

Capítulo 10

Clasificación y Filtrado de Información en la “Web Viva”

Carlos Hurtado Larraín

Gran parte de la Web corresponde a información estable o que cambia lentamente. Ésta incluye sitios corporativos y personales casi estáticos, conocimiento “enciclopédico” e información que se revisa poco a través del tiempo. Hay otra Web, llamada “Web viva”, que se refresca minuto a minuto, que está compuesta, principalmente, por sitios de noticias, weblogs y comunidades digitales. Lo que interesa a los usuarios de esta Web es lo novedoso, lo que apareció en el último día, en las últimas horas, o incluso minutos. Es la Web en la que nadie se baña dos veces en la misma información. El adjetivo “viva” no sólo apela a su dinamismo, sino a que su contenido, videos, fotografías, artículos, etc., es generado por comunidades digitales donde interactúan millones de personas en el mundo: la llamada Web 2.0 [14] con aplicaciones como Flickr, YouTube, Del.icio.us, Facebook, Twitter, etc. y los más de 70 millones de weblogs y variantes como videoblogs, linklogs y fotoblogs del planeta.

Este espacio de información fue recién tomado en cuenta por los principales buscadores de la Web (Google, Yahoo!, MSN) un par de años atrás. En ese entonces, la instantaneidad de la información no era requerimiento atendido por estos sistemas de búsqueda. Entregar información fresca era en

cierto modo incompatible con la tarea titánica de los buscadores de recolectar miles de millones de páginas en costosos recorridos de la Web. Mientras a fines del año 2005, los grandes buscadores sólo actualizaban el contenido de una página cada 10 ó 15 días, surgían buscadores como Technorati, Bloglines y Blogpulse, entre otros, que se posicionaron en la Web viva, conquistando un segmento de usuarios considerable en muy poco tiempo.

La dinámica de la Web viva se asemeja más a la forma en que la información viaja desde canales de comunicación en radio y televisión a las personas, que al concepto inicial de la Web como una gran biblioteca digital compartida. Sin embargo, los principios de la Web siguen operando con fuerza: red distribuida, con contenido enlazado (hipertexto), libertad de generar y consumir información, millones de canales y receptores latentes. En este capítulo explicaremos los conceptos que predominan en este nuevo contexto: canales, agregadores de información y sindicación de contenido, entre otros, y mostraremos el problema de filtrar información, una de las principales tareas para manejar la sobrecarga de información a la que este nuevo escenario nos expone.

Sindicación de Contenido

La Web viva es un espacio donde la información se disemina en forma automática y a gran velocidad. Aquí es común que una noticia publicada en un sitio local se propague casi en forma instantánea a cientos o miles de sitios en pocas horas y, casi en paralelo, sea recolectada por la mayoría de los buscadores. Esta instantaneidad es sostenida (aparte de la Web misma) por la infraestructura de “sindicación de contenido” de la Web. Sindicar contenido significa hacer disponible contenido para que otros puedan publicarlo, procesarlo o redistribuirlo. El concepto, mucho más antiguo que la Web mis-

ma, proviene del mundo de los medios de prensa, radio y televisión, donde contenido como fotografías, videos y noticias, entre otros, es diariamente sindicado alrededor del planeta.

La sindicación de contenido es una práctica cada día más extendida en la Web: compañías de música sindicán información sobre discografía que luego es publicada por sitios de comercio electrónico; bolsas de comercio sindicán información en línea sobre el valor de acciones que es procesada por portales financieros; la mayoría de las comunidades digitales emergentes están sindicando información con el objeto de llegar cada día a más usuarios.

