5. El Web mining frente a los principios generales del derecho

Por tratarse de una actividad de la ciencia que tiene corta data, el Web Mining no posee a nivel mundial una regulación específica. Sin embargo, al analizar la operación general de los algoritmos, técnicas y métodos involucrados, es posible conjeturar que al menos a nivel nacional, la actividad puede ser desarrollada al alero de las leyes de protección de datos y de telecomunicaciones, y que no es ajena a los principios generales del derecho. En este sentido rige plenamente el principio de buena fe, de protección de la dignidad humana y los derechos fundamentales, la seguridad jurídica, además de los principios del derecho informático como la neutralidad tecnológica, la equivalencia funcional y normativa, que resultan especialmente críticas en esta materia.

La creación de una regulación específica para la actividad del Web Mining a nivel científico, puede atentar gravemente a su desarrollo y a los beneficios que ha traído y seguirá trayendo para la humanidad. Sin embargo, lo anterior no exime a los los científicos del Web Mining de la responsabilidad de conocer cuales son los principios del derecho a través de los cuales pueden realizar sus investigaciones.

[5]Explanatory Report to the Convention on CyberCrime Sección Tercera párrafo 48 (traducción propia). Disponible en http://conventions.coe.int/Treaty/en/Reports/Html/185.htm

Una manifestación de estos principios en la ley 19.628 es el reconocimiento de los derechos de los titulares de datos personales, especialmente el derecho de acceso, consistente en la posibilidad de que el titular acceda en todo momento a sus datos personales que son objeto de tratamiento y en la medida que no se encuentre restringido este acceso por la Ley. Los derechos de modificación, cancelación y bloqueo de datos personales, que le permiten exigir al responsable del tratamiento que realice estas operaciones en las hipótesis previstas en la Ley.

Asimismo, es una manifestación de estos principios que se regule especialmente calidad del tratamiento de datos personales, que impone que estos sean tratados dentro del marco de una finalidad legítima y declarada por parte de quien realiza tratamiento s o de quien es responsable por dichos tratamientos. También es manifestación de este principio la temporalidad del tratamiento de datos, el deber de secreto que afecta a las personas que en razón de su trabajo participan en operaciones de tratamiento de datos personales y acceden a datos de esta naturaleza, obligación que persiste aún cuando la persona cese en las funciones.

Finalmente son manifestaciones de este principio las normas sobre acciones para reclamar por un tratamiento indebido de datos personales, como la acción de habeas data y de indemnización de perjuicios que puede ser entablada por la persona afectada por el tratamiento de datos personales. También, en la ley general de telecomunicaciones, estos principios se plasman a través del artículo 7 que prevé como una de las finalidades de los organismos reguladores del sector, es el procurar el pleno respeto a los derechos de los usuarios de los sistemas de telecomunicaciones, asimismo cuando sanciona la interceptación e intervención de las comunicaciones.

Capítulo 8

Conclusiones

Entia multiplicanda non sunt sine necessitate.
Guillermo de Ockham

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.

Por otro lado la personalización de los sitios no debe dar paso a la restricción de la libertad de navegación de los usuarios. Encasillarlos a ver sólo lo que el sistema de recomendación dice que es lo correcto, es casi como tratar a los usuarios como borregos que se dirigen a un lugar incierto y donde se les quiere extraer el mayor beneficio. Por este motivo, se plantea que la creación de sitios con capacidades de adaptación debe considerar la sugerencia como forma de incentivar al usuario a visitar una página y bajo ningún motivo un ocultamiento de la información que, a través de otro medio, como un buscador, sería pública y accesible. Lo anterior debe ser especialmente tomado consideración por los profesionales de la información habida cuenta de los desarrollos actuales en la consagración de un derecho a la neutralidad en la red, con tutela judicial efectiva, del que se derivarán sanciones específicas a las empresas que mediante cualquier sistema tecnológico impidan el libre acceso a toda la vasta gama de contenidos disponibles en la web.

