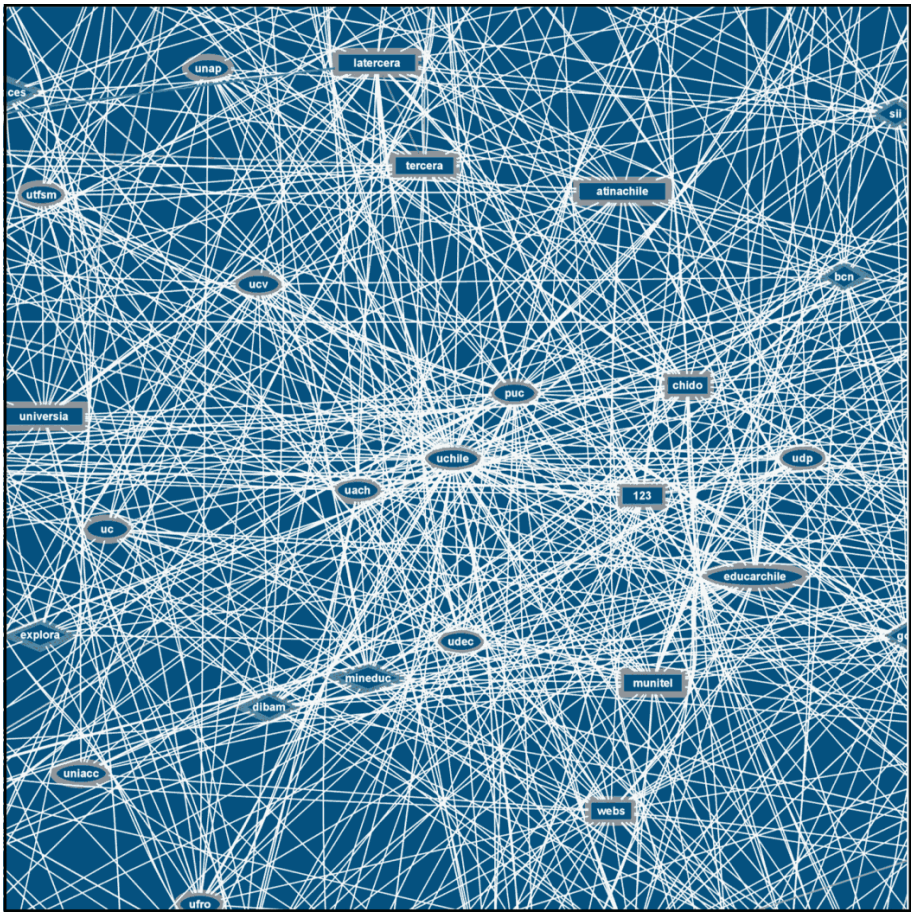
Cómo funciona La Web

Centro de Investigación de la Web

Universidad de Chile



ciw.cl centrodeinvestigaciónde laweb
departamento de ciencias de la computación
UNIVERSIDAD DE CHILE
INICIATIVA CIENTIFICA MILENIO ICM

Cómo funciona la Web

Centro de Investigación de la Web
Universidad de Chile

CIW es un Núcleo Científico Milenio

Cómo funciona la Web

Centro de Investigación de la Web
Departamento de Ciencias de la Computación
Universidad de Chile

CIW es un Núcleo Científico Milenio

Registro de Propiedad Intelectual Número 169174, Chile
ISBN: 978-956-319-225-1

Publicación Autoeditada.
Primera Edición, Junio 2008.
Santiago de Chile.

Editor General: Claudio Gutiérrez Gallardo.

Distribución gratuita de ejemplares impresos para colegios y universidades chilenos.
Distribución gratuita de versión digital a través de www.ciw.cl

Gráfico de la Portada: Eduardo Graells, en Características de la Web Chilena, estudio dirigido por Ricardo Baeza-Yates desde 2001.
Diseño de Portada: Javier Velasco M.
Diseño Interior: Mauricio Monsalve M.

Impreso en Gráfica LOM.

Índice

Introducción

El libro que el lector tiene en sus manos es producto de la experiencia del equipo de científicos que trabaja en el Centro de Investigación de la Web. Hemos querido exponer al gran público no sólo lo que hacemos, sino sobre todo, cómo funciona ese producto tan propio de nuestros tiempos que es la Web. Este libro pretende, a nivel de divulgación, dar a conocer las diferentes facetas que están detrás del funcionamiento de la Web.

Comenzamos con la cuestión básica: ¿Qué es la Web? En el capítulo 1 el profesor Claudio Gutiérrez desarrolla una respuesta a esta pregunta partiendo de las ideas de los creadores de la Web, quienes pensaron la Web como un gigantesco espacio de información universal, una suerte de biblioteca infinita.

Las preguntas inmediatas que siguen a la anterior son: ¿Qué estructura ha tomado este espacio? ¿Cómo está organizado? ¿Cómo es usado hoy por la gente? Esta y otras preguntas, junto con el análisis de la Web chilena, las responde el profesor Ricardo Baeza en el capítulo 2.

La Web es un espacio lógico que está construido sobre un soporte esencial: la red de comunicaciones conocida como Internet. El lector probablemente habrá escuchado usar estas dos palabras en forma intercambiable. En el capítulo 3, el profesor José Miguel Piquer expone la evolución y desarrollo de Internet. Esta gigantesca red mundial de comunicaciones tiene protocolos particulares y... ¡un gobierno!

Ya familiarizados con Internet, la Web y sus estructuras, pasamos a ocuparnos de uno de los temas fundamentales al hablar de la Web: ¿cómo buscar en ella? El profesor Gonzalo Navarro en el capítulo 4 nos explica cómo es posible que un "buscador" encuentre y categorice la información dispersa en casi infinitos sitios en la Web. Y de paso nos da algunas indicaciones de cómo presentar esta información al usuario. En la misma línea, el profesor Mauricio Marín en el capítulo 5 nos desvela las estructuras computacionales necesarias para desarrollar estas búsquedas, a saber, los clusters de computadores.

Íntimamente ligada con la búsqueda de información está la estructura que la información debe poseer. El profesor Marcelo Arenas nos introduce en el capítulo 6 en el moderno lenguaje para representar información: el estándar conocido como XML. Adentrándose algo en detalles técnicos, nos explica qué es este formato, cómo se usa, y cómo se extrae información de él.

Pero no sólo de texto vive el humano. Es cada vez más común manejar otros tipos de información. La profesora Andrea Rodríguez nos explica en el capítulo 7 cómo se presenta la información geográfica en la Web y cómo se la trata actualmente. En el capítulo 8, el profesor Javier Ruiz del Solar nos introduce el mundo de la información multimedial en la Web, un fenómeno de crucial importancia actualmente.

Por último, los capítulos 9 y 10 están dedicados a fenómenos recientes en el desarrollo de la Web. El investigador Javier Velasco nos presenta el tema de las redes sociales, cómo éstas han permeado la Web y se han transformado en una de sus aplicaciones estrella. Por otro lado, el profesor Carlos Hurtado nos muestra la "Web viva", es decir, aquella que cambia continuamente, donde juega un rol importante la suscripción a contenidos, el filtrado automático y el etiquetado de contenidos.

Esperamos haber cubierto los principales aspectos técnicos de este gran fenómeno que es la Web. Hemos intentado mantenernos en un lenguaje no técnico, aunque preciso. Para aquellos que quieren seguir informándose, conociendo y estudiando estos temas, hemos incluido al final de cada capítulo una bibliografía sobre cada tema.

Agradecemos a la Iniciativa Científica Milenio la posibilidad de poder llegar a un público más amplio que el que habitualmente tratamos (nuestros alumnos, colegas investigadores) y ojalá algún joven motivado por alguna de las ideas aquí presentadas se incline por investigar sobre la Web. Con ello habremos colmado nuestras expectativas.

Gonzalo Navarro
Director del Centro de Investigación de la Web
Santiago, Chile, Abril 2008.

Los Autores

Marcelo Arenas es profesor auxiliar del Departamento de Ciencia de la Computación de la Pontificia Universidad Católica de Chile. Obtuvo los grados de Licenciado en Matemáticas (1997), Magíster en Ciencias de la Ingeniería (1998) y el título de Ingeniero Civil de Industrias con Mención en Computación (1998) de la Pontificia Universidad Católica de Chile; y el grado de Doctor en Ciencia de la Computación de la Universidad de Toronto, Canadá (2005). Su investigación ha estado enfocada a distintos aspectos de la Web, tales como el desarrollo de metodologías para mejorar el diseño de las bases de datos XML, el desarrollo de una arquitectura para el intercambio de información XML y la construcción de lenguajes de consulta para la Web Semántica.

Ricardo Baeza-Yates es director de Yahoo! Research Barcelona, España y Yahoo! Research Latin America en Santiago, Chile. Hasta 2005 fue director del Centro de Investigación de la Web del Departamento de Ciencias de la Computación de la Escuela de Ingeniería de la Universidad de Chile, y catedrático ICREA en el Departamento de Tecnología de la Universitat Pompeu Fabra en Barcelona, España. Mantiene vínculos con ambas universidades como profesor jornada parcial. Sus intereses en investigación incluyen algoritmos y estructuras de datos, recuperación de información, minería de la Web, bases de datos de texto e imágenes, y visualización de software y bases de datos.

Claudio Gutiérrez es profesor asociado del Departamento de Ciencias de la Computación de la Universidad de Chile. Obtuvo la Licenciatura en

Matemáticas en la misma universidad, Magíster en Lógica matemática en la P. Universidad Católica de Chile, y Ph.D. en Computer Science en Wesleyan University, EE.UU. Su área de investigación es la lógica aplicada a la computación, bases de datos y Web Semántica. Ha obtenido premios al mejor artículo en conferencias de la Web Semántica los años 2005, 2006 y 2007. Actualmente es investigador asociado del Centro de Investigación de la Web.

Carlos Hurtado es doctor en Ciencias de la Computación de la Universidad de Toronto; Ingeniero Civil y Magíster en Ciencias de la Ingeniería de la Universidad Católica de Chile. Es profesor del Departamento de Ciencias de la Computación de la Universidad de Chile, donde dicta cursos y desarrolla investigación en las áreas de bases de datos, inteligencia artificial y minería de datos. Se ha desempeñado como investigador asociado del Centro de Investigación de la Web, del London Knowledge Lab y del Birkbeck College de la Universidad de Londres. Es socio y miembro del equipo de desarrollo de `orbitando.com`.

Mauricio Marín es investigador en el Centro de Investigación de Yahoo! de Santiago de Chile, e investigador asociado en el Centro de Investigación de la Web de la Universidad de Chile. Anteriormente fue profesor titular de la Universidad de Magallanes, Chile. Obtuvo una Maestría en Ciencias de la Computación en la Universidad de Chile y Doctorado en la Universidad de Oxford, Inglaterra. Sus áreas de interés en investigación son: procesamiento paralelo y distribuido de la información con aplicaciones en máquinas de búsqueda para la Web. Actualmente es Presidente de la Sociedad Chilena de Ciencia de la Computación.

Gonzalo Navarro obtuvo su Doctorado en Ciencias Mención Computación en la Universidad de Chile (1998). Actualmente es profesor titular y director del Departamento de Ciencias de la Computación de la misma Universidad. Ha dirigido diversos proyectos de investigación y hoy es director

del Núcleo Milenio Centro de Investigación de la Web. Sus áreas de interés son algoritmos y estructuras de datos, bases de datos textuales, compresión, y búsqueda aproximada. Es coautor de un libro sobre búsqueda en texto y de más de 200 artículos científicos.

José M. Piquer es profesor asociado del Departamento de Ciencias de la Computación de la Universidad de Chile, y director técnico de NIC Chile. Actualmente dirige el laboratorio de investigación de NIC Chile (NIC-labs), donde se desarrollan proyectos de cooperación con la industria (Entel PCS y SixLabs) sobre redes avanzadas como multimedia móvil, IPv6, IMS y redes de sensores. Obtuvo un Magíster en Ciencias, mención Computación en la Universidad de Chile (1986), y un Doctorado en Computación en l'École Polytechnique de París (1991). Es autor de más de 30 publicaciones internacionales.

M. Andrea Rodríguez Tastets tiene un Master (1997) y un Ph.D. (2000) en Ingeniería y Ciencias de la Información Espacial de la Universidad de Maine, EE.UU. Actualmente es profesora asociada en el Departamento de Ingeniería Informática y Ciencias de la Computación de la Universidad de Concepción e investigadora asociada en el Centro de Investigación de la Web de la Universidad de Chile. Andrea ha realizado trabajos en el área de recuperación de información basada en contenido geo espacial, acceso e indexación de información espacio temporal e integración semántica de datos heterogéneos.

Javier Ruiz-del-Solar es profesor asociado del Departamento de Ingeniería Eléctrica de la Universidad de Chile. Obtuvo el título de Ingeniero Civil Electrónico y el grado de Magíster en Ingeniería Electrónica de la Universidad Técnica Federico Santa María, y el grado de Doctor en Ingeniería de la Universidad Técnica de Berlín, Alemania. Sus áreas de investigación incluyen visión computacional, robótica móvil y búsqueda automatizada de

información multimedial en la Web. Ha obtenido premios al mejor artículo y a la innovación en los eventos de robótica móvil RoboCup 2004 y 2007. Actualmente es investigador asociado del Centro de Investigación de la Web, director del Laboratorio de Robótica de la Universidad de Chile y conferencista distinguido de la Sociedad de Robótica y Automatización del IEEE.

Javier Velasco, comunicador social, es uno de los pioneros en el campo de la Arquitectura de Información en Chile desde 2000. Ha trabajado en importantes proyectos Web en Chile y los Estados Unidos. También ha sido profesor adjunto en la Universidad de Maine, USA, y editor administrativo en la revista Boxes and Arrows. Desde 2003 integra parte del equipo CIW, donde ofrece cursos y consultorías en esta materia, y desde 2006 forma parte del laboratorio de Yahoo! Research en Santiago. Su trabajo se enfoca en el diseño de experiencia de usuario en sistemas de información, lo que incluye arquitectura de información, usabilidad, diseño de interacción, diseño de interfaces, diseño de información y estrategia en proyectos Web.

Capítulo 1
La Web como espacio de información universal

Claudio Gutiérrez

> Todo estaría en sus ciegos volúmenes. Todo: la historia minuciosa del porvenir, *Los egipcios* de Esquilo, el número preciso de veces que las aguas del Ganges han reflejado el vuelo de un halcón, el secreto y verdadero nombre de Roma, la enciclopedia que hubiera edificado Novalis, mis sueños y entresueños en el alba del catorce de agosto de 1934, la demostración del teorema de Pierre Fermat, los no escritos capítulos de *Edwin Drood*, esos mismos capítulos traducidos al idioma que hablaron los garamantas, las paradojas de Berkeley acerca del tiempo y que no publicó, los libros de hierro de Urizen, las prematuras epifanías de Stephen Dedalus que antes de un ciclo de mil años nada querrían decir, el evangelio gnóstico de Basílides, el cantar que cantaron las sirenas, el catálogo fiel de la Biblioteca, la demostración de la falacia de ese catálogo. Todo, ...
>
> J. L. Borges, *La Biblioteca Total*.

El sueño de la biblioteca infinita se ha hecho realidad: la Web hoy contiene lo que soñó Borges y bastante más. De hecho, se estima que la pieza promedio de información en la Web hoy día nunca será vista más que por su productor y sus amigos cercanos, y uno no puede ver más que un porcentaje minimal de lo que está publicado.

¿Cómo se logró esta fantástica biblioteca infinita? En este breve capítulo revisaremos los fundamentos conceptuales y técnicos que están en la base de la Web, y discutiremos sus alcances y limitaciones.

Es común que los términos *Web, Red* e *Internet* se usen intercambiablemente. Desde el punto de vista técnico son objetos completamente diferentes. Internet hace referencia a la red física que conecta diferentes computadores y lugares. Sus preocupaciones son protocolos de transmisión de datos (TCP IP), manejo de nombres de dominio, etc. y que lo tratamos en detalle en el capítulo 3. La Web hace referencia a la arquitectura lógica de la información que ha sido posible construir sobre esa red física. Confundirlos es como confundir el cerebro (una red neuronal) con el conocimiento que posee una persona. Todos tenemos casi el mismo material cerebral, pero los conocimientos y la información que cada uno posee difieren vastamente.

De fuentes aisladas a redes de información

La evolución del procesamiento de información ha ido desde unidades aisladas hasta una interconexión mundial hoy día a través de la Web.

Probablemente la mejor metáfora sea de nuevo la de una biblioteca. Allí hay información restringida al lugar físico donde funciona. Por un momento olvidemos los catálogos globales (¡productos de la Web también!), y pense-

mos cómo hace 50 años alguien buscaba información. Debía recorrer biblioteca por biblioteca, y correlacionar o comparar la información a mano. Por ejemplo, determinar los títulos de libros que estudian la vida de Andrés Bello. No podía "navegar" a través de la imagen virtual de todos los libros de todas las bibliotecas del mundo juntas. Sin embargo, la Web hizo posible esa realidad.

El desarrollo de la tecnología computacional ha sido clave en este proceso. Los computadores en sus inicios eran gigantescos armatostes que ocupaban pisos enteros de edificios, "centros" de procesamiento de información. La gente, técnicos, usuarios, etc. giraba en torno a ellos. La conexión entre dos de estos gigantescos aparatos era escasa o nula. Con el advenimiento de los computadores personales, llegó también la idea de que cada usuario poseedor de un PC pudiera "conectarse" con otros cercanos. Nacieron las redes locales. De esta idea hay un paso a pensar una red más grande, y finalmente una red "global". Y con esto, aparece el problema de cómo coordinar, integrar la información que está en cada uno de los nodos (computadores) de esta gigantesca red.

A comienzos de los noventa, Tim Berners-Lee [1] tuvo una idea genial: diseñar este sistema global de información de tal forma que cada usuario en un nodo pudiera navegar por el resto de forma totalmente automática, es decir, sin tener idea de cómo funciona el sistema del otro, qué sistema operativo tiene, qué lenguajes de programación usa, qué aplicaciones corre. Su experiencia en el CERN (ver figura 1.1) fue la gatilladora de esta simple idea, que es el origen de la Web. En palabras de Berners-Lee: "El concepto de la Web integró muchos sistemas de información diferentes, por medio de la formación de un espacio imaginario abstracto en el cual las diferencias entre ellos no existían. La Web tenía que incluir toda la información de cualquier tipo en cualquier sistema."