En la Web, la información sindicada es procesable por computadores, es decir, es fácil para un programa computacional sencillo, detectar los atributos más importantes de un artículo, video, imagen, etc. sindicado. Para que esto sea posible existen formatos que permiten describir la información sindicada. El más antiguo de estos formatos, “RDF Site Summary” (RSS), fue desarrollado por Ramanathan Guha, mientras trabajaba para Netscape, el año 1999. En poco tiempo, RSS derivó en una colección de formatos que incluye “Really Simple Syndication”, “RDF Site Summary” y “Rich Site Summary” [2]. En 2003 apareció un nuevo formato alternativo, Atom, apoyado por el consorcio de la Web (W3C) con la finalidad de unificar las propuestas anteriores. En la actualidad, RSS y Atom (en adelante usaremos el término RSS para referirnos a ambos formatos) compiten por establecerse como estándares *de facto* en la Web. El potencial de estos formatos es enorme, por ejemplo, hoy podemos recolectar RSS sindicado de diversas fuentes, combinarlo y procesarlo para producir nuevo RSS (lo que se denomina “mashup”) que a la vez podemos syndicar para que otros lo recolecten, y así sucesivamente, en una suerte de cadena alimenticia donde la información se

transforma, sintetiza y combina, desde sus fuentes hasta el usuario que la consume.

Canales y Agregadores de RSS

En la Web de la década pasada, los usuarios debían esforzarse por encontrar información, ya sea mediante buscadores o navegando enlaces. Hoy, podemos acceder a una gran cantidad de información de interés sólo esperando que ésta llegue a nosotros. Para que esto sea posible, las fuentes de información de la Web viva, llamados “canales”, publican RSS sobre información sindicada. Este RSS es recolectado en forma periódica y mostrado en la pantalla del usuario final por aplicaciones conocidas como “agregadores”. Estos sistemas entregan un flujo continuo de RSS, que referencian videos, fotografías, animaciones, artículos, noticias, etc, provenientes de canales tan diversos como medios de prensa, sitios de tecnología o weblogs.

En la actualidad, existe una oferta de cientos de agregadores RSS, la que incluye sistemas basados en la Web, como Yahoo! Pipes o Google Reader, o agregadores que se instalan como software cliente en computadores personales, PDA's o teléfonos móviles. Adicionalmente, los principales navegadores y lectores de correo electrónico están incorporando funciones de agregadores.

También hay agregadores que recolectan RSS para comunidades de usuarios. Este es el caso de Orbitando [12] (ver figura 10.1), que se enfoca en personas interesadas en contenido relacionado a Chile, o Topix [13], que se enfoca en una comunidad más amplia.



Figura 10.1: Portada de Orbitando [13].

Filtrado y Clasificación de Información

Los canales y agregadores nos permiten acceder a una enorme cantidad de información. Esta es sin duda una buena noticia. Clasificar y filtrar información son dos tareas fundamentales para manejar la sobrecarga de información en este nuevo contexto.

Filtrar información es la tarea de dejar pasar parte de ésta y bloquear otra de acuerdo a un objetivo. En algunas situaciones el objetivo es evitar información como contenidos no aptos para menores o publicidad no solicitada. Un ejemplo muy popular es el filtrado de correo electrónico no deseado (*spam*). En otros casos, necesitamos filtrar para descartar información irrelevante que constituye ruido. El filtrado de información también puede tener como objetivo personalizar y ajustar los agregadores de acuerdo a los intereses de un usuario o una comunidad de usuarios.

Clasificar es una tarea similar. En este caso, debemos decidir una o más categorías, entre un conjunto fijo de éstas, a las que asociamos determinada información, como cuando organizamos los archivos de nuestro computador en carpetas. Es común en la Web que las categorías sean tópicos, que incluso pueden formar estructuras jerárquicas donde los más específicos se conectan con los más generales. En otros casos, las categorías pueden referirse a alguna propiedad de la información como su tipo u origen. Por ejemplo, podríamos necesitar clasificar texto para detectar comentarios positivos y negativos. En el extremo derecho de la figura 10.1 se pueden ver las categorías en que un agregador clasifica RSS. Se consideran tópicos como política, negocios, tecnología, etc. y tipos de información como weblogs, videos, fotografías, *podcasts*, etc.