La publicación de cualquier información relativa a una persona en la Web, debería ser hecha bajo su consentimiento, sobre todo si se trata de datos personales y una vez publicado, debería existir algún mecanismo de retracto o eliminación. Al respecto, hay que considerar que los actuales motores de búsqueda almacenan en sus bases de

datos una fotografía de la Web, es decir, aunque la página ya no exista, es posible que aparezca dentro de las posibilidades de una búsqueda.

Eliminar información sobre una persona de un buscador, no es un problema técnico, de hecho hay casos jurisprudencia internacional donde se ha obligado estas empresas a eliminar ciertos contenidos. Sin embargo, tuvo que haber un juicio de larga data para que que la eliminación de información fuera posible, por lo que debería existir un mecanismo más eficiente para que el usuario común solicite que ciertos datos que le competen no aparezcan en las búsquedas.

En lo referente a la minería de los archivos de web log, existe la tecnología y los algoritmos necesarios como para reconstruir la sesión de un usuario, sin necesidad de requerir datos personales acerca de él, evitándose tanto la posibilidad de una sanción penal por acceso ilícito a sistemas ajenos como la circunstancia de generarse un tratamiento de datos inadecuado en atención a la normativa de protección existente. Lo anterior es la base para el análisis de perfiles y grupos de usuarios que luego permite la mejora del sitio en contenido y estructura. Entonces, no existe una justificación directa a la persecución que realizan algunos sitios web para obtener hasta el más mínimo detalle de quienes los visitan. Es cierto que con técnicas no invasivas de análisis de navegación y preferencias, no se logra extraer todo el beneficio que se podría de los usuarios, pero si se puede lograr un justo equilibro entre la maximización de las utilidades de la organización y la salvaguarda de la privacidad y libertad de navegación de los usuarios.

Finalmente, la pregunta que motiva esta tesis 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.

Glosario

Acceso Universal Tiene como objetivo el proveer las condiciones necesarias para que los usuarios puedan acceder a los servicios de telecomunicación básico. Dicho de otra forma, es hacer llegar la red de telecomunicaciones a las áreas más remotas y aisladas del país., 98

ADSL Asymmetrical Digital Subscriber Line o Línea Asimétrica de Suscripción Digital. Esta tecnología utiliza las líneas telefónicas para la transmisión de datos a una tasa muy elevada., 74

ANSI American National Standards Institute, 35

CERN Conseil Europeen pour la Recherche Nucleaire. Fundado en 1954 por 12 países europeos, es el mayor laboratorio de la historia destinado a análisis de las partículas fundamentales que conforman la materia. Subtiende un anillo concéntrico de 27 km de perímetro, el cual se encuentra una profundidad que, según los tramos, oscila entre los 50 y los 150 metros y está situado en la frontera entre Francia y Suiza, entre la comuna de Meyrin (en el Cantón de Ginebra) y la comuna de Saint-Genis-Pouilly (en el departamento de Ain). Cada cierta cantidad de metros, existe adosado un laboratorio. El compartir los datos e informes que se obtenían en cada laboratorio, fue el desafío que impulsó el desarrollo de la Web., 89

Checksum Es un contador del número de bits presentes en una unidad de transmisión, de tal forma que el receptor pueda chequear si la cantidad de bits recibidos coincide con los enviados. En caso de haber un calce, se asume que la transmisión fue exitosa. Protocolos como el TCP y UDP proveen la verificación de checksum en forma nativa, 66

DHCP Dynamic Host Configuration Protocol o Protocolo Configuración Dinámica de Servidor. Es un protocolo que a través de una aplicación permite a los dispositivos de una red LAN obtener los parámetros de configuración del TCP/IP (IP Address, IP del Router, IP del DNS, etc.)., 65

Dial Up Define una forma de conexión de un dispositivo a una red o a Internet a través de un módem que utiliza como canal de envío de datos a la red pública telefónica. El módem se encarga de transformar los datos a transmitir en pulsos audibles, que son los que pueden ser enviados en el ancho de banda que utiliza la telefonía. Por esta razón, se logra una conectividad que no supera los 56Kbps., 73