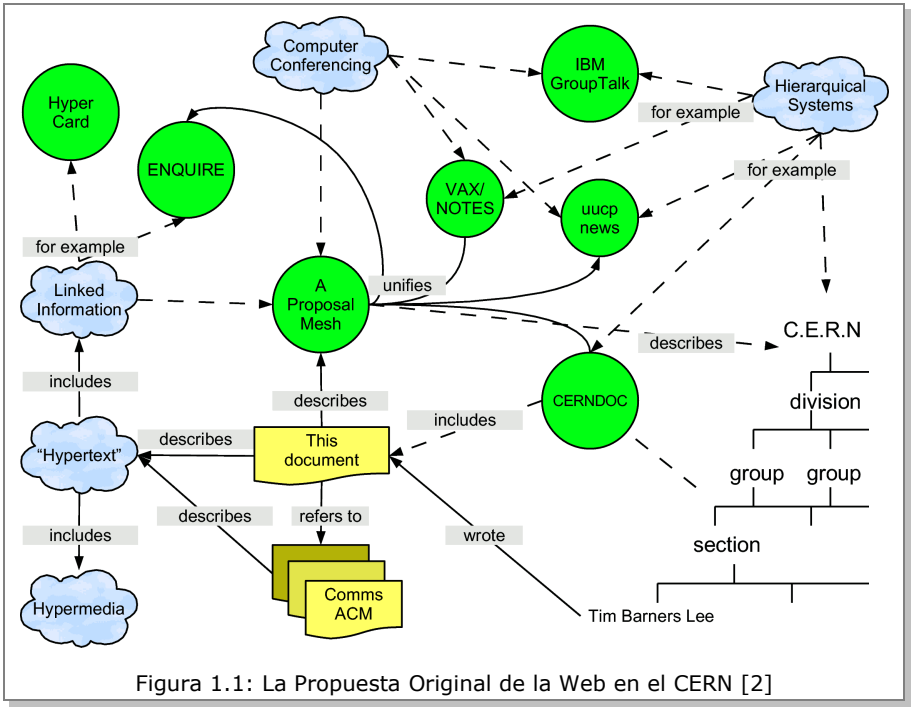


Figura 1.1: La Propuesta Original de la Web en el CERN [2]

Es así como la Web es hoy un gran espacio de información universal, una vitrina de acceso a casi -excluimos la de organizaciones como el Pentágono, etc.- toda la información existente en el mundo hoy en día.

Las bases lógicas de la Web

Desde el punto de vista técnico, los tres pilares básicos sobre los que se sustenta la arquitectura lógica de la Web son:

1. *Identificadores únicos (URI)*: en un mundo ideal, la suposición básica necesaria para poder referirse (referenciar) y hablar de (describir) todos los objetos, es que éstos tengan su nombre propio, que en términos técnicos se llama *identificador*. En la Web estos nombres propios se llaman *Identificadores Universales de Recursos* (URI por sus siglas inglesas).

Una versión más elemental de URI es la URL (*Localizador* universal de recursos), que corresponde a una *dirección* en la Web. La dirección es una de las formas de identificar un objeto, pero es bueno señalar que la noción de identificador es más amplia que la de dirección, por ejemplo para recursos móviles que no tienen dirección fija.

2. *Lenguaje universal para describir HTML:* Otra suposición básica para la comunicación universal es un lenguaje único, entendible por todos. Tim Berners-Lee diseñó el lenguaje HTML (siglas del inglés *Hyper Text Markup Language*, que a sus características de simplicidad de uso, suma una característica clave: el ser un lenguaje de *hipertexto*, es decir, que tiene un forma de anclar o redirigir al lector desde un punto cualquiera del texto a otro lugar. Estos son los famosos *links* o enlaces en la Web.

3. *Protocolo de transmisión de datos HTTP:* Desde un punto de vista más técnico, uno necesita un protocolo que permita *enviar* y *traer* información en HTML desde un lugar (sitio) a otro en esta gigantesca red que es la Web.

El protocolo HTTP (sigla del inglés Hyper Text Transfer Protocol) tiene varias características distintivas que lo han hecho muy perdurable. HTTP es un protocolo de transmisión entre clientes y servidores. El cliente, que puede ser un browser, un agente, o cual-

quier herramienta. El servidor es el que almacena o crea recursos como archivos HTML, imágenes, etc. Entre ellos puede haber varios intermediarios, como proxies, gateways y túneles. A través de instrucciones simples, pero poderosas, el cliente indica al servidor qué acciones realizar para recibir o entregar datos. Ver más detalles en capítulo 3.

La filosofía de la Web como espacio de información: la W3C

La Web fue creada con una cierta filosofía, una posición de principios frente a los desarrollos que se venían dando en materia de publicaciones, de desarrollo de software, de derechos de autor y de difusión. Esta filosofía puede resumirse en tres principios básicos: *todos pueden publicar*, *todos pueden leer*, *nadie debe restringir*.

¿Cómo lograr esto técnicamente? En esta dirección, se creó el Consorcio de la Web (W3C), una organización internacional que se propuso como sus dos objetivos primordiales el impulsar la *interoperabilidad* y *evolutividad* de la recientemente creada red universal de información. Para esto se comenzaron a generar estándares y protocolos. ¿Qué significan estos dos requerimientos en más detalle? En un famoso artículo, *Explorando la Universalidad* [3], Tim Berners-Lee desglosaba sus aspectos básicos:

- *Independencia de Dispositivo.* La misma información debe ser accesible desde diversos dispositivos. Esto significa, por ejemplo, que la visualización debe tener estándares que permitan acceder a la información desde casi cualquier formato de pantalla y audio. Una de

las bases para implementar esta *desiderata* es la separación de contenido y forma en la información.

- *Independencia de Software.* Hay muchos y diversos programas de software que se usan. Ninguno debe ser crítico para el funcionamiento de la Web. El desarrollo descentralizado del software ha sido clave para su crecimiento. Además, tema no menor, este postulado previene que la Web misma caiga bajo el control de una comunidad dada o algún gobierno usando el control del software.

- *Internacionalización.* Desde sus inicios, la Web no ha estado cargada a ningún país. Con la introducción de UNICODE, la última barrera que cargaba su desarrollo hacia los lenguajes occidentales ha sido barrida. (La diferencia clave entre el viejo HTML y el nuevo estándar XHTML, aparte de mejoras técnicas relacionadas con XML, es que XHTML está basado en UNICODE.)

- *Multimedia.* Los formatos disponibles para publicar deben estar abiertos a todas las facetas de la creatividad humana capaces de representar. En este sentido, soportar multimedia no representa sólo un par de avances tecnológicos, sino una filosofía de desarrollo de la Web.

- *Accesibilidad.* La gente difiere en múltiples cosas, en particular, en sus capacidades. La universalidad de la Web debe permitir que ella sea usada por la gente independientemente de sus discapacidades. De nuevo aquí la separación de contenido y forma de la información es un pilar básico.

- *Ritmo y razón.* Como dice TBL, la información varía desde un poema hasta una tabla en una base de datos. El balance entre procesamiento automático y humano debe estar presente. Por un lado, por

las cantidades y tipo de información actualmente disponible es impensable que ésta sea procesada sólo por seres humanos: se necesitan agentes automáticos. Por otra parte, es absurdo pensar que en algún momento los humanos serán prescindibles en el desarrollo y enriquecimiento de la Web. Hay que buscar los justos términos para cada aplicación.

- *Calidad.* Las nociones de calidad son subjetivas e históricas. Por ello es impensable que algún día *toda* la información vaya a ser de calidad. Aquí hay otro compromiso, y es que la tecnología de la Web debe permitirnos navegar y vivir entre información con diferentes niveles de calidad.

- *Independencia de escala.* La armonía a gran escala supone armonía en sus componentes. La Web debe soportar grandes y pequeños grupos. Debe permitir que la privacidad de la información de individuos y grupos pueda ser negociada por ellos mismos, y permitir que cada grupo se sienta seguro en el control de su espacio. Hay que lograr un balance entre un gigante monolítico y una diversidad que pueda llevar al aislamiento completo de cada uno.

La Web Semántica

Uno de los problemas más importantes que aparece con la Web es el de determinar qué "significa" cada dato que está en la Web. Es prácticamente imposible para un usuario chileno entender una página en chino o tailandés. Y viceversa. El problema es aún más dramático: es muy difícil para un humano encontrar la información que necesita. Los buscadores funcionan de manera puramente "sintáctica", es decir, no "entienden" las palabras. ¿Qué hacer?

Tradicionalmente eso era resuelto por *catalogadores*, personas especializadas que agregaban *metadatos* (etiquetas que explicitan información) a los libros: qué tema trata, dónde está ubicado, cuál es el autor, etc. Estos metadatos están accesibles en un catálogo en las bibliotecas. En la Web, como ya veíamos, no tenemos catálogo, ni menos catalogadores. Con los volúmenes de información que cada día crecen, es imposible que humanos se preocupen de clasificar la información. Además, porque el modelo de la Web es distribuido, quienes publican tienen diversas visiones sobre cómo clasificar sus objetos.

Para los profesionales de la información, el principal desafío hoy es cómo manejar esta extraordinaria cantidad de datos que crece día a día. Estamos comenzando a ver los problemas: los motores de búsqueda a menudo no contestan lo que buscamos; hay dificultades para filtrar la información; la heterogeneidad de los datos y los contenidos; desde el punto de vista de quien publica, se ha convertido en un problema hacer visible la visible, tanto en formato como en contenido. Han habido avances en los niveles estructurales y sintácticos con el estándar XML y sus tecnologías aledañas. Desafortunadamente, al nivel del significado (semántica) aún estamos muy por debajo de las necesidades. Estamos lejos de responder preguntas como "todos los museos que exhiban trabajos de Guayasamín" o "¿Cuál es la biblioteca que tiene la mejor colección de los escritos de Gandhi?" o "¿Cuál es la compañía que ofrece el mejor mapa de Isla de Pascua desde el punto de vista precio/resolución?" Un motor de búsqueda estándar (como Google, Yahoo!, etc.) no puede responder tales consultas. Pero tampoco ningún agente las podría responder hoy en día. Sin embargo, la información está allí: hay que relacionarla y agregarla. La limitación obedece a la falta de habilidad de las máquinas para entender el significado y las relaciones entre las partes de información que recolectan. Hoy en día los humanos agregamos el contexto, interpretamos y damos sentido a la información que existe

en la Web. En otra dirección, otro ejemplo de estas limitaciones es la dificultad para diseñar e implementar una tarea tan natural como organizar todos los recursos educacionales de un país, de tal forma que resulte sencillo para cada estudiante y profesor el publicar y obtener la información que requieran. Se necesitan vocabularios comunes, descripción precisa de los datos expuestos, publicación distribuida, búsquedas automatizadas. En una frase: debido a las enormes dimensiones, la Web se ha convertido en una torre de Babel no sólo al nivel del lenguaje natural, sino esencialmente al nivel del significado, contradiciendo las ideas por las cuales fue creada. ¿La solución? Pavimentar el camino para la construcción de agentes de software que puedan procesar información de la Web por nosotros. La noción de *Web Semántica* [4] es transformar la Web actual de tal forma que la información y los servicios sean entendibles y usables tanto por computadores como por humanos. La Web Semántica creará el ambiente necesario donde los agentes de software puedan rápidamente realizar tareas sofisticadas y ayudar a los humanos a encontrar, entender, integrar, y usar la información en la Web.

Metadatos y RDF

La característica distintiva de la Web Semántica será un *lenguaje estándar de metadatos* y *ontologías*, que permitirán que agentes de software encuentren el significado de la información en páginas Web, siguiendo enlaces a las definiciones de términos claves y reglas para razonar acerca de ellas lógicamente. Los *metadatos* son datos descriptivos acerca de un objeto o recurso, sea éste físico o electrónico. Las *ontologías* son especificaciones formales de vocabulario y conceptos compartidos para un dominio.

Aunque el concepto de metadatos es relativamente nuevo, los conceptos subyacentes han estado rondando desde que se organizaron grandes colecciones de información. En áreas tales como catalogación en bibliotecas y

museos han sido usados por décadas, por ejemplo, el DCC (Dewey Decimal Classification), OCLC (On Line Computer Library Center), Dublin Core. Una manera útil de pensar acerca de los metadatos es "la suma total de lo que uno puede decir acerca de cualquier objeto de información a cualquier nivel de agregación". Hay muchos tipos de metadatos, y los usos más comunes se refieren a documentación de *copyrights* y accesos legales, versionamiento, ubicación de información, indización, descripción de condiciones físicas de recursos, documentación de software, autentificación, etc.

En la Web, los metadatos también han jugado un rol importante en áreas como catálogos de propósito general (Dublin Core, Open Directory Project, Wikipedia), sindicación y *rating* (Rich Site Summary RSS, Platform for Internet Content PICS), colecciones personales (música, fotos), privacidad, etc. Y los más populares hoy son simplemente *tags*, es decir, etiquetas; un lenguaje que no tiene verbos ni adjetivos. Simplemente nombres. Todos estos metadatos son sectoriales y usan una diversidad de modelos y lenguajes.

Por el contrario, se necesita un lenguaje de significados (de metadatos) universal. Este es RDF [5] (del inglés *Resource Description Framework*), que es un lenguaje diseñado para soportar la Web Semántica, de la misma manera que HTML es el lenguaje que ayudó a iniciar la Web. El modelo de RDF es simple: el universo a modelar (la Web) es un conjunto de *recursos* (esencialmente todo puede tener una URL); el lenguaje para describirlo es un conjunto de *propiedades* (técnicamente predicados binarios); las descripciones son *oraciones* similares en estructura al modelo sujeto-predicado-objeto, donde el predicado y el objeto son recursos o cadenas de caracteres. Así, por ejemplo, uno puede afirmar "El creador de `http://www.picarte.cl` es Claudio Gutiérrez". El vocabulario de las *propiedades* para este lenguaje pue-

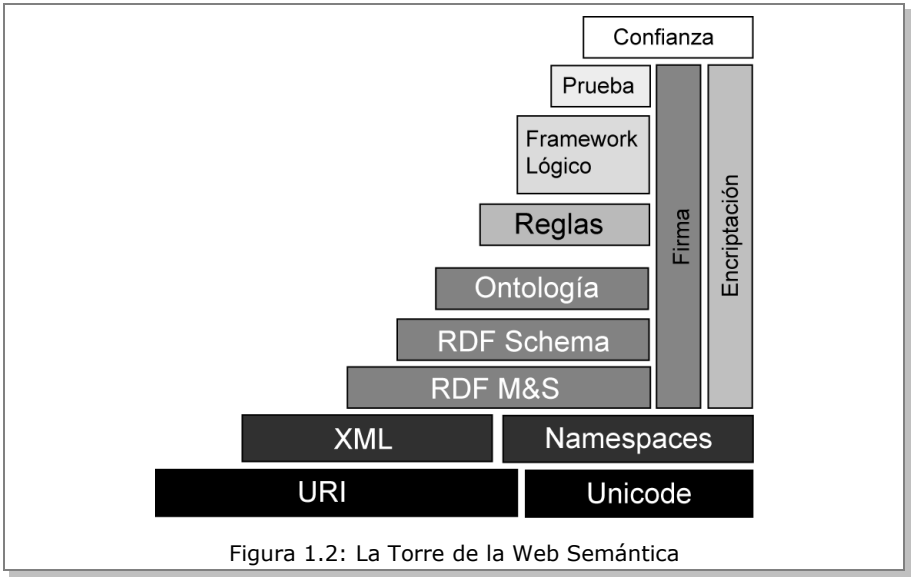


Figura 1.2: La Torre de la Web Semántica

de ser definido siguiendo las líneas dadas en los esquemas RDF (*RDF Schema*), y básicamente son codificaciones de ontologías a diferentes niveles.

El Futuro de la Web

No es fácil predecir los desarrollos futuros de la Web. El proyecto inicial de Tim Berners-Lee incluía el desarrollo de capas sucesivas para permitir el intercambio global de información y conocimiento. Luego de la estructura básica que conocemos, vendrá una capa de *semántica*, de metadatos. Esta capa permitiría procesar la información semi-automáticamente, es decir, permitiría a agentes de software procesar la información en paralelo a los humanos. (Nótese que la Web actual está hecha casi en su totalidad para que seres humanos la naveguen.)

La Web por supuesto ha evolucionado en miles de direcciones, muchas no previstas, como redes sociales, blogs, etc. Muchos han llamado al conjunto de estos desarrollos "novedosos" no previstos *Web 2.0*. En los capítulos siguientes trataremos varias de estas facetas.

El futuro está abierto. Hoy en día no es posible predecir los usos futuros de la Web, y aquí ya entramos al campo de la ciencia ficción.

Para saber más

- Tim Berners-Lee, *Tejiendo la Red*, Siglo Veintuno Eds., España, 2000.
- Tim Berners-Lee, Ora Lassila *La Web Semántica*, Scientific American, 2002.
- La World Wide Web Consortium (W3C) ha dispuesto una breve guía introductoria, en español, sobre la web semántica: http://www.w3c.es/Divulgacion/Guiasbreves/WebSemantica

Referencias

1. "CERN: Where the web was born." Page at the CERN. http://public.web.cern.ch/public/en/About/Web-en.html
2. Tim Berners-Lee. Information Management: A Proposal (1989). http://info.cern.ch/Proposal.html - http://www.w3.org/History/1989/proposal.html
3. Tim Berners-Lee. The World Wide Web - Past Present and Future: Exploring Universality. http://www.w3.org/2002/04/Japan/Lecture.html
4. W3C Semantic Web Activity: http://www.w3.org/2001/sw/
5. Resource Description Framework (RDF) / W3C Semantic Web Activity: http://www.w3.org/RDF/

Capítulo 2
Anatomía de la Web

Ricardo Baeza Yates

Introducción

¿Qué estructura tiene la telaraña mundial de computadores o World Wide Web? (la Web de ahora en adelante, aunque no me queda claro si es femenino o masculino). Nadie sabe. Crece más rápido que la capacidad de ella misma para detectar sus cambios. Sus conexiones son dinámicas y muchas de ellas quedan obsoletas sin ser nunca actualizadas. El contenido de la Web es hoy de miles de terabytes (un terabyte o Tb es un billón de megabytes) de texto, imágenes, audio y video. Para aprovechar esta gran base de datos no estructurada es importante poder buscar información en ella, adaptándose al crecimiento continuo de la Web.