Hoy en día, los usuarios comunes de agregadores sólo pueden filtrar manualmente una fracción mínima del flujo de información que pueden recibir. También es poco práctico pensar en editores que hagan este trabajo, como suele ocurrir en medios de prensa tradicionales. El Open Directory Project [11], una ambiciosa iniciativa de comprometer editores humanos para clasificar la Web, gozó de gran popularidad en sus inicios a fines de los noventa, pero su impacto decreció en los últimos años.

Los Primeros Filtros Automáticos

A fines de los ochenta, tomó fuerza el desarrollo de programas que filtran en forma automática. Uno de los primeros de estos sistemas, *CONSTRUE*, implementado inicialmente para la agencia de noticias Reuters, permitía programar filtros basados a reglas modeladas por expertos. Por ejemplo, la siguiente regla, mencionada con frecuencia en libros del area, determina si un artículo es o no relevante para la categoría “trigo”:

```
if ( (trigo and predio) or (trigo and commodity) or
    (quintal and exportar) or (trigo and tonelada)
    or (trigo and invierno and not suave))
then clase=relevante
else clase=irrelevante
```

El antecedente de la regla (la condición a la izquierda del símbolo “*then*”) usa operadores lógicos como *and*, *or* y *not*. Cada término de esta condición es verdadero si el término aparece en el artículo. En el ejemplo, si el artículo satisface el antecedente de la regla, es clasificado como relevante, en caso contrario es clasificado como irrelevante.

Algunos experimentos iniciales mostraron que la tasa de error de un filtro generado por CONSTRUE podía ser menor a 10%. A pesar de estos resultados positivos, por distintos motivos, el método de CONSTRUE se tornó rápidamente impracticable en la mayoría de las aplicaciones donde se utilizó. En primer lugar, el tiempo y costo que toma tener expertos definiendo reglas es alto. Más aún, si lo que se considera relevante cambia, los expertos deben intervenir de nuevo las reglas, y en algunos casos el trabajo debe hacerse desde cero. La información es en general dinámica y las reglas de un filtro deben evolucionar constantemente. Por ejemplo, el interés de una comunidad a la cual se enfoca un agregador puede estar en constante cambio, o debemos reprogramar el filtro continuamente para incorporar nuevos términos.

Si bien sistemas como CONSTRUE permiten programar sistemas que filtran en forma automática, hoy es claro que el problema de fondo es mucho más complejo: requerimos de sistemas que aprendan a filtrar en base a una adaptación continua las necesidades de información de los usuarios. No solamente es importante automatizar el proceso de filtrado sino también el proceso de construcción y adaptación de un filtro.

Filtros que Aprenden y se Adaptan

Disciplinas como estadística, aprendizaje de máquinas, reconocimiento de patrones y, últimamente, minería de datos [3,4,5] son la base para desarrollar filtros de información que aprenden y se adaptan en base a la experiencia. Para que este proceso de aprendizaje se lleve a cabo, debemos contar con información ya filtrada, es decir, ejemplos positivos y negativos, denominada *datos de entrenamiento*, que se pueden generar por expertos o vía *feedback* de usuarios comunes. Estos datos se usan para entrenar o *inducir* el filtro. Una forma de pensar en este proceso es que a medida que incluimos más datos en el entrenamiento, el sistema incorpora nuevas *reglas*, siempre teniendo cuidado de que éstas se puedan generalizar a información más allá de los datos de entrenamiento. La figura 10.2 muestra un ejemplo de un proceso de entrenamiento de un modelo para clasificar vinos.