DNS Domain Name System o Sistema de Nombre de Dominio. Se trata de un sistema de bases de datos distribuido, cuya función esencial es resolver (traducir) los nombres usados en los servicios presentes en Internet, a una dirección IP (Ej: wi.dii.uchile.cl → 146.83.5.58)., 67

E-business Negocios electrónicos que distinguen a cualquier actividad empresarial que utilizan la Internet como medio para sus transacciones comerciales., 1

ECMA European Computer Manufacturers Association, 35

Gateway Gateway o puerta de enlace, es un equipo de interconexión de redes que permite la comunicación entre redes de protocolos distintos. Normalmente, se utiliza como sinónimo del Router, por cuanto la mayoría de las redes presentes en Internet utiliza el mismo protocolo de comunicación., 39

Guía de ondas Medio sobre el cual se confina una señal para su propagación, por ejemplo un cable., 16

HTML Hyper Text Markup Language o Lenguaje de Marcas de Hipertexto. Es el lenguaje en el cual se construyen las páginas Web y está compuesto por diversos marcadores o **tags**, los cuales van rodeados por corchetes angulares ($<, >$). Cada tag es interpretado por el navegador para incluir un objeto dentro de la página, especificar un tipo de letra, etc., 90

HTTP HTTP de HyperText Transfer Protocol o Protocolo de Transferencia de Hipertexto es el método mediante el cual se transfieren las páginas web desde un sitio hacia el navegador del usuario., 5

Hub Hub o concentrador es un equipo para la interconexión computadores. Su operación se basa en la retransmisión de los datos que recibe desde un computador hacia el resto de los conectados a la red que implementa. Lo anterior genera muchas colisiones al momento que dos o más computadores tratan de transmitir sus datos, por lo que su uso se ha ido discontinuando y han sido desplazados por otros dispositivos más idóneos para asegurar un adecuado uso del ancho de banda., 36

IEEE Institute of Electrical and Electronics Engineers, 35

Internet Móvil Conjunto de tecnologías que permiten la conexión a Internet a través de dispositivos móviles, como pueden ser los celulares, Palm, netbooks y por supuesto los notebooks. Su implementación ha sido realizada principalmente utilizando la capacidad de redes instaladas para la telefonía celular. En estos momentos se están utilizando fuertemente las redes de tercera generación (3G), pero se espera que en un par de años más se produzca un cambio hacia redes de cuarta generación, cuya característica principal es la velocidad de conexión., 76

ISO International Standards Organization, 35

ISP Internet Service Provider o Proveedor del Servicio Internet. Se trata de empresas cuyo negocio es proporcionar acceso a Internet, a través de un contrato pecuniario que comúnmente es por una cuota fija mensual. El ISP provee al contratante de un nombre de usuario, una contraseña y un método de acceso a su red., 73

LAN Local Area Network o Red de Área Local. De esta forma se conceptualiza al grupo de computadores y dispositivos asociados que comparten una línea de comunicación común, ya sea un cable o un router inalámbrico. Este tipo de red fue desarrollada para compartir recursos que son costosos y de alta demanda, como por ejemplo una impresora avanzada. Usualmente, existe un servidor que almacena aplicaciones y datos de la red, con la finalidad de que los usuarios siempre puedan acceder a la última versión de estos. Su principal característica es la distancia geográfica para la cual fueron creadas. En efecto, una red LAN está pensada para atender los requerimientos de computadores de un mismo edificio, lo que en la práctica puede significar muy pocos usuarios, por ejemplo en una casa, o miles de estos, a través de una red de fibra óptica., 96

MAC Media Access Control address o dirección de control de acceso al medio, es un identificador hexadecimal de 48 bits que identifica de manera única a una tarjeta de red., 35