Al igual que Internet, la red de computadores que interconecta el globo, que ya sobrepasó los 430 millones de computadores conectados en más de 220 países durante 2006, los servidores Web también crecen en forma exponencial desde 1993 (un servidor Web es el software que administra un sitio Web). Lamentablemente nadie sabe su número exacto, pues no es posible a partir de un nombre de dominio saber si es o no un servidor Web (la mayoría comienza con www, pero muchos lugares no siguen esta convención). Además un mismo computador puede manejar distintos servidores y también existen servidores virtuales (un mismo servidor Web puede manejar

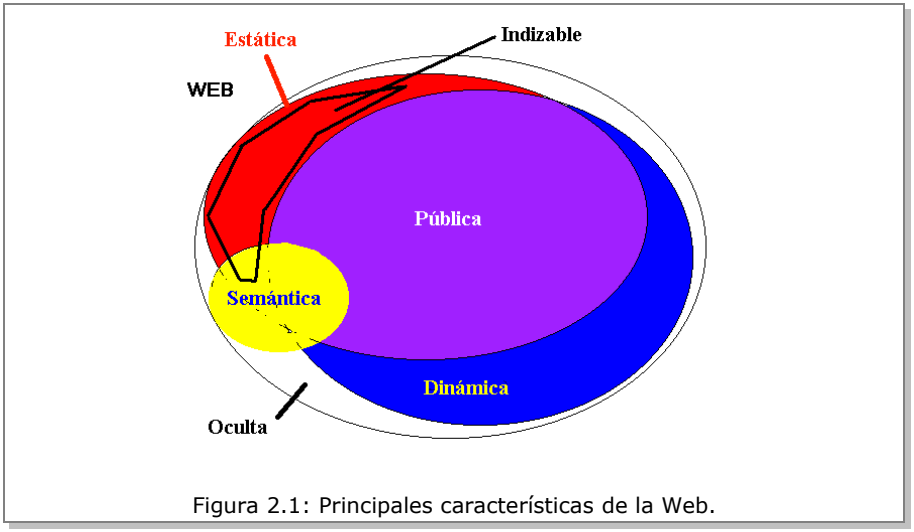


Figura 2.1: Principales características de la Web.

lógicamente otros servidores). En el año 2000, el número de servidores sobrepasó los 10 millones y en mayo de 2007 ya llegaban a los 120 millones.

Conceptos Básicos

La Web es compleja: hay páginas estáticas y dinámicas, públicas y privadas, con o sin metadatos, que representan la semántica de la Web, tal como se muestra en la Figura 2.1.

Las páginas *estáticas* son aquellas que existen todo el tiempo en un archivo en algún servidor Web. Las páginas *dinámicas* son aquellas que se crean cuando una persona interactúa con un servidor Web, por ejemplo la respuesta a una consulta en un buscador o el resultado de rellenar un formulario en un sitio de comercio electrónico. Actualmente, la mayor parte de la Web es dinámica, y como en algunos sitios se puede generar un número no

acotado de páginas dinámicas (por ejemplo, un calendario), la Web que podemos crear es infinita.

Las páginas *públicas* son las que todas las personas pueden ver y las *privadas* son las que están protegidas por una clave o se encuentran dentro de una Intranet. Como cada persona tiene acceso a distintas páginas privadas, la Web pública depende del observador. En particular cada buscador refleja una Web pública distinta. Algunos sitios tienen información semántica que ayuda a los buscadores y se estima que un 5% de ellos tiene información fidedigna. Sin embargo, más son los sitios que tienen información falsa, lo que se llama *spam de Web*.

Minería Web

Para caracterizar la Web debemos realizar un proceso de minería de datos de la Web, lo que en inglés se llama *Web mining*. Una metáfora sería excavar la Web y es posible hacerlo en distintas partes de ella: en su contenido, en su estructura y en su uso. El contenido y la estructura se recolectan con un software que recorre las páginas de la Web y siguen sus enlaces, un programa que en inglés se llama *crawler*. El uso se obtiene de la información que dejan las personas al usar un sitio Web, que se almacena en una bitácora. A continuación detallamos brevemente cada uno de estos casos.

Excavando el Contenido

Lo más simple es recuperar información a través de buscadores como Google o Yahoo!. Pero es posible también usar análisis de lenguaje natural para entender parcialmente la semántica del texto, extraer otros objetos como imágenes o audio, aprovechar las marcas de HTML para transformar el contenido o extraer datos específicos, o mejorar los resultados de los bus-

cadores agrupando páginas similares. Uno de los problemas principales es cómo encontrar las páginas que poseen el contenido que necesitamos, pues sólo encontrar todas las páginas que son indexables ya es difícil (ver figura 2.1).

Desenredando la Estructura

La estructura de la Web es compleja y evoluciona en el tiempo. Hay desde sectores altamente conectados hasta islas que sólo conocen algunos buscadores. La estructura puede ser usada por los buscadores para jerarquizar los resultados (en base a las páginas más referenciadas usando heurísticas como Pagerank) o para encontrar grupos de páginas que se apuntan entre sí y representan comunidades de personas con intereses similares. El problema principal en este caso es entender el proceso de evolución y su relación con las personas que participan en él.

Analizando el Uso

Analizar las bitácoras de acceso (*logs*) a un sitio Web es lo más interesante desde el punto de vista comercial. Por ejemplo, una página que nunca es visitada tal vez no tiene razón de ser, o si páginas muy visitadas no están en los primeros niveles, esto sugiere mejorar la organización y navegación del sitio. Por lo tanto, es importante detectar patrones de acceso y sus tendencias. Esta detección puede ser genérica o para un usuario específico (lo que permite personalizar sitios en forma dinámica) y los resultados pueden ser usados para recomendar servicios o productos. El problema principal en este caso es poder diferenciar a los usuarios y cuándo se conectan o desconectan (determinar sesiones).

El Principio del Mínimo Esfuerzo

George Kipling Zipf era un lingüista de Harvard y publicó en 1949 su libro sobre el principio del mínimo esfuerzo un año antes de su deceso, a la prematura edad de 40 años. El descubrimiento inicial de Zipf en 1932 fue que si uno contaba el número de veces que se usaba cada palabra en distintos textos en inglés, y las ordenaba de más frecuente a menos frecuente, se cumplía que la frecuencia de la palabra *i*-ésima, multiplicada por *i*, era igual a una constante *C*, y la constante *C* dependía del texto escogido. Actualmente es necesario elevar *i* a un exponente *t* mayor que 1 y cercano a 2 para muchos textos existentes, en particular en la Web. Graficando esta curva usando una escala logarítmica en ambos ejes, ella se convierte en una recta con pendiente negativa *t* [1].

Zipf explica estos resultados empíricos como una condición humana, donde siempre es más fácil escribir una palabra conocida que una menos conocida. Fenómenos similares aparecen en otros ámbitos como el número de citas bibliográficas a un artículo dado o las poblaciones de las ciudades. Diversos autores, entre ellos Mandelbrot y Miller, argumentaron más tarde que en realidad la ley de Zipf representa la consecuencia de las leyes de las probabilidades en procesos asociados a codificación de información donde hay mucho de azar. Sin querer tomar partido en esta disputa científica, cierta o no cierta, la ley de Zipf aparece frecuentemente en la práctica y refleja bien la actitud natural de minimizar el esfuerzo, exceptuando los casos extremos, que serían en el ejemplo inicial, usar muy pocas palabras o usar muchas. Tal vez esta ley sólo explica la diversidad humana, la que se inclina más por la pereza que por la erudición. De hecho, que *t* sea ahora alrededor de 1.8 para textos en inglés, indica un mayor sesgo en esa diversidad, y una degradación en el tiempo de la riqueza del vocabulario que usamos al escribir.

La Web como un Proceso Humano

La Web es el producto del trabajo colaborativo de millones de personas. Si hay algún fenómeno donde el principio del mínimo esfuerzo aparecería si existiera, es la Web. Aparte de la distribución de palabras en la Web, las siguientes medidas siguen una curva de Zipf:

- Tamaños de las páginas o de otros tipos de archivos (imágenes, audio, etc.). En este caso la ley no se ajusta bien al comienzo, porque hacer páginas con muy poco texto produce el pudor de la vergüenza que contrarresta al mínimo esfuerzo.
- Número de enlaces que salen de una página. En este caso la curva no se ajusta muy bien en los extremos, porque hacer una página con muy pocos enlaces cae en el caso del punto anterior y, por otra parte, hay páginas con muchos enlaces producidas en forma automática.
- Número de enlaces que llegan a una página. La mayoría de las páginas tienen sólo un enlace a ellas y hay pocas páginas con muchos enlaces.
- Fecha de actualización de las páginas, existen más páginas nuevas o modificadas que viejas.
- Número de componentes conexos de distinto tamaño. Es decir, grupos de páginas en las que se puede navegar de cualquier página a otra página. Esto representa en cierta medida el número de páginas de un sitio Web: muchos sitios tienen pocas páginas, pocos sitios muchas páginas.
- Uso de las palabras en las consultas a un buscador (confirmado experimentalmente en TodoCL.cl). El resultado es que la mayoría de las preguntas son muy simples.

Lo anterior se propaga a otras medidas, como tráfico en la red, uso de proxies, etc. ¿Es todo esto una casualidad producto del azar o un fenómeno del comportamiento humano?. Sin duda la respuesta es que esta ley es resultado del proceso humano de creación de la Web.

Caracterizando la Web

Estructura y Visibilidad

¿Cuántas referencias tiene una página HTML? (HTML es un acrónimo para Hyper Text Markup Language; el lenguaje usado para estructurar páginas Web). Más del 75% de las páginas tiene al menos una referencia, y en promedio cada una tiene entre 5 y 15 referencias. La mayoría de estas referencias son a páginas en el mismo servidor. De hecho, la conectividad entre sitios distintos no es muy buena. En particular, la mayoría de las páginas no son referenciadas por nadie y las que sí son referenciadas, lo son por páginas en el mismo servidor.

Considerando sólo referencias externas (entre sitios distintos), más del 80% de las páginas tienen menos de 10 referencias a ella. Otros sitios son muy populares, teniendo decenas de miles de referencias a ellos. Si contamos sitios que referencian a sitios, aparecen ODP (www.dmoz.org), el directorio abierto, y el directorio de Yahoo! en los dos primeros lugares. Estos sitios son los que conectan la Web. Por otro lado, hay algunos sitios que no son referenciados por nadie (están porque fueron incluidos mediante el envío directo de una dirección Web a Yahoo! u otros buscadores, pero que realmente son islas dentro de la Web). En este mismo sentido, las páginas personales también se pueden considerar como entes aislados en muchos casos. Asimismo, la mayoría de los sitios (80%) no tiene ninguna referencia

hacia páginas en otros servidores. Esto significa que una minoría de los servidores mantiene toda la carga navegacional de la red. Estadísticas recientes indican que el 1% de los servidores contienen aproximadamente el 50% del volumen de datos de la Web, que se estimaba mayor a 20,000 millones de páginas durante 2006.

Tamaños y características

¿Cómo es una página Web promedio? Una página de HTML promedio tiene alrededor de 5 a 7 kilobytes (alrededor de mil palabras). Si agregamos audio o video, este promedio aumenta. De hecho, la distribución de tamaños sigue una distribución de Zipf. En otras palabras, aunque la mayoría de los archivos son pequeños, existe un número no despreciable de archivos grandes; y hasta 50 kilobytes predomina el volumen de las imágenes. Desde allí hasta 300 kilobytes son importantes los archivos de audio. Más allá de este límite, llegando a varias decenas de megabytes, tenemos archivos de video. Los formatos más populares (en base a la extensión del nombre de archivo) son HTML, GIF, TXT, PDF, PS y JPG, entre otros.

¿Cómo es una página HTML? Alrededor de la mitad de ellas no tiene ninguna imagen. Un 30% no tiene más de dos imágenes y su tamaño promedio es de 14Kb. Por otra parte hay un porcentaje no despreciable (mayor al 10%) de páginas con más de 10 imágenes. La razón es que son imágenes tipográficas, como por ejemplo puntos rojos, líneas de separación de color, etc. La mayoría de las páginas usan HTML simple. Sólo un porcentaje pequeño sigue todas las normas y otro porcentaje mayor (alrededor del 10%) es sólo texto. Finalmente, la calidad del texto deja mucho que desear, pues hay errores de tipeo, errores que viene de la conversión de imágenes de documentos a texto, etc. Más aún, la información contenida puede estar obsoleta,

puede ser falsa o engañosa. Hay que tener esto en mente cuando usamos una página Web como fuente de información o la referenciamos.

Los Sitios Impenetrables

Estos sitios son aquellos que contienen una o más páginas donde un buscador no puede extraer los enlaces a las páginas internas porque no usan HTML sino un diseño gráfico basado en un programa. Es decir, la estética es prioritaria pero por ignorancia mata su contenido. Según el último estudio de la Web Chilena [4], estos son el 21% de los sitios, es decir más de 25 mil sitios. Esto incluye sitios que usan Flash en su portada, otros que son o hacen una llamada a un programa y unos pocos que usan mapas de imágenes anti-cuados. Muchos de estos sitios tienen una portada impenetrable de más de 100Kbs de código, sin contar imágenes, así que además son poco visibles, pues en un módem normal tardarían al menos 30 segundos en cargarse.

Uno puede perdonar que la mayoría de las empresas chilenas no sepan que Flash o Javascript mal usado convierte sus sitios en bóvedas de seguridad. Sin embargo, hay casos en que esto es imperdonable:

- La empresas de tecnologías de la información no pueden apelar a la excusa de ser ignorantes.
- Los sitios de gobierno deben ser los más públicos, visibles[1] y fáciles de encontrar de la Web.
- Las empresas donde la información es uno de sus valores fundamentales. ¡Y sin mencionar las empresas de este tipo que no tienen sitio Web!

1 Por ejemplo, cuando su portada hace difícil la navegación.

La Web como un Grafo

Imaginemos que por cada persona que conocemos existe una conexión directa entre ella y sus amigos. Por ejemplo, un número telefónico. Si hacemos esto para todas las personas del mundo, tenemos un *grafo* (como los de la Figura 2.2) muy grande. En ese grafo podemos ahora medir "distancias" entre dos personas usando el número mínimo de llamadas telefónicas que necesita una persona para contactar con otra. Por ejemplo, si la persona que quiero contactar está en China es posible que si yo conozco una persona que conoce a una persona en China, el número de llamadas sea pequeño (en el mejor caso, sólo tres llamadas). La distancia máxima entre dos personas se llama el *diámetro del grafo*, usando una analogía geométrica. A mediados de los sesenta, Milgram realizó un famoso experimento utilizando paquetes de correo y estimó que el diámetro dentro de Estados Unidos era 6.

Para que un grafo tenga un diámetro pequeño debe tener muchas conexiones. Si todas las conexiones existen, el diámetro es 1. Por otra parte, un grafo aleatorio tiene un diámetro mucho mayor. Un modelo de grafo que representa bien este fenómeno es aquel en el que cada persona está conectada con todas las personas cercanas (geográficamente) y sólo con algunas lejanas de manera aleatoria y con una distribución de probabilidad uniforme. Este modelo se llama *small-world* o mundo pequeño, valga la redundancia, y también representa bien la red neuronal de un gusano y la red eléctrica del oeste de Estados Unidos, entre otros casos [2].

A finales de los 90, Albert, Jeong y Barabási midieron la distancia (número mínimo de enlaces para llegar de una página a otras) entre 330 mil páginas de la Web [5]. Con esto aproximaron el diámetro con una función logarítmica en el número de páginas. Al extrapolar esta función, considerando que el número de páginas Web es de más de mil millones de páginas, obtu-

vieron que el diámetro de la Web es aproximadamente 19. Es decir, con 19 clicks del ratón llegamos a cualquier página Web del planeta. Ellos y otros autores sugieren que un buscador podría aprovechar esto para encontrar rápidamente la página deseada. Sin embargo, esto significa saber qué enlace seguir, un problema que no es trivial.

Aunque el modelo de mundo pequeño podría ser válido en la Web, este modelo no explica cómo una persona que sólo tiene conocimiento local puede saber a quién contactar para encontrar a otra persona. Recientemente, Kleinberg [6] ha modificado el modelo original, de tal modo que las conexiones lejanas no siguen una distribución uniforme, sino que una que es inversamente proporcional al cuadrado de la distancia. Esta distribución es óptima en el sentido que minimiza el número promedio de llamadas que haría una persona para contactar a otra, y explica lo que ocurre en la práctica.

La Web es más que un simple conjunto de documentos en distintos servidores, ya que existen relaciones de información entre los documentos mediante los enlaces que establecen entre ellos. Esto presenta muchas ventajas, tanto para los usuarios, a la hora de buscar información, como para los programas que recorren la Web a la hora de buscar contenido para recolectar (probablemente para un motor de búsqueda). Debido a esto se plantea la Web como un modelo de grafo dirigido, en el que cada página es un nodo y cada arco representa un enlace entre dos páginas.

En general las páginas enlazan a páginas similares, de modo que es posible reconocer páginas mejores que las demás, es decir, páginas que reciben un número mayor de referencias que lo normal. En base a esto la Web tiene una estructura que se puede clasificar como *red libre de escala*. Dichas redes, al contrario de las redes aleatorias, se caracterizan por una distribución dispareja de enlaces y porque dicha distribución sigue una ley de Zipf. Los nodos altamente enlazados actúan como centros que conectan muchos de los

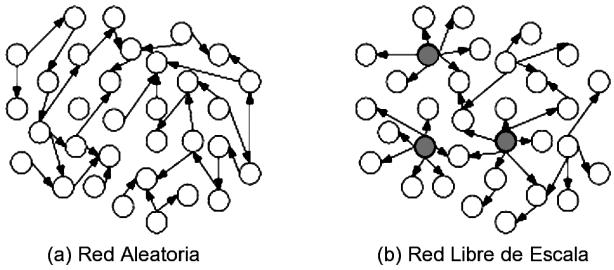


Figura 2.2: Ejemplos ilustrativos de una red aleatoria y una red libre de escala. Cada grafo tiene 32 nodos y 32 enlaces.

otros nodos a la red, como se ilustra en la Figura 2.2. Esto quiere decir que la distribución de los enlaces es muy sesgada: unas pocas páginas reciben muchos enlaces mientras que la mayoría recibe muy pocos o incluso ninguno.

Conectividad

Para conocer qué páginas Web apuntan a una página dada es necesario recorrer toda la Web, algo que los grandes buscadores hacen periódicamente. El primer estudio de la estructura del grafo de la Web fue realizado a partir de dos recorridos de Altavista en Mayo y Octubre de 1999, cada uno de más de 200 millones de páginas (entre un 20% y un 25% de la Web en esa época) y 1.500 millones de enlaces. Sólo almacenar y procesar el grafo equivalente es todo un desafío.