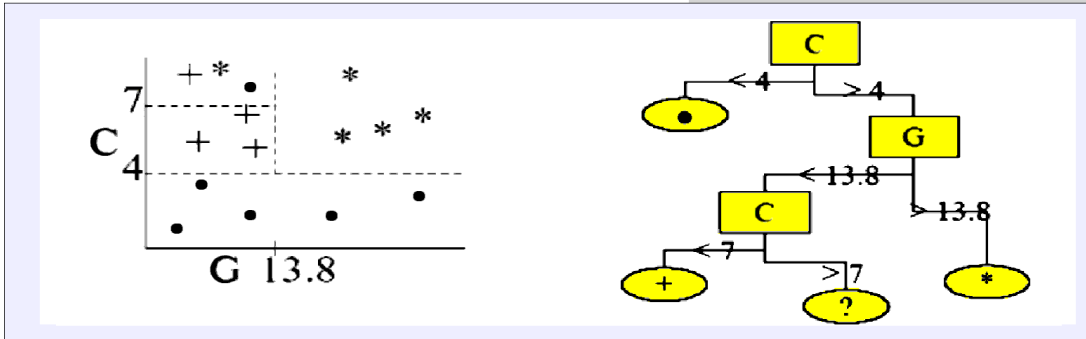
En este proceso es muy importante evaluar el desempeño del sistema creado, es decir, medir su capacidad para predecir correctamente las categorías de nueva información que se presenta. En términos simples, esto se hace separando de los datos de entrenamiento un nuevo conjunto, llamado “datos de prueba”, que usamos para medir la tasa de error. En general, es importante distinguir distintos tipos de error (falsos positivos y falsos negativos). Por ejemplo, en un agregador de contenido para niños es mucho más grave el error de dejar pasar información no apta que muestra violencia o pornografía, que el error de descartar alguna información adecuada.

Hoy en día existen cientos de técnicas para desarrollar filtros de información, algunas de las cuales han alcanzado tasas de error menores a un 10% en diversos experimentos. Entre estas están los árboles de decisión, máquinas de soporte vectorial, redes neuronales, redes bayesianas, discriminantes lineales, regresión logística, etc. En la actualidad, estas técni-

PROBLEMA

Imaginemos que deseamos clasificar el vino en tres categorías: *, + y •. Nuestros datos de entrenamiento son 15 vinos ya clasificados por enólogos expertos. Luego de usar algunas técnicas estadísticas, vemos que el color (C) y grado alcohólico (G) permiten diferenciar los tipos de vino. En efecto, al graficar los distintos tipos de vinos, vemos que C y G separan bastante bien los tipos de vino en rectángulos.

Un problema esencial al construir un clasificador es encontrar las variables que discriminan o separan las clases. Estas variables forman “espacio vectorial” donde cada objeto a ser clasificado se representa como un punto. En nuestro ejemplo, este espacio está definido por las variables C y G.



EL PROCESO DE APRENDIZAJE DEL FILTRO

El proceso de aprendizaje o inducción del filtro, en este caso un árbol de decisión, comienza con encontrar una variable y una condición sobre ésta que mejor separe las clases. La variable la ponemos al tope del árbol y la condición en los arcos que salen de éste. En nuestro modelo hemos elegido la variable C y la condición es $C \leq 4$, $C > 4$. Esta elección la hacemos en base a medidas estadísticas (las dos más conocidas son la *entropía* y el *índice Gini*). En la primera decisión del árbol, los datos se dividen en dos conjuntos, el de la izquierda sólo contiene vinos en la clase • y por lo tanto es muy “puro”. Debido a que el conjunto asociado al nodo de la derecha no define una región “pura”, el modelo debe seguir siendo refinado, ahora usando los datos de esa región (es decir todos los vinos con $C > 4$).

Este proceso se repite, agregando más decisiones, hasta que se llegan a regiones puras o con pocos datos. En el árbol de arriba, la región de pocos datos se representa con “?” debido a la falta de certeza sobre su clase. Cuando el árbol está terminado, hemos inducido el conocimiento de los enólogos en el modelo.

Figura 10.2: Construcción mediante aprendizaje de un árbol de decisión para filtrar vinos.

cas son usadas con éxito en distintas aplicaciones, no sólo en el contexto de la Web, sino en problemas tan variados como reconocimiento de voz, clasificación de imágenes telescópicas en astronomía o evaluación de riesgo financiero.