Modulación de onda Proceso mediante el cual una onda, denominada portadora, sufre variaciones en función de otra onda, conocida como moduladora, la cual contiene información que se desea transmitir. La modulación se puede realizar variando la amplitud, frecuencia o fase de la portadora., 15

NAT Network Address Translation o Traducción de Dirección de Red. Estándar para la utilización de una o más direcciones IP para conectar varias computadoras a una red (especialmente Internet). Cada computadora tiene una dirección IP distinta (generalmente no válida para Internet). Fue desarrollada por la IETF., 100

NCP Network Control Protocol o Protocolo de Control de Red., 54

Networking La idea detrás de toda red, es compartir recursos que son escasos o costosos. En este sentido, Networking simboliza una forma de trabajo en que las personas hace un uso eficaz y eficiente de la red social existente entre ellos, para compartir un bien tan preciado como lo es el conocimiento. Generalmente se hace uso intensivo de los medios electrónicos para potenciar y mejorar el networking., 84

NIC Network Information Center o Centro de Información de la Red. Se trata de una institución encargada de la asignación de los dominios en Internet, ya sean estos genéricos, de empresas, personas naturales, etc. Básicamente existe un NIC por cada país, el cual delega subdominos bajo su terminación, que en el caso de Chile es .cl., 69

NIST National Institute of Standards and Technology, 35

Nodo Así se denomina en Internet a un dispositivo conectado a la red, que posee un nombre y dirección real. Un sinónimo frecuentemente utilizados es hots., 87

OSi Open Systems Interconnection o Interconexión de Sistemas Abiertos. Se trata de una estandarización para el desarrollo de protocolos de comunicación propuesto en 1984 por el ISO.La idea es dividir el problema de la comunicación entre computadores en siete niveles, proporcionando a los fabricantes de dispositivos de red de un marco para el desarrollo compatible e interoperable entre distintas tecnologías., 40

POP3 Post Office Protocol 3 o Protocolo de Oficina de Correos versión 3. Es la más reciente versión del protocolo estándar para la recepción de correos electrónicos. POP3 permite la recepción y almacenamiento de los correos electrónicos para que puedan ser leídos por los usuarios, a través de aplicaciones tales como Eudora o Thunderbird., 72

Protocolo Del latín protocollum. Se trata de un conjunto de reglas o normas que posibilitan que se establezca una comunicación entre varios equipos o dispositivos. Antes de que se inicie la comunicación, los dispositivos deben acordar el protocolo de comunicación a usar (a esto se le denomina handshaking protocol)., 42

RFC Request For Comment - Petición de Comentarios. Con el objetivo de ir documentando los acuerdos y estándares sobre protocolos y en general avances en Internet, se crearon una seria de documentos en 1967, los RFC, que permite estudiar un acuerdo en específico y comentarlo., 80

Router El router o enrutador, ruteador o encaminador es un dispositivo para interconexión de interconexión de redes informáticas que permite asegurar el enrutamiento de datps entre redes o determinar la ruta que deben tomar los datos., 38

Sitio Web Conjunto de páginas y objetos web, tales como imágenes, sonidos, películas, etc., que se relacionan a través de hipervínculos, 1

SMTP Simple Mail Transfer Protocol o Protocolo Simple de transmisión de Correos. Este protocolo es usado para la transmisión de correos electrónicos entre computadores, usualmente sobre una red Ethernet. Su implementación se realiza a través de una aplicación servidora que recibe y transmite los correos electrónicos desde las aplicaciones clientes., 70

Switch Un switch o conmutador es un equipo de interconexión de redes de computadores que permite el envio de datos punto a punto entre dos computadores. En el caso de una red Ethernet, utiliza la dirección MAC para identificar exactamente a que computador debe enviar los datos transmitidos., 38

Url Uniform Resource Locator o Localizador de Recurso Uniforme. Es la forma en que se localiza un objeto en la Web. Su estructura está compuesta por el protocolo a utiliza, la dirección IP del dispositivo y finalmente camino donde está el objeto (protocolo://direcciónIP/directorio/objeto)., 90