Los resultados de este estudio mostraron que la fracción de páginas de la Web que son apuntadas por i páginas es proporcional a $1/i^{2.1}$, mientras que la fracción de páginas que tienen i enlaces es proporcional a $1/i^{2.7}$. Esto significa que el número de páginas muy apuntadas (populares) y el número

de páginas con muchos enlaces es muy pequeño. Estos valores son casi los mismos para los dos recorridos, pese a que entre ellos pasaron 6 meses.

Estructura

Para analizar la estructura de la Web se buscan las partes del grafo que están conectadas entre sí. El estudio ya mencionado, y el único realizado a nivel global, muestra que el núcleo o centro de la Web lo constituían más de 56 millones de páginas, existiendo un camino para ir de cualquier página a otra, con un largo máximo (diámetro) de al menos 28. En otras palabras, el camino más corto entre dos páginas en el peor caso implicaba visitar 28 de ellas. Esto contrasta con el modelo del mundo pequeño mencionado al comienzo que predecía un diámetro máximo de 20 páginas para toda la Web. En la práctica se encontraron caminos hasta de largo 900, lo que indica que el diámetro de la Web es mucho mayor. De todos modos, este número no es tan grande considerando que son cientos de millones de páginas.

La Figura 2.3 muestra la estructura de la Web de acuerdo al estudio mencionado. A la izquierda había 43 millones de páginas desde las cuales se podía llegar al centro, pero no viceversa. Del mismo modo, a la derecha había otras 43 millones que podían ser accedidas desde el centro, pero que no enlazaban páginas del núcleo. Alrededor de estos dos grupos hay tentáculos que contienen 44 millones de páginas y que son caminos sin salida, con la excepción de algunos tubos, que conectan el grupo de la izquierda con el de la derecha. Finalmente, tenemos 17 millones de páginas que están agrupadas en islas que no están conectadas al centro de la Web. Muchos se preguntarán cómo Altavista conocía estas islas si no están conectadas al resto de la Web y no pueden ser recorridas siguiendo enlaces. Es muy simple: estos son sitios Web que fueron directamente enviados al buscador y por lo tanto están en su índice aunque el resto del mundo no las conozca.

Los autores del estudio no hacen ninguna interpretación sobre esta estructura. En las investigaciones que hemos realizado en Chile, que muestran una estructura similar, el grupo de la izquierda son páginas más nuevas que aún no son demasiado conocidas y que si tienen éxito pasarán al centro de la Web, donde están las páginas consolidadas. En cambio, en el grupo de la derecha son páginas antiguas que no enlazan al centro de la Web porque en su época esas páginas no existían, pero sí fueron enlazadas por nuevas páginas. También incluyen muchos sitios Web que no tienen enlaces externos pero que se han preocupado de tener un enlace desde un buen sitio, por ejemplo vía enlaces publicitarios.

En Chile hemos encontrado que la proporción de sitios que son islas es muy alta, mucho mayor que en el estudio original, gracias a que conocemos todos los dominios `.cl`.

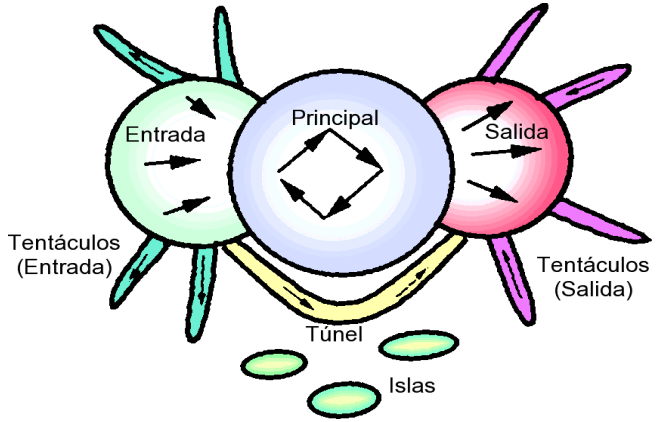


Figura 2.3: Estructura del grafo de la Web.

Dinámica de la Web

Más de la mitad de la Web ha nacido o ha sido actualizada en los últimos seis meses. Parte de ese crecimiento, alrededor de 20%, es replicándose a través de sitios espejos o mirrors u otros tipos de copias (en algunos casos plagio). Al mismo tiempo gran parte de la Web muere. Se estima que el tiempo promedio de vida de una página es alrededor de tres meses. Otra parte de la Web muta, ya sea a través de cambios de nombres de dominio, sitios, directorios o archivos. Es como un organismo caótico, como una colonia de bacterias que está sobrealimentada en algunas partes y en otras agoniza.

La dinámica violenta de la Web y su volatilidad tiene consecuencias importantes. Por ejemplo, sitios Web nuevos serán difíciles de encontrar sin campañas de publicidad, correo electrónico o a través de la comunicación verbal entre personas. Lo mismo para los buscadores. Además, los sitios nuevos tendrán menos sitios que los referencien, con los que son menos importantes para buscadores como Google o Yahoo! que usan los enlaces a un sitio para evaluar su importancia.

Un sitio nuevo generalmente comienza en ISLAS o IN. Luego, si es conocido, pasa al centro de la Web o MAIN. Si luego decide no apuntar a un sitio importante o no es actualizado pasa a la derecha u OUT, o peor aún, se convierte nuevamente en isla. Los componentes más estables en Chile están en MAIN y OUT que tienen el 35% de todos los sitios. En la figura 2.4, mostramos la dinámica de la estructura de la Web Chilena. Los tonos claros indican la procedencia de los sitios antiguos, mientras que los oscuros representan los sitios nuevos. El tamaño del tarro de basura indica la cantidad de sitios que desaparecen.

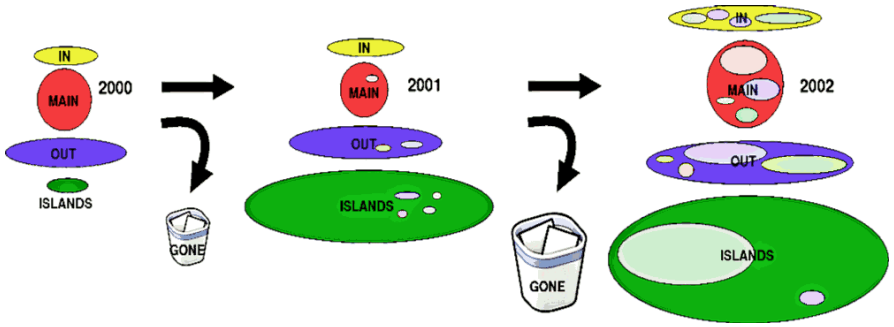


Figura 2.4: Dinámica de la estructura del grafo de la Web Chilena entre los años 2000 y 2002 [3].

La Web Chilena

Definimos como sitio Web chileno aquel que termina en .cl o el cual su IP pertenece a un proveedor chileno de Internet. El último estudio realizado con datos de 2006 mostró los siguientes resultados: La Web chilena está compuesta por más de 170.000 sitios, y estos contienen más de 7 millones de páginas. Muchas de sus características son muy similares a las de la Web global en general.

- El 14% de los sitios están conectados entre sí a través de enlaces y tienen el 53,3% de las páginas. Por otro lado, el 49,5% de los sitios está completamente desconectado en términos de enlaces, pero representan sólo el 14% de las páginas.
- Un sitio promedio tiene 43 páginas, contenidas en 0,304 MiB, con 1,56 referencias desde otros sitios.
- Un dominio promedio tiene 1,08 sitios y 46,61 páginas, contenidas en 0,328 MiB.

- Cerca de 1/4 de las páginas chilenas fue creada o actualizada en el último año, lo que implica un alto grado de crecimiento y dinamismo.
- Alrededor del 80% de las páginas de Chile está en español y cerca del 17% en inglés. Otros idiomas tienen una presencia muy leve.
- Los sustantivos que más aparecen en la Web chilena son: Chile, producto, usuarios, servicio y mensaje. También aparecen Santiago, Web, blog, región e información.
- Los países más referenciados desde Chile son Argentina, España, Alemania, Reino Unido y México, y en general el número de referencias a países extranjeros está relacionado con el volumen de intercambio comercial.
- Los sitios que reciben más enlaces son `sii.cl`, `uchile.cl`, `mineduc.cl`, `meteochile.cl` y `bcentral.cl`.
- Los proveedores de hosting con mayor número de sitios son IFX Networks, VirtuaByte, T-Chile, Telefónica Internet Empresas, DattaWeb y PuntoWeb.

Respecto a la calidad de las páginas y sitios:

- De todos los sitios, el 20% más grande de ellos contiene el 99% de la información en la Web chilena, medida en el número de bytes contenidos en sus páginas.
- Cerca del 21% de los sitios de Chile no son fáciles de encontrar ya que están hechos con tecnologías no visibles para los motores de búsqueda, como Flash y Javascript.
- Unas pocas páginas acaparan la mayoría de los enlaces. De hecho, sólo el 3% de las páginas tienen algún valor de contenido en

términos de estar referenciadas desde otros sitios. Sin embargo, estas páginas están repartidas en el 35% de los sitios Web.

- Cerca de 5% de los enlaces ya no existen.

Respecto a las tecnologías Web:

- De los servidores que entregan información, el servidor Web más utilizado es Apache con 66,7%, seguido por Microsoft Internet Information Server con 32,8%.
- De los servidores que entregan información, el sistema operativo más utilizado es Unix, con 48,5%, seguido por Microsoft Windows con 38,5%. Además, Linux es utilizado en un 12% de los servidores.
- El generador de páginas dinámicas más usado es PHP con 75% de participación en el mercado.
- El formato de documentos más usado es PDF, con 53% de participación, seguido por XML con un 21%.
- Aproximadamente hay una disponibilidad del doble de archivos con paquetes de software para Linux que para Windows en la Web chilena.

Para saber más

- Centro de Investigación de la Web, http://www.ciw.cl
- Google Labs, http://labs.google.com
- Search Engine Watch, http://www.searchenginewatch.com
- TodoCL, el buscador chileno, http://www.todocl.cl
- Web Information Retrieval resources, http://www.webir.org
- World Wide Web Consortium, http://w3c.org
- Yahoo! Research, http://research.yahoo.com

Referencias

1. Information on Zipf's Law. http://www.nslij-genetics.org/wli/zipf/
2. S. Boccaletti et al. "Complex Networks: Structure & Dynamics." Physics Reports, Elsevier. 2006.
3. Ricardo Baeza-Yates, Barbara J. Poblete, Felipe Saint-Jean. "Evolución de la Web Chilena 2001-2002." Centro de Investigación de la Web. 2003. http://www.ciw.cl/recursos/estudio2002/estudio2002html.html
4. Ricardo Baeza-Yates, Carlos Castillo, Eduardo Graells. "Características de la Web Chilena 2006." http://www.ciw.cl/material/web_chilena_2006/index.html
5. R. Albert, H. Jeong and A-L. Barabási. "Diameter of the World Wide Web" Nature 401, 130. 1999.
6. J. Kleinberg et al. "The Web as a graph: measurements, models, and methods." Proceedings of the 5th International Computing and combinatorics Conference, 1999.

Capítulo 3
Internet

José Miguel Piquer

El desarrollo de Internet[2]

En las décadas de 1970 y 1980 los computadores se desarrollaban rápidamente mientras iba siendo claro que existía la necesidad de interconectarlos en redes mundiales, básicamente para poder enviar *mail* desde una parte del mundo a cualquier otra; necesidad básica de la comunidad científica que hasta ese momento sólo disponía de un lento y poco confiable sistema de cartas internacionales para intercambiar ideas y trabajos escritos.

Sin embargo, estas redes se desarrollaban en torno a un tipo determinado de computador: existían la redes de computadores IBM (BITNET), Digital (DECNET), Unix (UUCP), etc. En Chile nos conectamos a la red BITNET y a la red UUCP en 1986. Ambas conexiones llegaban a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, pero BITNET llegaba al

2 En el lenguaje coloquial, muchas veces el lego usa intercambiablemente las nociones de "Internet" y "Web". Desde un punto de vista técnico es necesario diferenciarlas. Una analogía puede ayudar a aclarar la intuición de esta diferencia: el sistema de transporte de pasajeros terrestre está basado en una red de carreteras. Pero el transporte de pasajeros y la red de carreteras son dos cosas completamente diferentes, con problemas diferentes. Lo mismo ocurre para la Web respecto de Internet.

Centro de Computación (en el segundo piso en Blanco Encalada 2120) y UUCP al Departamento de Ciencias de la Computación (en el primer piso de la misma dirección). Estas redes eran incompatibles entre sí, y no teníamos forma de enviar mails desde la una hacia la otra, por lo que tuvimos por un tiempo un sistema de interconexión que consistía de una persona con un diskette que subía y bajaba las escaleras con el mail de una red hacia la otra.

La necesidad clara de construir un sistema interconectado mundial entre todas estas redes fue uno de los motores fundamentales de Internet. El mismo nombre lo indica: el objetivo era construir una inter-red; una red de redes. Internet conquistó el mundo a través de dos tecnologías clave: el protocolo Internet (IP), que permitía conectar a Internet a cualquier tecnología de red existente; y al sistema de nombres de dominio que permitió tener direcciones de correo electrónico únicas e independientes de la tecnología usada. En 1986, en la Universidad de Chile teníamos varias direcciones de mail, las que ocupaban la casi totalidad de la superficie de nuestras tarjetas de visita. Si el nombre de usuario era jperez, en la tarjeta figuraba la siguiente lista:

```
UUCP: ...!seismo!uchdcc!jperez
BITNET: jperez@uchcecvm.BITNET
DECNET: uchvax.DECNET::jperez
X.400: S=jperez; P=uchdcc; A=d400; C=cl;
```

Al comenzar a usar nombres de dominio, la dirección de correo se volvió única (`jperez@dcc.uchile.cl`) y se ha mantenido así por 20 años, a pesar de que la tecnología física de interconexión ha cambiado múltiples veces. Para lograr esto, la Universidad de Chile tuvo que inscribirse como la organización a cargo de administrar el dominio `.CL`, ya que fue la primera en requerir un nombre de este tipo en Chile.

Hoy resulta difícil imaginar la informalidad de esos años, pero todo esto ocurría sin apoyo oficial de ningún tipo, y era simplemente el esfuerzo de un grupo de investigadores motivados tanto en Chile como en el extranjero para que Internet funcionara y se desarrollara.

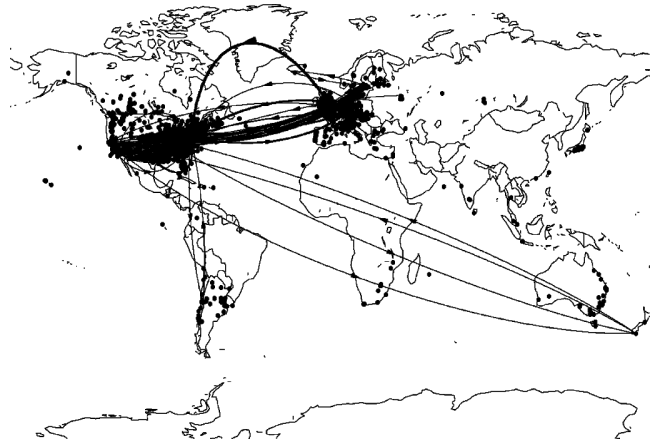
Durante muchos años el dominio `.cl` creció muy lentamente (ver figura 3.1-b). Al cabo de 10 años, comenzaron a aparecer las inscripciones masivas de nombres y hubo que crear una organización formal que administrara los nombres (NIC Chile), un sistema de cobros por dominio y un sistema de administración de los conflictos que surgen en torno a un nombre. NIC Chile continúa operando el dominio .cl bajo el alero de la Universidad de Chile hasta el día de hoy.

En el mundo, los nombres de dominio han sido uno de los principales puntos de conflicto entre el sector privado, el público y la comunidad internacional. Aunque se ha ido avanzando mucho y se han creado organizaciones con bastante apoyo para administrarlos a nivel mundial, aun persisten muchas discusiones en torno a la operación del sistema, su relación con las marcas y la propiedad intelectual y el rol de los gobiernos en los dominios de país.

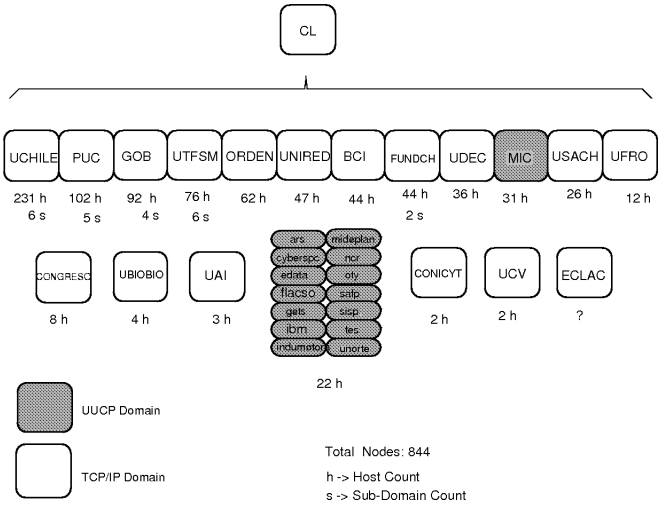
Arquitectura

Para que la Web funcione, se requiere de una Internet que provea básicamente la funcionalidad que permita que cualquier computador conectado a Internet pueda conectarse a un servidor identificado por la URL utilizada.

Parte de esa funcionalidad la provee el ISP (Internet Service Provider) y otra parte la provee mi computador y otra el servidor web de destino.



(a)



(b)

Figura 3.1: (a) Flujo de las *News* en 1993, (b) nombres inscritos bajo el dominio .cl en 1993. Tomados de [1].

La arquitectura Internet divide esta funcionalidad en cuatro servicios:

1. Traducción de nombre de dominio a dirección IP (DNS)

   Este es el servicio inicial que se invoca para traducir un nombre de dominio (como `www.ciw.cl`) a una dirección IP (como `146.83.4.11`), que es básicamente un número único que se requiere para poder llegar al computador destino. Este servicio es crucial para el funcionamiento eficiente de la Web, puesto que todo nombre debe ser traducido antes de poder conectarnos al servidor. La operación requiere de varios servidores de nombres (DNS) que responden por cada dominio, proveiendo redundancia y rapidez en las respuestas.