Nuevas ideas y mejoras se desarrollan en la actualidad para bajar las tasas de error ¿Podremos tener sistemas computacionales con capacidades de aprendizaje y desempeño similar a seres humanos? Para ello necesitamos desarrollar sistemas que emulen capacidades cognitivas humanas como comprensión de lenguaje natural, captura de sentido común y otras formas de procesamiento avanzado para llegar a la semántica de la información.

Filtrado Colaborativo

Un enfoque radicalmente distinto y de mucha aplicación en la actualidad, conocido como “filtrado colaborativo” [6], se basa en la idea de que la información relevante para un usuario es también relevante para otros usuarios con preferencias similares. Una comunidad de usuarios puede en conjunto actuar como un gran filtro espontáneo, si combinamos e interpretamos adecuadamente las acciones de cada uno de sus miembros.

El filtrado colaborativo no es más que la sistematización de un método de sentido común que aplicamos a decisiones de la vida diaria. Por ejemplo, si intentamos seleccionar una película para ver en el cine, podríamos primero buscar personas con gustos similares a los nuestros, para luego elegir alguna películas preferidas por estas personas. Esta elección, en muchos casos, será más acertada que la que haríamos después de conocer información intrínseca de las películas. El método de filtrado colaborativo es útil en especial cuando es complejo y costoso analizar la información a procesar, como sucedería si ésta está compuesta por videos, imágenes, audio, etc.

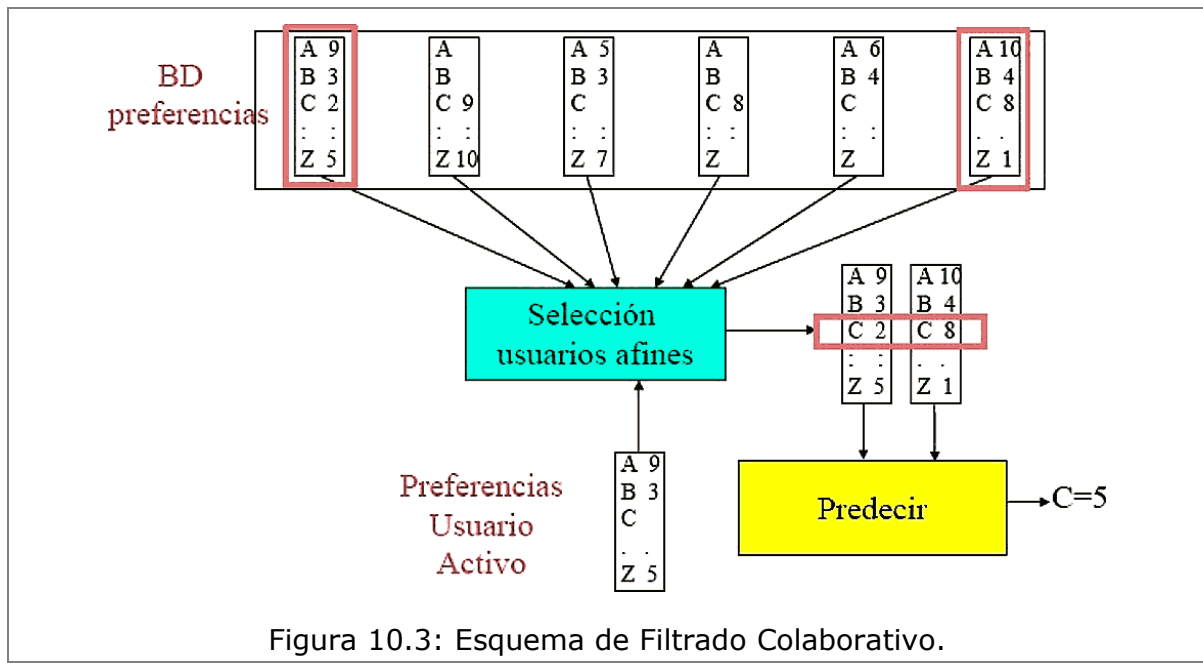


Figura 10.3: Esquema de Filtrado Colaborativo.