WAN Wide Area Network o Red de Área Amplia. Así se le denomina a la red que permite unir computadores y dispositivos asociados, separados por geográficamente por áreas distantes en kilómetros, como puede ser una región, provincia e incluso un país. Usualmente estas redes vienen a interconectar redes más pequeñas, como pueden ser LANs corporativas. El ejemplo clásico de este tipo de red es justamente Internet., 96

WiFi Wireless Fidelity o Fidelidad Inalámbrica. Se trata de un grupo de estándares para redes inalámbricas basado en las especificaciones IEEE 802.11, desarrollado para la creación de redes LAN inalámbricas, pero que también permite la conexión a Internet., 35

Wimax Worldwide Interoperability for Microwave Access o Interoperabilidad Mundial para Acceso por Microondas. Realiza la transmisión de datos usando ondas de radio. Su principal ventaja es que permite dar acceso a un canal de datos a sectores donde la densidad de usuarios es muy baja, lo que haría inviable, del punto de vista económico, el uso de otros medios de transmisión como cables o fibra óptica. Para garantizar el estándar y la interoperabilidad, se creó Wimax Forum, organismo encargado de la certificación de los equipos que proveen distintos fabricantes, los cuales están calibrados para funcionar entre las frecuencias de 2,5 y 3,5 Ghz., 76

World Wide Web La Web es un conjunto de archivos escritos en lenguaje HTML que se entrelazan entre si a través de hipervínculos. Su transmisión sobre Internet se realiza a través del protocolo http, o versiones segurizadas de este, 1

Bibliografía

[1] Ararteko. *Derechos Humanos y Nuevas Tecnologías*. Defensoría del Pueblo del País Vasco, San Sebastián, España, 2003.

[2] R. Arrieta-Cortés and C. Reusser-Monsálvez, editors. *Chile y la Protección de Datos Personales. ¿Están en crisis nuestros derechos fundamentales?* Universidad Diego Portales, Santiago, Chile, 2009.

[3] L.A. Ballesteros-Moffa. *La Privacidad Electrónica*. Tiranto Lo Blanch, Valencia, España, 2006.

[4] 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.

[5] R. Herrera Bravo. Reflexiones sobre los delitos informáticos motivadas por los desaciertos de la ley chilena. *Revista de Derecho Informático Alfa-Redi*, 5, Diciembre 1998.

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

[7] J.H. Cheong and M.C. Park. Mobile internet acceptance in korea. *Internet Research*, 15(2):125–140, 2005.

[8] 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.

[9] 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.

[10] M.A. Davara-Rodríguez. *La Protección de Datos Personales en el Sector de las Comunicaciones Electrónicas*. Universidad Pontificia Comillas, Madrid, España, 2003.

[11] República de Chile. Ley general de telecomunicaciones. `http://www.bnc.cl/`, 1982. [En línea; visitado el 01-Febrero-2010].

[12] República de Chile. Modifica el código penal, el código de procedimiento penal y el código procesal penal en materia de delitos de pornografía infantil. `http://www.bnc.cl/`, 2004. [En línea; visitado el 23-Febrero-2010].

[13] República de Chile. Sistema nacional de registro de adn. `http://www.bnc.cl/`, 2004. [En línea; visitado el 19-Diciembre-2009].

[14] Reública de Chile. Ley sobre protección de datos de carácter personal. `http://www.bnc.cl/`, 2002. [En línea; visitado el 19-Diciembre-2009].

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

[16] Real Academia de la Lengua Española. Definición de telecomunicación. `http://www.rae.es/`, 2006. [En línea; visitado el 19-Diciembre-2009].

[17] Unión Internacional de Telecomunicaciones. Convenciones sobre el uso del espectro radioeléctrico. `http://www.itu.int/`, 2009. [En línea; visitado el 10-Enero-2010].

[18] V. Drummond. *Internet, Privacidad y Datos Personales*. Reus, Madrid, España, 2004.