   Este servicio es provisto en parte por el ISP, quien debe proveernos de un servidor de nombres inicial a quien enviarle nuestras consultas, y en parte por servidores por cada dominio. En el ejemplo, hay un grupo de servidores para `.cl` y otro para `ciw.cl`, los que responden con la dirección IP de `www.ciw.cl`.

2. Conexión y Transporte (socket)

   Una vez obtenida la dirección IP del servidor establecemos una conexión con él, que permite enviar y recibir datos en forma confiable. Esto se hace a través de un `socket` que es la parte más compleja del sistema porque implementa un protocolo de corrección de errores que permite transmitir información sobre una Internet que pierde datos, que los desordena y a veces incluso los duplica.

   La inteligencia del `socket` radica sólo en los extremos de la conexión: el navegador y el servidor. El resto de la red no interviene en este servicio, y eso es fundamental para mantener a Internet como

un servicio barato y eficiente, dado que la complejidad principal la ejecutan los computadores en los extremos. Esto contrasta con la red telefónica que es todo lo contrario: los teléfonos son tontos y toda la inteligencia y complejidad radica en la red misma, lo que la hace mucho más cara.

Este servicio no es provisto por el ISP.

3. Ruteo de paquetes IP

El servicio básico que me debe proveer un ISP es el ruteo de los datos que fluyen entre el navegador y el servidor, los que van en paquetes separados los unos de los otros y que deben pasar a través de varias redes potencialmente en países y continentes diferentes.

Este es el servicio fundamental que me provee el ISP.

4. Protocolo HTTP

Este es el diálogo que se establece entre el navegador (Internet Explorer, Mozilla Firefox, Opera, etc.) y el servidor web en el otro extremo una vez que están conectados. El protocolo permite intercambiar contenidos de todo tipo, como texto, páginas web, imágenes, audio, video, etc. Toda la web está basada en HTTP.

El protocolo original fue desplegado en Internet en 1991 y rápidamente le cambió la cara a Internet; pasó de terminales de texto a navegadores muy parecidos a los actuales.

En resumen, el navegador envía una URL al servidor, quien le responde con el contenido almacenado para esa URL de manera que el navegador lo interprete y decida qué hacer con éste. El diálogo HTTP termina al terminar esa transferencia.

El gobierno de Internet

En inglés se habla de "Internet Governance", que más que un gobierno es una forma de control y supervisión del sistema que nos dé garantías de que esto funcione en forma estable para todos.

En un inicio, cuando Chile se conectó a Internet en 1992, un par de personas controlaban los servicios y asignaban recursos casi sin formalidad alguna. Solicitamos[3] a Jon Postel, quien manejaba los nombres de dominio, que nos asignara la administración de `.cl` ya que estaba vacante. Nos dio la respuesta positiva rápidamente.

Esto ha cambiado mucho y hoy es muy complejo el tema de la administración y asignación de responsabilidades en Internet. En esto participa la comunidad Internet completa, los gobiernos y los organismos internacionales como las Naciones Unidas. Al ser de alcance global, Internet no debe ser controlada por ningún país en particular, pero la comunidad le teme mucho a una administración burocrática tipo Naciones Unidas.

Por ahora, el organismo que intenta administrar esta discusión y los recursos de Internet es ICANN, que es una fundación sin fines de lucro con residencia en California, Estados Unidos. Su autoridad es bastante cuestionada, pero todos respetan sus procedimientos para garantizar la estabilidad operacional de Internet. A modo de ejemplo, `.cl` es uno de los pocos dominios de país que tiene un acuerdo marco firmado con ICANN especificando las responsabilidades de cada parte.

Existe una gran batalla de poder en torno a Internet en la actualidad [2]. Algunos opinan que los países deben tomar control sobre sus recursos al ser un servicio básico, los organismos internacionales consideran que deben

3 Jorge Olivos, Patricio Poblete y yo.

existir leyes globales para regirla y los usuarios sólo queremos que siga funcionando. Afortunadamente, a estas alturas no es fácil tomar acciones locales para ninguno de los actores y se requiere un cierto consenso para llevar a cabo cualquier cambio, lo que da algunas garantías de que el sistema siga operando en forma estable por muchos años más.

Para saber más

- Para saber más sobre el gobierno de Internet, visite el sitio de ICANN: http://www.icann.org
- NIC Chile (http://www.nic.cl) se encarga de administrar los nombres de dominio en Chile.

Referencias

1. Ricardo Baeza-Yates, José M. Piquer, Patricio V. Poblete. "The Chilean Internet Connection or I Never Promised You a Rose Garden." INET '93. http://www.nic.cl/inet93/paper.html
2. .CL. Wikipedia the Free Encyclopedia: http://en.wikipedia.org/wiki/.cl
3. Internet Governance. Wikipedia the Free Encyclopedia: http://en.wikipedia.org/wiki/Internet_governance

Capítulo 4
Buscando en la Web

Gonzalo Navarro

Se dice que los más jóvenes no tienen idea de cómo era buscar información antes que que existiera la Web. Eso es sólo parte de la verdad. Los menos jóvenes tampoco recordamos gran cosa. Nos resulta un ejercicio de imaginación muy difícil recordar cómo vivíamos cuando, ante cualquier consulta, desde cultural hasta de entretenimiento, no podíamos escribir un par de palabras en nuestro buscador favorito y encontrar inmediatamente montañas de información, en general muy relevante.

Para operar este milagro no basta con Internet. Ni siquiera basta con la Web. El ingrediente imprescindible que se necesita son los *buscadores* o *máquinas de búsqueda*. Estos buscadores, cuyos representantes más conocidos hoy son probablemente Google [1], Yahoo! [2] y Microsoft MSN [3], son los que conocen en qué páginas de la Web aparecen qué palabras (y saben bastante más). Sin un buscador, deberíamos conocer las direcciones Web de todos los sitios de bibliotecas, o de turismo, o de cualquier tema que nos pudiera interesar, y los que no conociéramos sería como si no existieran. En un sentido muy real, los buscadores *conectan* la Web, pues existen grandes porciones de la Web a las que no se puede llegar navegando desde otra parte, a menos que se use un buscador. No es entonces sorprendente que casi un tercio del tiempo que los usuarios pasan en Internet lo dediquen a hacer búsquedas.

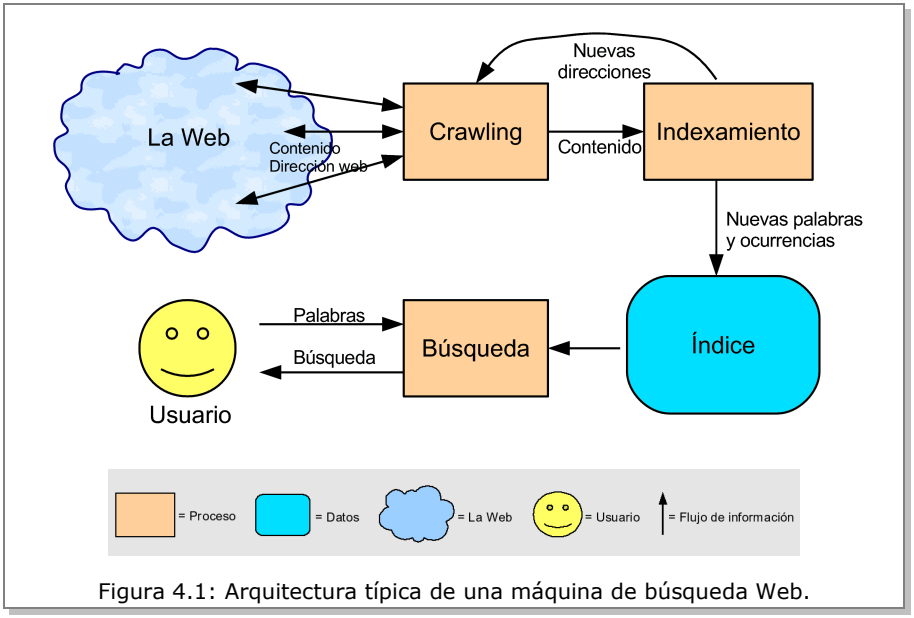


Figura 4.1: Arquitectura típica de una máquina de búsqueda Web.

Esto nos da una primera idea del gigantesco desafío tecnológico y científico que supone desarrollar un buscador. Debemos resolver cuestiones básicas como ¿qué páginas debería conocer un buscador? ¿Qué debería almacenar de esas páginas? ¿Qué tipo de preguntas debería aceptar? ¿Qué debería responder a esas preguntas? ¿Cómo debería mostrar la información? Y ésas son sólo las preguntas más elementales.

Para ordenar la discusión comencemos mostrando la arquitectura típica de una máquina de búsqueda, en la figura 4.1. En ésta, la Web y los usuarios son el mundo exterior al buscador. Todo lo que está a la derecha es parte del buscador.

En el *crawling* se recolectan páginas de la Web, ya sea nuevas o actualizadas. El proceso de *indexamiento* es el que extrae los enlaces que parten de las páginas leídas y realimenta el crawling con nuevas direcciones para visitar, mientras que almacena en el *índice* la información para qué palabras aparecen en qué páginas, junto con una estimación de la importancia de tales ocurrencias. La *búsqueda* usa el índice para responder una consulta, y luego presenta la información al usuario para que éste navegue por ella [4].

Crawling: ¿qué páginas debería conocer un buscador?

Se llama *crawling* al procedimiento de visitar páginas para ir actualizando lo que el buscador sabe de ellas. Un *crawler* es un programa que corre en la máquina del buscador y que solicita a distintos computadores de Internet que le transfieran el contenido de las páginas Web que él les indica. Para estos computadores es casi lo mismo que un crawler o un ser humano visite sus páginas: debe enviarle el contenido de la página solicitada.

¿Qué páginas debería conocer un buscador? ¡Es tentador responder que todas! Pero lamentablemente esto no es posible. La Web cambia demasiado seguido: un porcentaje alto de las páginas cambia de un mes a otro, y aparece un porcentaje importante de páginas nuevas. Internet no es lo suficientemente rápida: se necesitan meses para transmitir todas las páginas de la Web al buscador. Es simplemente imposible mantener una foto actualizada de la Web. ¡Ni siquiera es posible explorarla al ritmo al que va creciendo! La foto que almacena un buscador es siempre incompleta y sólo parcialmente actualizada. No importa cuántos computadores usemos para el buscador. Los mayores buscadores hoy ni se acercan a cubrir el total de la Web. ¡Es incluso difícil saber cuál es el tamaño real de la Web! Esto es aún

peor si consideramos la llamada *Web dinámica*, formada por páginas que se generan automáticamente a pedido (por ejemplo, al hacer una consulta al sitio de una línea aérea), y que son potencialmente infinitas. Y esto considerado que se refieren sólo a la Web pública (de acceso gratuito).

Algunos números pueden dar una idea de las magnitudes involucradas. En 2005 se estimaba que la Web contenía 11.500 millones de páginas, de las cuales los mayores buscadores cubrían a lo sumo el 70%. Algunos estudios calculan que la Web dinámica, por otro lado, puede llegar a los 500 mil millones de páginas.

Querer mantener una foto de la Web al día puede compararse con querer estar al tanto de todo lo que ocurre en todas partes del mundo, hasta los menores detalles locales, mediante leer el diario continuamente. Van ocurriendo más novedades de las que es posible ir leyendo. Podemos pasarnos todo el tiempo leyendo detalles insignificantes y perdiéndonos los hechos más importantes, o podemos tener una política más inteligente de seleccionar las noticias más relevantes, y postergar (tal vez para siempre) la lectura de las menos relevantes.

Un tema fundamental en un buscador es justamente el de decidir qué páginas debe conocer, y con cuánta frecuencia actualizar el conocimiento que tiene sobre cada página. Un crawler comienza con un conjunto pequeño de páginas conocidas, dentro de las cuales encuentra enlaces a otras páginas, que agrega a la lista de las que debe visitar. Rápidamente esta lista crece y es necesario determinar en qué orden visitarlas. Este orden se llama "política de crawling". Algunas variables relevantes para determinar esta política son la importancia de las páginas (debería actualizar más frecuentemente una página que es más importante, lo que puede medirse como cantidad de veces que la página se visita, o cantidad de páginas que la apuntan, o frecuencia con que se buscan las palabras que contiene, etc.), y la frecuencia

de cambio de las páginas (el crawler debería visitar más frecuentemente una página que cambia más seguido), entre otras.

Indexamiento: ¿qué debería almacenarse de las páginas?

El *indexamiento* es el proceso de construir un *índice* de las páginas visitadas por el crawler. Este índice almacena la información de manera que sea rápido determinar qué páginas son relevantes a una consulta.

¿No basta con almacenar las páginas tal cual, para poder buscar en ellas después? No. Dados los volúmenes de datos involucrados (los mayores buscadores hoy indexan más de 3 mil millones de páginas, que ocupan varios terabytes), es imposible recorrer una a una todas las páginas almacenadas en un buscador para encontrar cuáles contienen las palabras que le interesan al usuario. ¡Esto demoraría horas o días para una sola consulta!

El buscador construye lo que se llama un *índice invertido*, que tiene una lista de todas las palabras distintas que ha visto, y para cada palabra almacena la lista de las páginas donde ésta aparece mencionada. Con un índice invertido, las consultas se pueden resolver mediante buscar las palabras en el índice y procesar sus listas de páginas correspondientes (intersectándolas, por ejemplo). La figura 4.2 ilustra un índice invertido.

Los buscadores grandes deben procesar hasta mil consultas por segundo. Si bien este trabajo puede repartirse entre varios computadores, la exigencia sigue siendo alta. El mayor costo para responder una consulta es el de leer de disco las listas de páginas apuntadas por el índice invertido. Es posible usar técnicas de compresión de datos para reducir el espacio en que se representan estas listas. Con esto se logra ganar espacio y velocidad si-

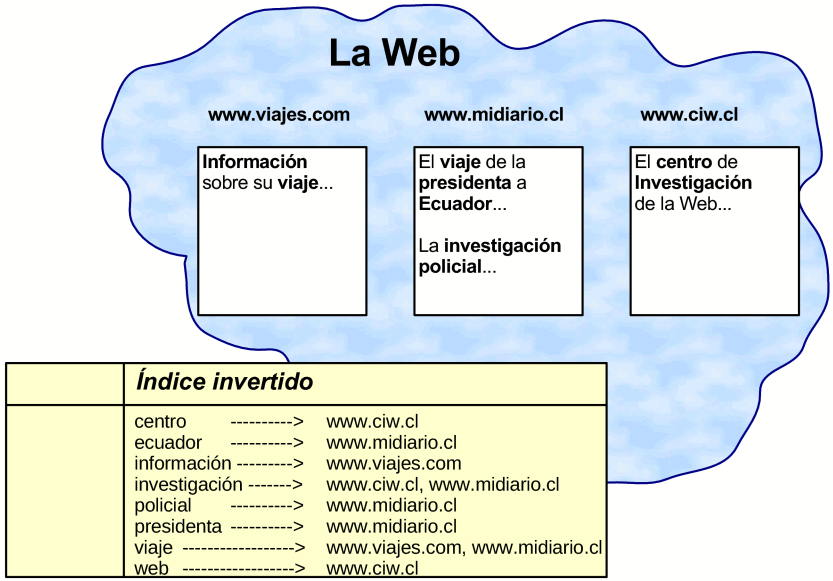


Figura 4.2: Ejemplo de un índice invertido para tres páginas Web.

multáneamente. Pueden hacerse también otras cosas, como precalcular las respuestas a las consultas más populares.

Búsqueda: ¿qué preguntas debería responder, y cómo?

Hemos estado considerando que el usuario escribe algunas palabras de interés y el buscador le da la lista de las páginas donde aparecen estas palabras. La realidad es bastante más complicada. Tomemos el caso más elemental, de una consulta por una única palabra. Normalmente hay millo-

nes de páginas que contienen esa palabra, y está claro que el usuario no tiene la menor posibilidad de examinarlas todas para ver cuáles satisfacen su necesidad de información. De alguna manera el buscador debe *ordenar* las respuestas por su supuesta *relevancia* a la consulta.

Existen muchas formas de calcular esta relevancia, que dan lugar a mejores o peores heurísticas. Por ejemplo, uno puede considerar que una página donde la palabra buscada aparece varias veces es más relevante que otra donde aparece una vez. Pero si la palabra aparece más veces en una página que es mucho más larga que otra, entonces tal vez la palabra no sea tan importante en esa página. También uno puede considerar cuan importante es la página en sí (por ejemplo si es muy visitada, o muy apuntada por otras). Los buscadores utilizan fórmulas matemáticas para calcular la relevancia que tienen en cuenta estos aspectos.

Existen técnicas más sofisticadas, por ejemplo llevar información de cómo se comportaron otros usuarios cuando hicieron esta misma consulta (por ejemplo, el buscador puede saber que la gran mayoría de los usuarios que buscaron `mp3` terminaron yendo a ciertos sitios específicos). Esto se llama *minería de consultas* y es extremadamente útil para dar buenas respuestas a consultas que no dicen mucho. También puede usarse información posicional, por ejemplo si la palabra aparece en el título de la página o de los enlaces que la apuntan, puede ser más relevante que si aparece cerca del final.

La situación se complica cuando la consulta tiene varias palabras, donde algunas pueden ser más importantes que otras. Normalmente las ocurrencias de palabras que aparecen en muchos documentos, como los artículos y preposiciones, son poco importantes porque no sirven para discriminar. Para peor, sus listas de ocurrencias en los índices invertidos son muy largas, ocupando espacio inútil. Por ello muchos buscadores las omiten

de sus índices (intente buscar `and` en su buscador favorito). La forma de combinar el peso de las distintas palabras da lugar también a mejores o peores heurísticas. Por ejemplo los buscadores en la Web normalmente muestran sólo páginas donde aparecen todos los términos, como una forma de eliminar respuestas irrelevantes. Asimismo, los mejores dan preferencia a páginas donde las palabras aparecen cercanas entre sí.