El método de filtrado colaborativo se explica, a *grosso modo*, en la figura 10.3. Contamos con una base de datos de preferencias donde cada rectángulo representa las notas (de 1 a 10) con que califica cada usuario un conjunto de artículos (denotados de A a Z). Un usuario particular, que llamaremos X, también ha evaluado algunos artículos, pero no conoce el artículo C. Entonces el sistema puede predecir una nota para este artículo que refleje la opinión de X. Para hacer esto en una primera etapa, se identifica un grupo de usuarios afines a X, por ejemplo, buscamos a aquellos cuyas notas tengan mayor correlación con las notas de X. Como resultado de esta etapa, seleccionamos dos usuarios. Finalmente, el sistema predice la nota de X como un promedio simple de las notas para el artículo C de los dos usuarios seleccionados.

La técnica de filtrado colaborativo tiene en la actualidad muchas aplicaciones debido a la proliferación de comunidades digitales en la Web que registran información de preferencias de sus usuarios. Estas preferencias

pueden ser implícitas, como selecciones (“clicks” o compras de productos), o explícitas, como comentarios o notas. Dos casos de aplicaciones muy citadas son el sistema de recomendación de productos de Amazon y Netflix, un sistema Web recomendador de películas. El método de filtrado colaborativo es la base de las nuevas generaciones de agregadores que permiten portadas de información personalizadas.

El Rol de los Tags

Otro enfoque colaborativo para clasificar y filtrar se basa en el fenómeno de “etiquetado social” (“social tagging”) que es la acción de usuarios de la Web de marcar recursos con “etiquetas” (“tags”), es decir, con términos que confieren semántica a los recursos. Las etiquetas representan entidades como personas, eventos, lugares, conceptos, etc. Gran parte de la información de la Web viva está sujeta a un intenso etiquetado social. Las etiquetas se publican en los archivos RSS asociados a información sindicada y pueden ser vistas como categorías de sistemas de clasificación, llamados folksonomías (neologismo que combina la palabra griega “clasificar” con la alemana “pueblo”) que, a diferencia de las taxonomías clásicas, evolucionan con gran dinamismo producto de la creación y desaparición continua de etiquetas.

La figura 10.4 muestra “nubes de etiquetas” de Orbitando (izquierda) y Technorati (derecha). Estas estructuras muestran las etiquetas más populares asociadas a una colección de documentos. El tamaño de cada etiqueta en la nube nos dice su peso o popularidad en la colección de documentos.

En la actualidad, las nubes de etiquetas son estructuras muy populares. Sin embargo, debido a que las etiquetas se crean libremente, las nubes pueden ser caóticas (como por ejemplo la nube de Technorati que se muestra en la figura 10.4 (derecha)), debido a sobreposición (dos o más etiquetas con



muchos documentos comunes), sinonimia (dos etiquetas o más que significan lo mismo), polisemia (una etiqueta con más de un significado) y otros problemas. Adicionalmente, no siempre disponemos de etiquetas. Un área extensa de investigación, denominada “extracción de información” [8], estudia el problema de generar etiquetas desde colecciones de texto plano e identificar relaciones semánticas entre ellas.

Conclusión

La Web viva ha generado una nueva dinámica de acceso a la información que está presentando desafíos científicos y tecnológicos importantes. En este contexto, la información “fluye” desde canales hacia agregadores que la deben filtrar y clasificar para finalmente presentarla a los usuarios.

Hoy, la mayoría de la información en la Web tiene las propiedades de un flujo. Los sistemas computacionales que filtran deben tener la capacidad de adaptarse continuamente a éste y a los requerimientos cambiantes de los

usuarios. Estos sistemas deben ser capaces de interpretar información como selecciones, votos, transacciones y etiquetas para sacar provecho de la dinámica social y colaborativa de la Web actual.

Agradecimientos. Se agradece a Carlos Orrego y José María Hurtado por sus aportes y sugerencias que contribuyeron a mejorar este artículo.