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

[20] Internet Engineering Task Force. The internet engineering task force. `http://www.ietf.org`, 2009. [En línea; visitado el 2-Enero-2010].

[21] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proc. 9th ACM Conference on Hypertext and Hypermedia*, pages 225–234, 1998.

[22] Web Intelligence Research Group. Papers and technical reports on web intelligence. `http://wi.dii.uchile.cl/`, 2009. [En línea; visitado el 19-Diciembre-2009].

[23] A. Harb, M. Plantiè, and G. Dray. Web opinion mining: How to extract opinions-from blogs? In *Proc. 9th CSTST*, pages 211–217, 2008.

[24] J.M. Huidobro-Moya. *Redes y Servicios de Telecomunicaciones*. Thomson Editores, Madrid España, 2002.

[25] 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.

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

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

[28] B. Leiner, V. Cerf, D. Clark, R. Kahn, L. Kleinrock, D. Lynch, Jon Postel, L. Roberts, and S. Wolff. A brief history of the internet. *ACM SIGCOMM Computer Communication Review*, 39(5):22–31, 2009.

[29] D. Lyon. *El Ojo Electrónico. El auge de la sociedad de la vigilancia*. Alianza, Madrid, España, 1995.

[30] 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.

[31] C.A. Meneses. Delitos informáticos y nuevas formas de resolución del conflicto penal. *Revista de Derecho Informático Alfa-Redi*, 51, Octubre 2002.

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

[33] 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.

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

[35] C.P. Márquez-Escobar. *El Ojo Ve, el Poder Mira. La arquitectura para la vigilancia y el fin de la privacidad*. Pontificia Universidad Javeriana, Bogotá, Colombia, 2004.

[36] J. Nielsen. User interface directions for the web. *Communications of ACM*, 42(1):65–72, 1999.

[37] N. García Noguera. Delitos informáticos en el código penal español. *Delitos Informáticos*, 2002.

[38] H. Nyquist. Certain topics in telegraph transmission theory. *Trans. AIEE*, 47(1):617–644, 1928.

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

[40] P. Poblete. La verdadera y real historia de internet en chile. `www.dcc.uchile.cl/~ppoblete/sigloxxi-27Feb96.html`, 1996. [En línea; visitado el 19-Diciembre-2009].

[41] A.E. Pérez-Luño. *La Tercera Generación de Derechos Humanos*. Aranzadi, Cizur Menor (Navarra), España, 2006.

[42] 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.

[43] C.E. Serra-Uribe. *Derecho a la Intimidad y Videovigilancia Policial*. Laberinto, Madrid, España, 2006.

[44] C. E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37(1):10–21, 1949.

[45] Internet Society. The internet society. `http://www.isoc.org/isoc/`, 2009. [En línea; visitado el 2-Enero-2010].

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

[47] A.S. Tanenbaum. *Computer Networks*. Prentice Hall., New Jersey USA, 2002.

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

[49] H.T. Tavani and J. Moor. Privacy protection, control of information, and privacy-enhancing technologies. *Computers and Society*, 1:6–11, 2001.

[50] A. Vedder. Kdd: The challenge to individualism. *Ethics and Information Technology*, 1:275–281, 1999.

[51] 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.

[52] J.D. Velásquez. Conversión de texto a voz. Master's thesis, Departamento de Ciencias de la Computación, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile, 2001.

[53] J.D. Velásquez and Lorena Donoso. Web mining: Análisis sobre la privacidad del tratamiento de datos originados en la web. *Revista Ingeniería de Sistemas*, 23(1):5–26, 2009.

[54] 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.

[55] J.D. Velásquez and V. Palade. Building a knowledge base for implementing a web-based computerized recommendation system. *International Journal of Artificial Intelligence Tools*, 16(5):793–828, 2007.

[56] 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.

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

[58] 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.

[59] C. Villagrasa-Alcaide. *Nuevas Tecnologías de la Información y Derechos Humanos*. Cedecs, Barcelona, España, 2003.