Pero la verdad es que en la Web hay mucha, mucha más información de la que se puede obtener mediante buscar documentos que contengan ciertas palabras. Esta limitación se debe a que no es fácil implementar búsquedas más sofisticadas a gran escala. Conseguir responder consultas más complejas a escala de la Web es un tema actual de investigación. Algunos ejemplos son:

1. Buscar por contenido en fotos, audio o video. Imagínese mostrar una foto de su promoción y poder encontrar otras fotos de las mismas personas en la Web, incluso sin recordar sus nombres. O tararear una parte de una melodía (incluso con errores) y encontrar el mp3 para poder bajarlo. Existen técnicas para hacer esto, pero no a gran escala. Los buscadores ofrecen búsqueda de fotos, pero basada en palabras que una persona se encarga de asociar a cada foto durante el crawling.
2. Hacer preguntas complejas que se pueden inferir de la Web. Por ejemplo preguntas como ¿cuál es la farmacia más cercana que venda un antigripal a un precio inferior a $ 3.000? y ¿qué universidades dictan una carrera de Diseño Gráfico de 5 años en la Región Metropolitana? Responder este tipo de preguntas requiere normalmente de cierta cooperación de quien escribe las páginas.

3. Hacer consultas con componente temporal, como ¿qué ocurrió con el seguimiento en los medios de comunicación a las consecuencias de la guerra en el Líbano en los meses siguientes a su finalización? Esto requiere llevar una cuenta histórica de los contenidos de la Web a lo largo del tiempo.

Interacción con el Usuario: ¿cómo presentar la información?

Ya vimos que las respuestas que se muestran al usuario son sólo una mínima parte de las que califican. Los buscadores normalmente presentan una lista de las primeras páginas según el orden que han hecho en base a la consulta. En esta lista se indica la dirección de la página (para que el usuario pueda visitarla con un click) y usualmente el *contexto* del texto donde las palabras aparecen. Esto ayuda al usuario a saber rápidamente si las palabras aparecen en la forma que esperaba (por ejemplo `investigación` puede referirse a científica o policial).

Poder mostrar un contexto requiere que el buscador no almacene sólo el índice invertido, sino también el contenido completo de las páginas que indexa. Si bien el espacio es barato, esto es un requerimiento bastante exigente, ¡pues el buscador debería tener suficiente almacenamiento para duplicar toda la Web en sus discos! Por ejemplo, para reducir el espacio, el buscador puede evitar almacenar las imágenes. La compresión de datos es también útil para aliviar este problema.

Los buscadores suelen ser lo suficientemente buenos como para que, en un gran porcentaje de las veces, lo que busque el usuario esté entre las primeras respuestas que ofrece. De todos modos es posible pedirle que

entregue el siguiente conjunto de respuestas, y el siguiente, hasta hallar lo que uno busca. La experiencia normal es que, si la respuesta no está en las primeras páginas, es raro que esté más adelante. En esos casos es mejor reformular la consulta, por ejemplo haciéndola más específica (si se encontraron demasiadas páginas irrelevantes) o más general (si se encontraron muy pocas respuestas). Por ejemplo, en la figura 4.2, si buscáramos `investigación` encontraríamos tanto la página del Centro de Investigación de la Web como la noticia policial. Refinando la consulta a `investigación policial` tendríamos mejor precisión. Esta iteración es frecuente en las sesiones con los buscadores, y con el tiempo el usuario aprende a formular consultas más exitosas.

Existen formas mucho más sofisticadas de presentar la información, pero nuevamente es difícil aplicarlas a sistemas masivos como la Web. Asimismo suele ocurrir que las interfaces demasiado "inteligentes" resultan ser demasiado complejas para la mayoría de la gente. Incluso los lenguajes de consulta más sofisticados, donde se puede indicar que las palabras *A* y *B* deben aparecer, pero no *C*, normalmente están disponibles en los buscadores Web, pero se usan muy raramente. La regla en este caso es que la simplicidad es lo mejor.

Para saber más

- El sitio www.searchenginewatch.com está dedicado a las estadísticas sobre las principales máquinas de búsqueda en la Web.
- Los sitios http://www.press.umich.edu/jep/07-01/bergman.html y http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ están dedicados a estudiar el crecimiento de la Web, y en general de la cantidad de información disponible en el mundo.
- El sitio www.todocl.cl es el buscador chileno Todo.cl.

Referencias

1. Google. http://www.google.com
2. Yahoo! http://www.yahoo.com
3. Microsoft MSN. http://www.msn.com
4. Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley-Longman, 1999. Capítulo 13.

Capítulo 5
Manejo de grandes volúmenes de información utilizando Clusters de computadores

Mauricio Marín

Se estima que la cantidad de texto almacenado en los distintos sitios Web del mundo es del orden de centenas de Terabytes, y la cantidad de información disponible crece día tras día. En este escenario es evidente que almacenar y procesar toda esa información utilizando un sólo computador es prácticamente imposible. Lo que hacen los buscadores actuales es utilizar muchos computadores para resolver los distintos pasos involucrados en la producción de una respuesta a una consulta de usuario [1]. A este conjunto de computadores se les llama *cluster*.

Un cluster está compuesto de un conjunto de computadores interconectados mediante una red que les permite enviarse mensajes entre ellos (ver figura 5.1). Estos mensajes se utilizan para recolectar la información necesaria para resolver una determinada tarea como por ejemplo la solución a una consulta de un usuario. En el cluster cada computador tiene su propia memoria RAM y disco para almacenar información. Cada computador puede leer y escribir información en su propia memoria y si necesita información

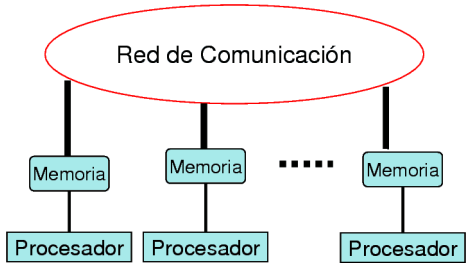


Figura 5.1: *P* computadores con memoria local e interconectados mediante una red.

almacenada en otro computador debe enviarle un mensaje y esperar la respuesta.

Un ejemplo que muestra la manera en que esto trabaja es el siguiente (ver figura 5.2). Supongamos que existen P computadores de un cluster que necesitan tener en su memoria un libro de N páginas para poder trabajar, pero dicho libro se encuentra almacenado en un sólo computador, digamos el computador 1. Lo que puede hacer el computador 1 es dividir el libro en P partes cada una de N/P páginas y enviar una parte distinta a cada uno de los P-1 computadores del cluster. Luego de este paso cada computador queda con una parte distinta del libro. Luego, en un segundo paso, cada computador envía a todos los otros la parte de tamaño N/P que tiene almacenada en su memoria. Al final de este paso todos los computadores quedan con una copia completa del libro.

Una estrategia alternativa es simplemente hacer que el computador que tiene el libro envié un mensaje a cada uno de los P-1 restantes computadores con una copia del libro. El resultado final es el mismo, pero es menos eficiente que el primer método porque no existe el paralelismo que se produce en el

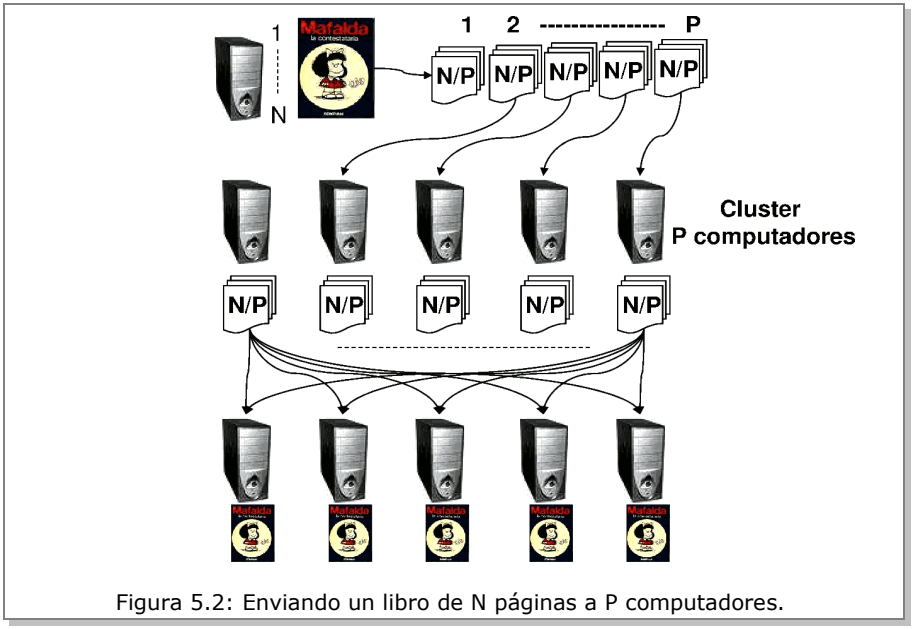


Figura 5.2: Enviando un libro de N páginas a P computadores.

segundo paso cuando todos al mismo tiempo están enviando una copia de su parte de tamaño *N/P* a todos los otros.

Máquinas de búsqueda y Clusters

En un cluster utilizado como máquina de búsqueda, cada computador tiene su propia memoria RAM y disco para almacenar una parte de la información del sistema completo. Por ejemplo, si tenemos una colección de texto bajado de la Web por el *crawler* que ocupa *N* bytes y tenemos un cluster con *P* computadores, entonces podemos asignar a cada uno de los *P* computadores una fracción *N/P* de los bytes de la colección. En la práctica si la colección

completa tiene *D* documentos o páginas Web, entonces a cada computador del cluster se le asignan *D/P* documentos.

En una máquina de búsqueda las consultas de los usuarios llegan a un computador recepcionista llamado *broker*, el cual distribuye las consultas entre los *P* computadores que forman el cluster (ver figura 5.3). Tal como se muestra en la figura 4.1, las máquinas de búsqueda utilizan un índice invertido para disminuir el tiempo de procesamiento requerido para obtener la respuesta a una consulta.

Dado que cada computador del cluster tiene un total de *D/P* documentos almacenados en su memoria, lo que se hace es construir un índice invertido en cada computador con los documentos almacenados localmente en cada uno de ellos. Entonces cada vez que el broker recibe una consulta de un usuario, este envía una copia de la consulta a todos los computadores del cluster (notar que podemos considerar un grupo grande de consultas como

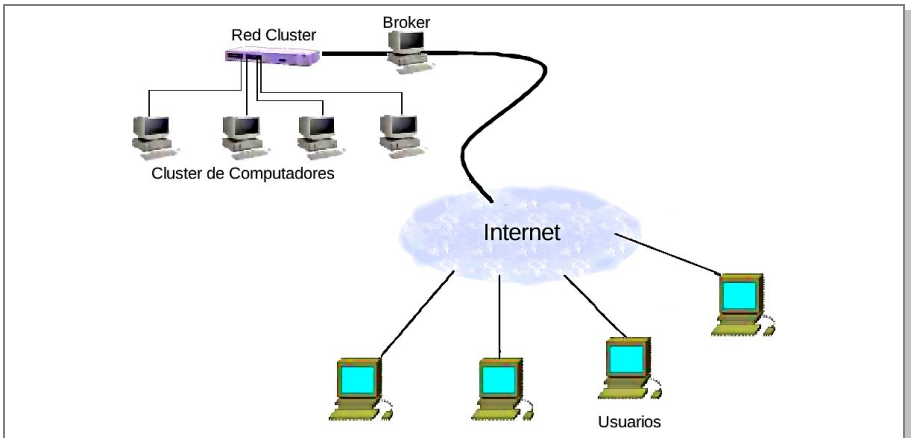


Figura 5.3: Organización de un cluster de computadores.

un libro y por lo tanto el broker puede distribuirlas de manera eficiente utilizando la estrategia de la figura 5.2). En el siguiente paso, todos los computadores en paralelo leen desde su memoria las listas invertidas asociadas con las palabras que forman la consulta del usuario. Luego se realiza la intersección de las listas invertidas para determinar los documentos que contienen todas las palabras de la consulta.

Al término de este paso todos los computadores tienen un conjunto de respuestas para la consulta. Sin embargo, la cantidad de respuestas puede ser inmensamente grande puesto que las listas invertidas pueden llegar a contener miles de identificadores de documentos que contienen todas las palabras de la consulta. Es necesario hacer un ranking de los resultados para mostrar los mejores K resultados al usuario como solución a la consulta.

Para realizar el ranking final de documentos es necesario colocar en uno de los computadores del cluster los resultados obtenidos por todos los otros. Esto con el fin de comparar esos resultados unos con otros y determinar los mejores K. Sin embargo, enviar mensajes conteniendo una gran cantidad de resultados entre dos computadores puede consumir mucho tiempo. Es deseable reducir la cantidad de comunicación entre computadores.

Ahora, si cada computador ha calculado los mejores resultados para la consulta considerando los documentos (listas invertidas) que tiene almacenados en su disco, entonces no es necesario enviarlos todos al computador encargado de realizar el ranking final. Basta con enviar a este computador los K mejores de cada uno de los P-1 computadores restantes. Es decir, el ranking final se puede hacer encontrando los K mejores entre los $K \times P$ resultados aportados por los P computadores.

Pero esto se puede mejorar mas aún y así reducir al máximo la cantidad de comunicación entre los computadores. Dado que los documentos están uniformemente distribuidos en los P computadores es razonable pensar que cada computador tendrá más o menos una fracción K/P de los mejores K resultados mostrados al usuario. Entonces lo que se puede hacer es trabajar por ciclos repetitivos o iteraciones. En la primera iteración todos los computadores envían sus mejores K/P resultados al computador encargado de hacer el ranking final. Este computador hace el ranking y luego determina si necesita más resultados de los otros computadores. Si es así entonces pide nuevamente otros K/P resultados y así hasta obtener los K mejores (ver figura 5.4). Esto porque si tenemos mala suerte podría ocurrir que para esa consulta en particular uno de los computadores posea los K mejores resultados que se le van a entregar al usuario, caso en que se necesitan P iteraciones para obtener la respuesta para el usuario. Pero es muy poco probable que esto ocurra para todas las consultas que se procesan en una máquina de búsqueda grande. En la práctica se requieren uno o a lo más dos iteraciones para la inmensa mayoría de las consultas, lo cual permite reducir considerablemente el costo de comunicación entre los computadores del cluster.

En las máquinas de búsqueda más conocidas se reciben alrededor de 600 consultas por segundo. Una manera de explotar al máximo la capacidad de los computadores del cluster es hacerlos trabajar en paralelo. Esto se puede lograr asignando los computadores para hacer el ranking de manera circular. Por ejemplo, el computador broker elige al computador 1 para hacer el ranking de la consulta q_1, al computador 2 para la consulta q_2, ..., el computador P para la consulta q_p, el computador 1 para la consulta q_{p+1}, y así sucesivamente de manera que en un instante dado podamos tener a P computadores haciendo el ranking de P consultas distintas en paralelo.

Recolección de páginas Web y Clusters

Para poder realizar consultas de información en una máquina de búsqueda necesitamos que ésta contenga información actualizada de la Web. Los buscadores comerciales tienen software en operación que está constantemente conectándose a los sitios Web de todo el mundo para bajar los documentos de los sitios e indexarlos (es decir, actualizar el índice invertido de la máquina de búsqueda) y ponerlos a disposición de los usuarios.

La Web mundial es inmensamente grande y los enlaces a Internet tienen limitaciones de velocidad de transferencia de datos, por lo tanto no es posible bajar toda la Web en un par de horas. Por ejemplo, actualmente bajar toda la Web Chilena toma de 4 a 5 días utilizando un solo computador conectado a un enlace de alta velocidad. Para bajar la Web mundial es necesario utilizar clusters de computadores cuyo número varía entre diez y veinte mil computadores y es un proceso que demora varias semanas.

Gran parte del éxito de una máquina de búsqueda descansa en su capacidad de almacenar la versión más reciente de la Web. Por lo tanto es necesario establecer un orden para realizar las visitas a los sitios Web de manera de recuperar primero los sitios que son de mayor interés para los usuarios. Una manera de asignar una medida de "interés" para los sitios Web es suponer que los sitios que son más "apuntados" por otros sitios reconocidos como importantes son también interesantes para los usuarios. Un sitio *a* es apuntado por otro sitio *b*, si en el sitio *b* hay páginas Web que tienen enlaces o referencias a las páginas del sitio *a*.

La primera página de un sitio Web es llamada *home-page*. Una o más páginas son descubiertas si, cuando bajamos una página, ésta contiene enlaces a páginas nuevas que no han sido consideradas anteriormente. Entonces

si bajamos el home-page de un sitio podemos descubrir nuevas páginas desde los enlaces que esta página tiene.

Una estrategia para recuperar las páginas Web de los distintos sitios en orden de importancia es calcular un número real que indica la importancia de cada página que se descubre. La próxima página a bajar es la que presenta un mayor valor numérico en ese instante. Por ejemplo, podemos usar la siguiente regla para numerar las páginas. Inicialmente les damos el valor 1 a todos home-pages conocidos. Cada vez que se baja un home-page le repartimos de manera equitativa el valor 1 a todas las páginas a las que el home-

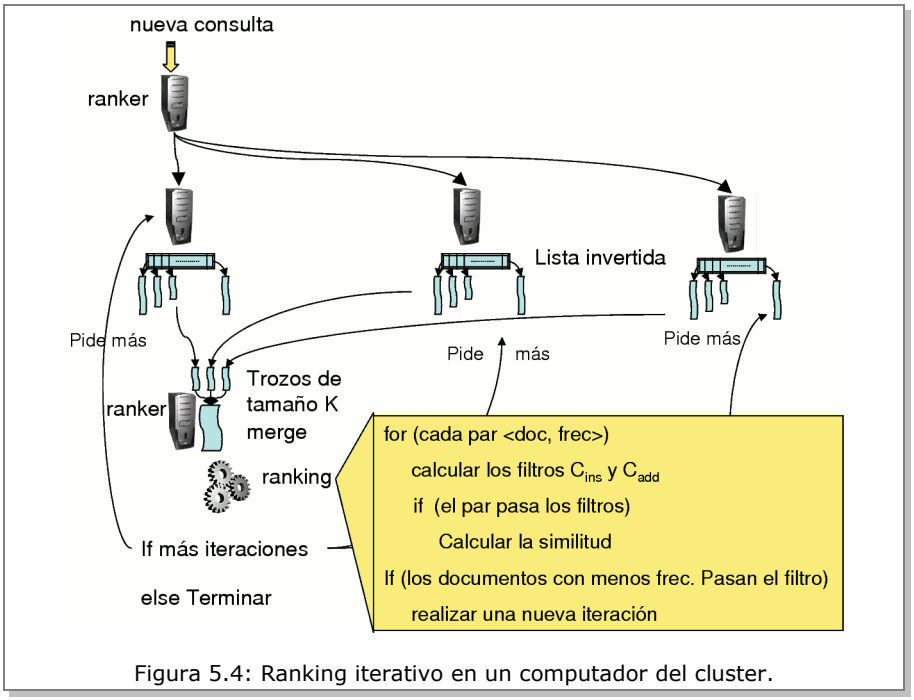


Figura 5.4: Ranking iterativo en un computador del cluster.

page apunta (páginas referenciadas por los enlaces del home-page). A su vez, estas páginas de segundo nivel apuntan a otras páginas y hacemos lo mismo, es decir, el valor de estas páginas es repartido a las páginas apuntadas por ellas y así sucesivamente. La figura 5.5 muestra un ejemplo de dos sitios con home-pages dados por las páginas *A* y *B*. En este ejemplo, la página C es la tercera página a ser bajada puesto que recibe el valor 1/3 desde la página *A*, y el valor 1/4 desde la página *B*.