Para saber más

- ◆ En el sitio Desarrollo Web hay un tutorial sencillo sobre RSS: <http://www.desarrolloweb.com/articulos/2101.php>
- ◆ KD Nuggets es un sitio dedicado a la minería de datos, descubrimiento de información y minería Web. <http://www.kdnuggets.com/>

Referencias

1. Soumen Chakrabarti. Mining the Web Discovering Knowledge from Hypertext Data. Morgan-Kaufmann Publishers, 2002.
2. Ben Hammersley. Content Syndication with RSS. O'Really, 2003.
3. R. Feldman, J. Sanger. The Text Mining Handbook: Advanced Approach in Analyzing Unstructured Data. Cambridge University Press, 2007.
4. D. Hand, H. Mannila, P. Smyth Principles of Data Mining. The MIT Press, 2001.
5. J. Han and M. Kamber. Data Mining Concepts and Techniques. Morgan Kauffman Publishers, 2001.
6. John S. Breese; David Heckerman; Carl Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering Fourteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufman, 1998.
7. P. Jackson, I. Moulinier. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. John Benjamins Publishing Co. 2002.
8. GroupLens Research. Movielens. <http://movielens.umn.edu>
9. Nielsen/NetRatings. <http://www.netratings.com>
10. NewsMap. www.marumushi.com/apps/newsmap
11. Open Directory Project. www.dmoz.com

12. Orbitando. www.orbitando.com
13. Topix. www.topix.net
14. Tim O'Reilly. What Is Web 2.0. O'Reilly Network. Septiembre, 2005.
15. Fabrizio Sebastiani Machine learning in automated text categorization. ACM Computing Surveys (CSUR) archive Volume 34, Issue 1, March 2002.

Capítulos y Autores

1- La Web como Espacio de
Información Universal

Claudio Gutiérrez

2- Anatomía de la Web

Ricardo Baeza Yates

3- Internet

José Miguel Piquer

4- Buscando en la Web

Gonzalo Navarro

5- Manejo de Grandes Volúmenes de
Información utilizando Clusters de
Computadores

Mauricio Marín

6- XML: Transformando la Web en una
Base de Datos

Marcelo Arenas

7- Uso y Búsqueda de Información
Geográfica en la Web

Andrea Rodríguez

8- Multimedia en la Web

Javier Ruiz del Solar

9- Redes Sociales

Javier Velasco

10- Clasificación y Filtrado de
Información en la “Web Viva”

Carlos Hurtado Larraín



9 789563 192251

Cómo funciona la Web

Internet llega a Chile en 1992 y desde entonces, su crecimiento ha sido explosivo. Pocos años después se desarrollarían las aplicaciones que permitirían a todos los usuarios aprovecharla, y es lo que se conoce como Web.

Actualmente la mayoría de las personas en Chile se conectan a Internet y hacen uso de la Web diariamente, o al menos, en forma semanal. Pero ¿cómo funciona la Web?

En este libro de difusión, los investigadores del Centro de Investigación de la Web nos explican los detalles del funcionamiento de Internet y la WWW, abriendo los enigmas de los buscadores de Internet, la Web social y el futuro de la Web Semántica.

Centro de Investigación de la Web

El Centro de Investigación de la Web, como su nombre lo indica, es un Centro de investigación y desarrollo del Departamento de las Ciencias de la Computación de la Universidad de Chile, dedicado a la investigación básica en Ciencia de la computación, particularmente enfocado a la investigación en la Web.

Sus principales áreas de investigación son la recuperación de información, la minería en la Web, los aspectos lógicos y semánticos de la Web, y el procesamiento de información geográfica y multimedial.

El CIW es reconocido como uno de los centros de mayor calidad en este campo a nivel mundial, atrayendo a los más destacados investigadores internacionales a su equipo.

Con este libro, el CIW busca acercar su trabajo a los jóvenes chilenos, explicando los conceptos básicos de su trabajo en términos simples.