[60] L. Wang, X.Zhao, D. Pei, R. Bush, D. Massey, and L. Zhang. Protecting bgp routes to top-level dns servers. *IEEE Transactions on parallel and distributed systems*, 14(9):1–10, 2003.

[61] A. Weiss. Net neutrality?: there's nothing neutral about it. *netWorker*, 10(2):18–25, 2006.

[62] L.V. Wel and L. Royakkers. Ethical issues in web data mining. *Ethics and Information Technology*, 6:129–140, 2004.

[63] El Mercurio: Ciencia y Tecnología. Chile es el primer país del continente en probar la conexión 4g. `http://diario.elmercurio.com/2010/01/27/ciencia_y_tecnologia/ciencia_y_tecnologia/noticias/20F2319D-3CA3-45B4-9327-A4D63F10F591.htm?id=`, 2010. [En línea; visitado el 29-Enero-2010].

Índice alfabético

Árboles de Decisión, 118
1G,2G,3G,4G, 81

AAN, 117
Acceso Universal, 104
ADSL, 79, 107
ANSI, 37
ARPANET, 92
Artificial Neural Networks , 117

B2B, 105
B2C, 105
Bitnet, 93
Bits, 23

CERN, 95
Checksum, 23, 71
Clasificación, 116
Clustering, 116
Codificación Binaria, 23
Consejo Nacional de Nombres de Dominio y Números IP, 85
CSMA/CD, 36

DARPA, 92
Datagrama, 57, 75
DHCP, 70
Dial Up, 78
DNS, 73
Domain Name System, 73

E-business, 1
ECMA, 37
Espectro
    Electromagnético, 15
    Radioeléctrico, 17
European Organization for Nuclear Research, 95

Firewall, 102

Gateway, 41

Host, 93
HTML, 96
HTTP, 6, 59
Hub, 38

IAB, 84
IANA, 74, 84
ICANN, 74, 84
IEEE, 37
IETF, 84
Internet, 6, 93
    Activities Board, 84

Assigned Numers Authority, 84
Common Address Notation, 60
Corporation for Assigned Names and Numbers, 84
Engineering Task Force, 84
móvil, 81
Research Task Force, 84
Internet Protocol, 56
Internet Society, 84
ISO, 37
ISOC, 84
ISP, 78, 94, 106, 138

K-Means, 118

LAN, 55, 92, 102

MAC, 37
Modulación
Amplitud, 16
Fase, 16
Frecuencia, 16

NAT, 106
NCP, 56
Network Control Protocol, 56
Network Information Center, 84
Networking, 90
NIC, 74, 84
NIST, 37
Nodo, 93
Nsfnet, 93
Nyquist, 20
Frencuencia de, 20

Open Systems Interconnection, 42
OSI, 42
OUI, 37

P2P, 105
Paquetización, 71
Patrones de Secuencia, 115
Personalización, 2, 105, 111, 139
POP3, 59, 77
Privacidad, 4, 139
Protocolo, 44, 91
Proxy, 102

Redes Neuronales Artificiales, 117
Reglas de asociación, 115
Resolución del nombre, 73
RFC, 86
Router, 41, 102

Señal, 10
Amplitud, 14
Análoga, 18
Digital, 19
Energía, 12
Frecuencia, 14
Muestreo, 23
Potencia, 13
Self Organizing Feature Maps, 118
Sesionización, 113

Shannon, 20
SMTP, 59, 75
SOFM, 118
Spyware, 103
Support Vector Machines, 118
SVM, 118
Switch, 40

Tablas de ruta, 72
TCP, 56, 71
TCP/IP, 6, 56, 71, 134
Teorema del Muestreo, 20
Transmission Control Protocol, 56

UDP, 59
Url, 96

WAN, 102
Web, 1
    Crawlers, 103
    Data, 138
    data, 111, 112
    mining, 111, 115
    Personalización, 2
    site, 1
    World Wide, 1
Web 2.0, 123
Web mining, 111, 115
WiFi, 37, 82
Wimax, 81