La manera de poner a muchos computadores a bajar la Web mundial es distribuir de manera equilibrada todos los home-pages conocidos en *P* computadores. Por ejemplo, en el caso de la figura 5.5 el home-page *A* es alojado en el computador 1 y el home-page *B* es puesto en el computador 2. De esta manera el computador 1 puede bajar la página *A* al mismo tiempo que el computador 2 baja la página *B*. Sin embargo, debe haber un punto de comunicación entre los computadores puesto que una vez que el computador 1 baja la página *A*, este ha numerado con 1/3 las tres páginas a las que apunta y por lo tanto podría elegir a cualquiera de estas tres como la siguiente página a bajar. Luego, si no hay comunicación entre los computadores 1 y 2, el computador 1 podría elegir una página distinta a la página *C* como la siguiente página a ser bajada. Una situación similar ocurre con la página *E* si el computador 1 no le envía mensajes al computador 2 indicando cambios en la numeración de las páginas del sitio *B*.

Una solución poco eficiente es hacer trabajar a los computadores en forma estrictamente sincrónica permitiéndoles bajar sólo una página para luego realizar el intercambio de mensajes. Sin embargo esto puede resultar en una sub-utilización del paralelismo disponible, puesto que no siempre ocurren casos como el mostrado en la figura 5.5. Para una Web inmensamente grande es más práctico permitir que los computadores trabajen bajando muchas

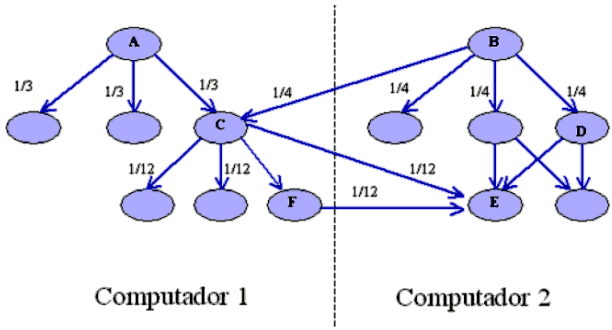


Figura 5.5: Dos sitios Web asignados a dos computadores.

páginas para luego iniciar una fase de intercambio de mensajes y re-numeración de páginas. Claramente hay una situación de compromiso entre la cantidad de páginas que les dejamos bajar antes de iniciar la fase de comunicación, y el error que se puede cometer al re-numerar tardíamente.

Esto muestra que algunas veces hacer trabajar a muchos computadores en paralelo de manera eficiente involucra pensar en la solución a problemas que no surgen cuando se trabaja con un solo computador. En este caso podemos alcanzar gran eficiencia permitiendo el error pero de forma controlada. Por ejemplo, cada computador puede bajar un número n de páginas y al finalizar la fase de comunicación determinar la magnitud del error cometido y, en base a esa evaluación, ajustar el valor de n para el siguiente ciclo.

Para saber más

- Una presentación en el Centro de Investigación de la Web sobre el mismo tema: www.ciw.cl/material/tw07mmarin.pdf
- El artículo "Web Search for a Planet: The Google Cluster Architecture," de Luiz Barroso, Jeffrey Dean y Urs Hoelzle, comenta la arquitectura de clusters de Google: http://labs.google.com/papers/googlecluster.html

Referencias

1. Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval Addison-Wesley-Longman, 1999. Capítulos 9 y 13.
2. Luiz Barroso, Jeffrey Dean, Urs Hoelzle. "Web Search for a Planet: The Google Cluster Architecture." IEEE Micro, Marzo/Abril 2003 (vol. 23, nro. 2). Páginas 22-28. http://labs.google.com/papers/googlecluster.html

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

Marcelo Arenas

Una de las razones para la popularización de la Web ha sido el desarrollo de una infinidad de páginas que entregan distintos servicios; buscadores como Yahoo! y Google, grandes repositorios de información como Wikipedia, tiendas electrónicas como Amazon, diarios y revistas electrónicas, página personales, etc. Bajo este desarrollo ha estado HTML, un lenguaje que permite estructurar tanto la información como las posibilidades de navegación en una página Web.

Durante los últimos años, la cantidad de información almacenada en la Web ha ido creciendo de manera dramática. Hoy ningún usuario tiene la capacidad de recorrer la Web entera en busca de información, y es necesario utilizar buscadores automáticos como Yahoo! y Google para poder revisar una fracción significativa de esta red.

Nadie puede negar la importancia y utilidad que tienen los buscadores para encontrar información en la Web. Sin embargo, muchos usuarios pueden decir que su experiencia con ellos no ha sido completamente satisfactoria. A medida que las consultas que se quiere realizar son más complejas, la búsqueda de información puede requerir de varios, o muchos, intentos en los cuales es necesario jugar con distintos parámetros. Piense por ejemplo en la

consulta "dé la lista de libros de Ariel Rubinstein". Para realizar esta consulta basta con poner "Ariel Rubinstein" en un buscador y usar los primeros elementos de la lista de respuesta (probablemente el primero) para encontrar la página de este autor, y ahí la lista de sus libros. Pero ahora piense en la pregunta "dé la lista de libros de Ariel Rubinstein y sus precios". ¿Qué colocaría en un buscador para encontrar la respuesta? Peor aun, piense en una pregunta como la siguiente "dé la lista de libros de Ariel Rubinstein que han bajado de precio en los últimos años". ¿Cómo se puede buscar esta información usando Yahoo! o Google?

¿Por qué los buscadores tienen dificultades en los ejemplos anteriores? Una de las razones es el uso de HTML; este es un lenguaje que permite desplegar información que es fácil de entender para los usuarios, pero que en general es difícil de interpretar para los computadores. Estas dificultades ya pueden verse en ejemplos tan sencillos como el siguiente:

```
<html>
  <body bgcolor="#FFFFFF">
     <center>
        <h2> Todo Libros </h2>
     </center>
     <ul>
       <li><b>Teor&iacute;a de Juegos.</b>
       Martin Osborne y Ariel Rubinstein. Precio: 16000.</li>
     </ul>
  </body>
</html>
```

Este archivo es usado para mostrar la lista de libros vendidos por la librería "Todo Libros". Nótese que este archivo ha sido indentado (espaciado) de manera que sea fácil visualizar la estructura jerárquica del documento. Por ejemplo, `<li>` corresponde a un ítem en la lista definida por `<ul>`. En un

browser tal como FireFox o Explorer, esta lista será desplegada de la siguiente forma:

Todo Libros

- **Teoría de Juegos.** Martin Osborne y Ariel Rubinstein.
  Precio: 16000.

Para un usuario la información en esta lista es fácil de entender; es claro que hay una lista de libros, cada uno con sus autores y su precio. Sin embargo, para un computador esta información no es tan clara. Una de las razones es que el computador no tiene la información de contexto, o meta-información, que tiene el usuario. ¿Cómo puede un computador deducir que está frente a una lista de libros? Y aun si sabe esto, ¿cómo puede extraer información desde el documento, por ejemplo los precios de los libros? Es importante notar aquí que el documento HTML no tiene ninguna indicación sobre donde buscar esta información, simplemente dice cómo debe ser desplegada la lista de libros. Así, el computador debe tratar de interpretar el texto para poder extraer la lista de precios. Por ejemplo, puede buscar la palabra "Precio" y el número que lo sigue (o antecede). Aunque en este caso esto puede dar buenos resultados, la situación puede volverse más complicada si la lista contiene varios precios para un mismo libro (precio sin descuento, con descuento por compra electrónica, con descuento a clientes frecuentes, etc), o aun más complicada si se requiere de hacer algunos cálculos para saber el precio final (precio después del 15% de descuento por compra electrónica).

La búsqueda de información en la Web puede mejorarse si los formatos usados para almacenar información pueden ser fácilmente interpretados por

los computadores. Una propuesta para hacer esto es el uso de XML, como se verá en las siguientes secciones.

XML: Un lenguaje para almacenar información

Un documento XML (eXtensible Markup Language [2]) es similar a un documento HTML; está compuesto por marcadores, o "tags", que están anidados como en el caso de HTML. La mayor diferencia es que los marcadores de HTML tienen significados predefinidos, tales como `<title>` y `<ul>`, mientras que los de XML son definidos por el usuario. Por ejemplo, el siguiente es un documento XML que almacena la misma información que el documento HTML mostrado en la sección anterior:

```
<?xml version="1.0"?>
<libreria>
  <nombre>Todo Libros</nombre>
  <libro>
     <titulo>Teor&iacute;a de Juegos</titulo>
     <autor>
        <nombre>Martin</nombre>
        <apellido>Osborne</apellido>
     </autor>
     <autor>
        <nombre>Ariel</nombre>
        <apellido>Rubinstein</apellido>
     </autor>
     <precio>16000</precio>
  </libro>
</libreria>
```

Como puede verse, el documento está compuesto por marcadores tales como <libreria>, <libro> y <autor>. Un marcador con nombre <a> es cerrado por uno con nombre . Los nombres de los marcadores fueron definidos por un usuario, y la única restricción que deben cumplir, como en el caso de HTML, es que deben estar correctamente anidados; si leyendo el documento de arriba hacia abajo <autor> aparece después de <libro>, entonces el marcador </autor> que lo cierra debe aparecer antes que el marcador </libro> que cierra a <libro>, vale decir, <autor> debe estar completamente contenido dentro de <libro>. A través de esto se especifica que <autor> es uno de los autores de <libro>.

Los marcadores del documento XML fueron diseñados para mostrar de forma clara la información sobre un libro. Si un computador quiere buscar el título de un libro, entonces basta con que busque el marcador <titulo>, y si quiere encontrar el precio del libro con título "Teoría de Juegos", entonces basta que encuentre un marcador <libro> que tenga "Teoría de Juegos" en <titulo>, y que después despliegue lo que se encuentra en el marcador <precio> dentro de ese libro. La forma en que la información está agrupada y los nombres de los marcadores le indican a un computador dónde buscar información.

XML entonces surge como una buena alternativa para almacenar información; un computador tiene mayores posibilidades de interpretar y extraer información desde este tipo de documentos. ¿Debemos entonces reemplazar HTML por XML? La respuesta es no. Estos dos lenguajes tienen distintas finalidades. Mientras HTML es usado para especificar cómo desplegar información en un browser, XML es usado para almacenar información y no contiene indicaciones de como mostrarla. Se tiene entonces que diseñar tecnologías que permitan sacar ventajas de los dos lenguajes. En la siguiente sección se verá cómo hacer esto.

Transformación de documentos XML

Una de las razones para la creación de XML fue tener un formato que permitiera intercambiar información en la Web. La idea es que si varias personas o empresas desean intercambiar datos sobre un tema común, por ejemplo libros, y usan formatos XML distintos para almacenar su información, entonces puedan intercambiar información de manera sencilla. La forma de hacer esto es usando algún lenguaje de transformación que permita cambiar de un formato a otro. Por ejemplo, si una empresa usa el formato:

```
<autor>
   <nombre>Martin</nombre>
   <apellido>Osborne</apellido>
</autor>
```

para almacenar los nombre de autores de libros, mientras otra usa un formato más simple donde el nombre es almacenado como una sola palabra:

```
<autor>Martin Osborne</autor>
```

entonces una regla de transformación desde el primer formato al segundo debe concatenar el nombre y apellido de un autor para generar su nombre como una sola palabra.

XML fue elegido como el lenguaje para intercambiar información por su gran flexibilidad, esencialmente cualquier documento XML es válido mientras la anidación de los marcadores sea correcta. El lenguaje elegido para especificar las transformaciones fue XSLT (Extensible Stylesheet Language Transformations [3]). Este es un lenguaje que busca patrones dentro de un documento e indica cómo reestructurarlos. Por ejemplo, busca el tag `<autor>`, y después indica que las palabras que aparecen dentro de `<nombre>` y `<apellido>` para este autor tienen que ser concatenadas.

XSLT no sólo permite hacer transformaciones entre documentos XML, en general permite generar cualquier tipo de documento desde un documento XML (HTML, texto plano, programa en algún lenguaje de programación como Java o C++, etc). En particular, hoy es usado por browsers tales como FireFox y Explorer para poder desplegar documentos XML. La idea aquí es simple: como XML es un mejor formato para almacenar información, conviene tener los documentos en la Web en este formato. Si un documento XML tiene que ser desplegado por un browser, entonces se usa un conjunto de reglas XSLT para generar un documento HTML desde la fuente XML, el cual es usado por el browser al desplegar la información. Veamos esto en el ejemplo anterior. Para indicar cuál es el programa XSLT a usar al desplegar un documento XML se usa una línea adicional en el documento:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="libreria.xslt"?>
<libreria>
   <nombre>Todo Libros</nombre>
   <libro>
      ...
   </libro>
</libreria>
```

En el campo `href="libreria.xslt"` se indica que se debe usar el archivo XSLT `libreria.xslt`. En la figura 6.1 se muestra parte del conjunto de reglas XSLT que es usado para transformar el documento XML, con información sobre libros en el documento HTML mostrado en la primera sección.

No se espera aquí que el lector pueda entender todos los detalles de un documento XSLT, pero sí que después de terminar esta sección tenga una idea de cómo funciona este lenguaje. Como puede verse en la figura 6.1, un documento XSLT está compuesto por una serie de patrones que son declarados a través del marcador `xsl:template`. Cada uno de estos patrones tiene

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

  <xsl:template match="/">
    <html>
      <body bgcolor="#FFFFFF">
        <center>
          <h2>
            <xsl:apply-templates select="/libreria/nombre"/>
          </h2>
        </center>
        <ul>
          <xsl:apply-templates select="/libreria/libro"/>
        </ul>
      </body>
    </html>
  </xsl:template>

  <xsl:template match="/libreria/nombre">
    <xsl:value-of select="."/>
  </xsl:template>

  <xsl:template match="/libreria/libro">
    ...
  </xsl:template>

  ...

</xsl:stylesheet>
```

Figura 6.1: Reglas XSLT para transformar un documento XML en HTML.

un atributo `match` que indica dónde se debe usar el patrón. Por ejemplo, el patrón:

```
<xsl:template match="/libreria/nombre">
  <xsl:value-of select="."/>
</xsl:template>
```

debe ser usado en todos los nodos del documento XML que son alcanzados siguiendo el camino `/libreria/nombre` desde el punto inicial del documento. Así, en el ejemplo se va a alcanzar el elemento con marcador `<nombre>`, que es hijo del elemento con marcador `<libreria>`. En el patrón de arriba, se utiliza `xsl:value-of` para indicar qué seleccionar desde este elemento, en este caso `Todo Libros` ya que se usa `select="."`.

Nótese que el documento XSLT tiene un solo patrón tal que `match="/"`. Este es el primer patrón que debe ser usado, y en él se indica que el documento a construir es de la forma:

```
<html>
  <body bgcolor="#FFFFFF">
    <center>
      <h2>
        <xsl:apply-templates select="/libreria/nombre"/>
      </h2>
    </center>
    <ul>
      <xsl:apply-templates select="/libreria/libro"/>
    </ul>
  </body>
</html>
```

En este documento HTML aparece dos veces `xsl:apply-templates`. Esto es usado para indicar que en esos puntos se debe colocar los resultados de aplicar los patrones correspondientes. Por ejemplo, en el caso de:

```
<h2> <xsl:apply-templates select="/libreria/nombre"/> </h2>
```

se debe usar el patrón que contiene la expresión `match="/libreria/nombre"`. Ya se había visto que este patrón genera como respuesta `Todo Libros`, por lo que al hacer el reemplazo se va a obtener:

```
<html>
  <body bgcolor="#FFFFFF">
    <center>
      <h2> Todo Libros </h2>
    </center>
    ...
  </body>
</html>
```

Si se compara esto con el documento HTML mostrado en la sección inicial, se dará cuenta que lo que se muestra arriba coincide con la primera parte del documento HTML inicial. Para construir el resto del documento se utiliza el patrón que contiene la expresión `match="/libreria/libro"`. Este patrón, y el resto del documento XSLT, son omitidos en la figura 6.1.

En el enfoque para almacenar información descrito en esta sección, los datos son almacenados en un archivo XML, el cual es desplegado en un browser usando un conjunto de reglas XSLT que indican cómo generar un archivo HTML desde el archivo XML original. Para sacar el mayor provecho a este enfoque, todavía nos falta indicar cómo se puede extraer información desde un documento XML. Esto se verá en la siguiente sección.

Extracción de información desde XML

En las secciones anteriores se mostró un enfoque para almacenar información en la Web en el cual los datos son almacenados en XML y mostrados a los usuarios en HTML (utilizando transformaciones escritas en XSLT). Se argumentó que éste era un buen enfoque porque permitía tener lo mejor de dos mundos: por una parte para un computador es más fácil interpretar información escrita en XML, y por lo tanto es más fácil extraer información desde este formato; y por otra parte HTML provee de buenas herramientas para desplegar información en la Web.

Para que el enfoque anterior pueda llevarse a cabo es necesario tener buenos lenguajes de consulta para XML. Estos lenguajes deben ser suficientemente expresivos como para permitir al usuario expresar consultas generales, y también deben estar acompañados de procedimientos eficientes para evaluar consultas. En esta sección se va a introducir XPath y XQuery, los dos lenguajes de consulta más populares para XML.

La primera versión estandarizada de XPath es de 1999 [4]. XPath puede ser considerado como el lenguaje de consulta más popular para XML, ya que forma parte de la mayor parte de los lenguajes de consulta para XML y, en particular, es parte de XQuery [1], como se verá más adelante. XPath provee una serie de herramientas que permiten navegar un documento XML, seleccionar elementos desde él y extraerlos para ser desplegados o usados por otras consultas. Una de las razones de la popularidad de XPath es que estas herramientas son simples de usar, y son lo suficientemente expresivas para poder manejar muchas de las consultas que los usuarios tienen en la práctica. Además, la estructura simple de este lenguaje ha permitido el desarrollo de procedimientos eficientes para evaluar consultas.

La mejor manera de entender XPath es a través de algunos ejemplos. Suponga que se está utilizando el documento XML con información sobre libros descrito en la sección 6.1, y que se ha utilizado repetidas veces en este capítulo. Si un usuario quiere extraer el nombre de la librería, entonces puede utilizar la siguiente consulta XPath:

```
child/?nombre/text()
```

Esencialmente una consulta en XPath consiste de un camino, y su respuesta es el conjunto de todos los elementos que pueden ser alcanzados en un documento XML, siguiendo el camino desde el primer elemento de este documento. En una consulta XPath se pueden utilizar palabras que tienen un significado reservado (`child` y `text()` en el ejemplo) o palabras cuyo significado está dado por un documento (`nombre` en el ejemplo). Además, en una expresión XPath se puede utilizar el símbolo `?` para indicar que se quiere chequear una condición. En el ejemplo, la palabra reservada `child` es utilizada para pasar de un elemento a sus hijos y `?nombre` indica que sólo se va a considerar los elementos con marcador `<nombre>`. De esta forma, utilizando la expresión `child` en el ejemplo se pasa de un elemento con marcador `<libreria>` a los que tiene marcadores `<nombre>` y `<titulo>`, y luego utilizando el test `?nombre` se selecciona el único elemento con marcador `<nombre>` hijo del elemento con marcador `<libreria>`. Finalmente se utiliza `text()` para extraer el texto almacenado dentro del elemento con marcador `<nombre>`, vale decir, `Todo Libros`.

Es importante destacar que para simplificar la presentación del lenguaje XPath, no se está usando aquí la sintaxis de XPath definida en [4], sino que una versión simplificada (pero que refleja la forma en que trabaja XPath).

Suponga ahora que se quiere extraer la lista de apellidos de todos los autores de libros. Para hacer esto, se puede utilizar la siguiente consulta:

```
descendant/?apellido/text()
```

La mayor diferencia con la consulta anterior es la utilización de la palabra reservada `descendant`, la cuál indica que se debe utilizar a los descendientes del primer elemento del documento, vale decir, a los elementos que son alcanzables utilizando los caminos `child`, `child/child`, `child/child/child`, etc. Nótese que esta consulta funciona incluso en casos en que la información sobre autores es dada de manera menos estructurada:

```
...
<primer_autor>
  <nombre>Martin</nombre>
  <apellido>Osborne</apellido>
</primer_autor>
<segundo_autor>
  <nombre>Ariel</nombre>
  <apellido>Rubinstein</apellido>
</segundo_autor>
...
```

En general, se considera una ventaja de XPath el que pueda funcionar sobre información semi-estructurada, ya que en la práctica la estructura de muchos documentos XML es irregular.

En este punto, el lector probablemente se ha dado cuenta de que la consulta anterior puede funcionar de manera incorrecta si el documento no sólo contiene apellidos de autores (por ejemplo, contiene los apellidos de la gente que trabaja en la librería). En ese caso se puede utilizar la consulta `descendant/?libro/descendant/?apellido/text()` que busca apellidos que aparezcan dentro de elementos con marcador `<libro>`.

Una de las limitaciones de XPath es la falta de herramientas para estructurar la información que se extrae; una consulta en XPath retorna un conjunto de elementos y no un documento XML. XQuery es un lenguaje más

completo, que usa XPath para navegar documentos XML y tiene herramientas para estructurar la información extraída como un documento XML [1]. En el siguiente ejemplo se muestra una consulta XQuery:

```
let $lib := doc("libreria.xml")
return
  <lista>
  {
    for $x in $bib/child/?libro
      for $y in $x/descendant/?apellido
        where $y/text() = Rubinstein
          return
            <libro>
            {
              <titulo> $x/descendant/?titulo/text() </titulo>
              <precio> $x/descendant/?precio/text() </precio>
            }
            </libro>
  }
  </lista>
```

Al igual que para el caso de XPath, en una consulta XQuery pueden aparecer elementos que tienen un significado predefinido y otros que deben ser interpretados en un documento XML. En la consulta anterior, `let` es utilizado para indicar que la variable `$lib` está ligada al documento `libreria.xml` (una variable en XQuery comienza con el símbolo `$`). Además, en esta consulta `for` es usado para indicar que una variable debe tomar todos los valores alcanzados al utilizar un camino en XPath. Por ejemplo, `for $x in $bib/child/?libro` indica que `$x` va a tomar como valor los elementos con marcador `<libro>` que son hijos del primer elemento del documento. Nótese que al igual que en un lenguaje de programación, las

instrucciones que utilizan `for` pueden aparecer anidadas. En la consulta anterior, `where` es usado para chequear una condición y `return` para indicar que algo debe estar en la salida de la consulta. Así, por ejemplo, en la condición `where $y/text() = Rubinstein` se chequea que el apellido del autor que se va a utilizar sea `Rubinstein`. Es importante destacar que en una consulta XQuery se puede indicar cómo se va a estructurar la respuesta colocando marcadores XML. En el ejemplo, `<lista>` es el marcador del primer elemento del documento de salida, y contiene como hijos una serie de libros con marcador `<libro>`.

Seguramente el lector ya se ha dado cuenta que la consulta anterior retorna la lista de libros escritos por Rubinstein con sus precios. Esta es una de las consultas que se planteó al principio de este capítulo, y para las cuales no era claro como responderlas si la información era almacenada en documentos HTML. Como se muestra en el ejemplo, si la información se almacena en formato XML, una simple consulta en XQuery puede bastar para extraer la información deseada. Incluso en el caso de la consulta más compleja vista al comienzo de este capítulo ("dé la lista de libros de Rubinstein que han bajado de precio en los últimos años"), una consulta en XQuery puede ser usada para extraer la información deseada.

Para recordar

¿Qué debería recordar el lector después de navegar por este capítulo? El lector debería estar satisfecho si la arquitectura presentada en la Figura 6.2 le resulta familiar.

En caso de que el lector no recuerde todos los componentes de la Figura 6.2, aquí damos un breve resumen de lo que se trató este capítulo. El lenguaje HTML es usado para indicar a un browser (tal como FireFox o Explorer) la

forma en que se debe desplegar la información. Aunque el resultado de desplegar esta información es fácil de entender para los usuarios (como vemos a diario en las páginas Web que visitamos), es, en general, difícil de entender para un computador. Para solucionar este problema, XML ha surgido como un lenguaje para almacenar información, que es de fácil procesamiento para un computador. Es importante destacar que XML no ha venido a reemplazar HTML, muy por el contrario se ha convertido en su complemento; la información se almacena en XML y se despliega utilizando HTML, lo que nos permite tener lo mejor de estos dos mundos. Una serie de tecnologías han sido desarrolladas para sacar el máximo de provecho al matrimonio entre HTML y XML. Por una parte, es necesario utilizar el lenguaje de transformación XSLT para poder desplegar como HTML información que es guardada como XML. Por otra parte, lenguajes de consulta tales como XPath y XQuery son utilizados para extraer y analizar información que es almacenada en XML.

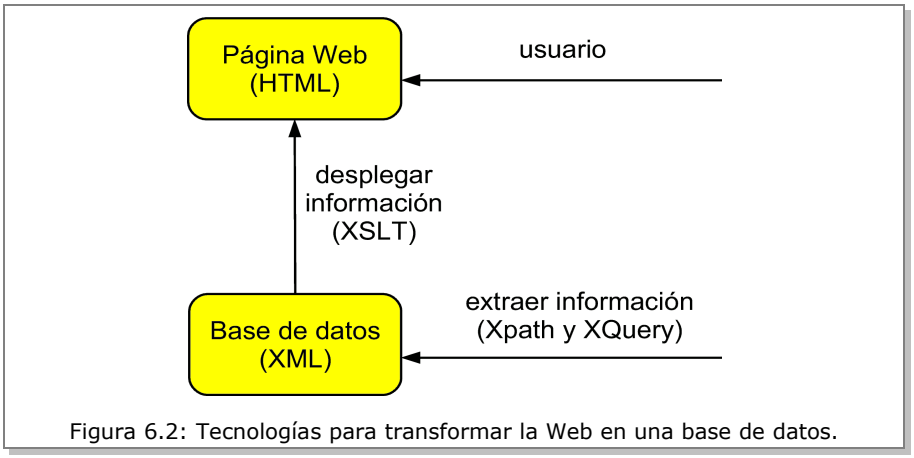


Figura 6.2: Tecnologías para transformar la Web en una base de datos.

Para saber más

- El sitio de la World Wide Web Consortium o simplemente W3C (http://www.w3.org/) es un buen lugar para informarse de los avances en las tecnologías Web como XML.
- El sitio http://www.w3schools.com/ tiene tutoriales sobre HTML, XML, XSLT, XPath, Xquery, etc.

Referencias

1. S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie y J. Siméon. XQuery 1.0: An XML Query Language. Recomendación de la W3C, enero 2007, http://www.w3.org/TR/xquery/
2. T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, F. Yergeau. Extensible Markup Language (XML) 1.0. Recomendación de la W3C, agosto 2006, http://www.w3.org/TR/2006/REC-xml-20060816/
3. J. Clark. XSL Transformations (XSLT) Version 1.0. Recomendación de la W3C, noviembre 1999, http://www.w3.org/TR/xslt
4. J. Clark y S. DeRose. XML Path Language (XPath) Version 1.0. Recomendación de la W3C, noviembre 1999, http://www.w3.org/TR/xpath

Capítulo 7
Uso y Búsqueda de Información Geográfica en la Web

Andrea Rodríguez

Si bien es cierto que en sus primeros tiempos la Web contenía esencialmente documentos textuales, hoy en día y en forma creciente la Web contiene también información en forma de imágenes, mapas, audio y videos. Esto amplía las posibilidades para que buscadores tradicionales incorporen nuevas facilidades en la búsqueda de información y formas de presentar los resultados de estas búsquedas. Un ejemplo de esto es el ya conocido Google Earth [4], el cual nos brinda la posibilidad de combinar imágenes satelitales, mapas, levantamientos de terreno o edificaciones en 3 dimensiones para poder entregar información referente a lugares específicos (ej. hoteles, hospitales, etc.), explorar información geográfica general en forma de videos o mapas (ej. paisajes, mapas de transporte, etc.) y compartir lugares de interés agregando información adicional (ej. fotos, notas, etc.).

Lo que hace que Google Earth no sea sólo un conjunto de imágenes es que éstas tienen la semántica dada por un dominio particular de información, en este caso, como información geográfica. Lo interesante es que si sabemos de qué trata la información, entonces podemos usar propiedades tí-

picas de su dominio que nos ayudan a conseguir una mejor búsqueda y recuperación de información. Consideremos un ejemplo sencillo en el que queremos encontrar hoteles en la ciudad de Pucón. Una búsqueda tradicional recuperaría los documentos que contienen las palabras hotel y Pucón, independientemente de que Pucón corresponda a una entidad geográfica que, por tanto, está cercana a otras localidades de características similares como Villarrica o Caburgua. Más aún, eventualmente un usuario podría realizar consultas que van más allá de la referencia de nombre de un lugar, ya sea por medio de la especificación de la relación de un lugar con otro (ej. hoteles cerca de Pucón) o bien estableciendo áreas geográficas de interés (ej. hoteles en la región de Los Lagos). Uno podría ir más lejos y tratar de combinar dominios de información como, por ejemplo, información geográfica y temporal, tal como sería el caso de consultar por acontecimientos ocurridos durante una cierta época y en una región determinada.

El objetivo de este capítulo es describir un caso concreto en el cual el dominio de información ha dado lugar a aplicaciones particulares en la Web. Este es el caso de información geográfica, para la cual describiremos su representación y uso en la Web.

¿Cuál es el tipo de información geográfica en la Web?

Existen diferentes formas de información geográfica en la Web (figura 7.1), las que podemos clasificar primariamente en tres tipos:

- Imágenes, en particular, imágenes satelitales que representan una vista de la superficie terrestre.
- Mapas digitales, en los que esencialmente se dibujan objetos en un espacio geográfico.

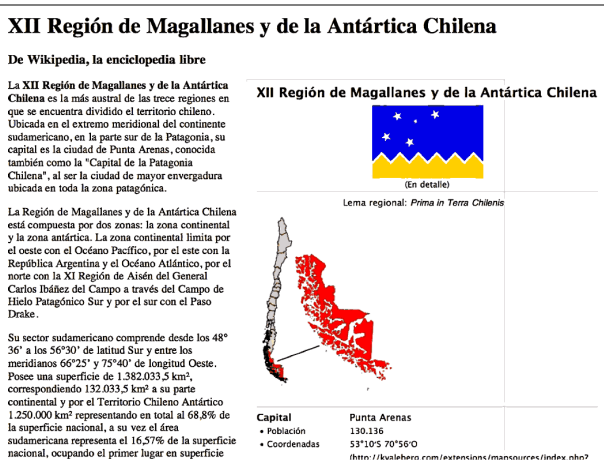
- Textos, en los que las referencias a localizaciones geográficas se dan, principalmente, mediante nombre de lugares y terminologías en lenguaje natural para las relaciones con otros lugares.

Con el primer y segundo tipo de información geográfica uno puede asociar información a la localización de un lugar por las coordenadas que describen su latitud y longitud. Tal tipo de información se utiliza típicamente en la visualización y manipulación de mapas a través de *servicios Web de información geográfica o Geo Web Services*. Google Maps [5] cae dentro de este tipo de servicios proveiendo una plataforma base de cartografía e imágenes satelitales que nos permiten situarnos en un punto particular del espacio.

El segundo tipo de información geográfica también representa elementos en el espacio geográfico aunque, implícitamente, mediante referencias que no están basadas en un sistema de coordenadas, sino que, más bien, se asocian a diferentes aspectos de un documento en la Web, específicamente:

- dónde fueron creados los documentos,
- de qué tratan o a qué se refieren los documentos,
- dónde residen los usuarios de los documentos.

A modo de ejemplo de estos tipos de referencias geográficas, un usuario podría requerir manejar las páginas de la Web Chilena o encontrar documentos que hagan referencia a Concepción o a alguna entidad geográfica relacionada a esa ciudad, o bien determinar los documentos que son usualmente visitados por usuarios ubicados en Concepción. El manejo de estas referencias geográficas han impulsado la extensión de las máquinas tradicionales de búsqueda, dando origen a las denominadas *máquinas de búsqueda Web geográfica* o *Geo Web Search Engines*.



XII Región de Magallanes y de la Antártica Chilena

De Wikipedia, la enciclopedia libre

La **XII Región de Magallanes y de la Antártica Chilena** es la más austral de las trece regiones en que se encuentra dividido el territorio chileno. Ubicada en el extremo meridional del continente sudamericano, en la parte sur de la Patagonia, su capital es la ciudad de Punta Arenas, conocida también como la "Capital de la Patagonia Chilena", al ser la ciudad de mayor envergadura ubicada en toda la zona patagónica.

La Región de Magallanes y de la Antártica Chilena está compuesta por dos zonas: la zona continental y la zona antártica. La zona continental limita por el oeste con el Océano Pacífico, por el este con la República Argentina y el Océano Atlántico, por el norte con la XI Región de Aisén del General Carlos Ibáñez del Campo a través del Campo de Hielo Patagónico Sur y por el sur con el Paso Drake.

Su sector sudamericano comprende desde los 48° 36' a los 56°30' de latitud Sur y entre los meridianos 66°25' y 75°40' de longitud Oeste. Posee una superficie de 1.382.033,5 km², correspondiendo 132.033,5 km² a su parte continental y por el Territorio Chileno Antártico 1.250.000 km² representando en total al 68,8% de la superficie nacional, a su vez el área sudamericana representa el 16,57% de la superficie nacional, ocupando el primer lugar en superficie

XII Región de Magallanes y de la Antártica Chilena

(En detalle)

Lema regional: *Prima in Terra Chilenis*

Capital	Punta Arenas
• Población	130.136
• Coordenadas	53°10'S 70°56'O

(a)

(b)

Figura 7.1: Tipos de información geográfica: (a) página Web con geo-referencias textuales y (b) servicio Web de mapas.

Servicios Web de información geográfica y máquinas de búsqueda pueden combinarse en un mismo sistema. Por ejemplo, uno puede pensar en tener información cartográfica que sirve de base de visualización de documentos textuales manejados por máquinas de búsqueda Web geográficas (ej. Google Earth).

Servicios web de información geográfica

Servicios Web de información geográfica (GWSs: Geospatial Web Services) son componentes modulares de aplicaciones que pueden ser publicadas, localizadas e invocadas a través de la Web, de modo de accesar y procesar datos de una variedad de repositorios de datos geográficos (figura 7.2). GWSs implementan tareas de procesamiento geográfico, tales como visualización cartográfica o planificación de rutas.

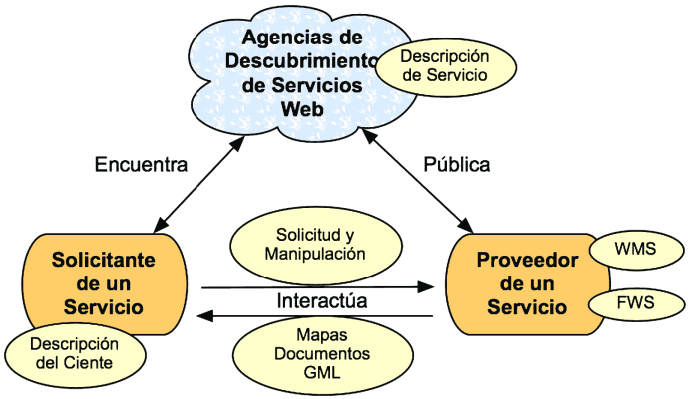


Figura 7.2: Componentes de una arquitectura de servicios Web de información geográfica compuesta de: el solicitante de información, el agente descubridor de los servicios que satisfacen a un solicitante y el proveedor de información tanto en forma de mapa (WMS) o en documentos de objetos geométricos o GML (FWS).

La tecnología de GWSs se basa en una serie de protocolos estándares derivados del XML a diferentes niveles de especificación, desde especificación de información geométrica hasta descripción de metadatos. Debido a que el número de GWSs disponibles hoy en día en la Web crece de manera rápida y continua, adicionalmente a los problemas intrínsecos de la gran diversidad en la forma de describir y representar información geográfica, *descubrir* los servicios que contienen los datos geográficos de interés entre todos los servicios disponibles es una tarea central para el desarrollo de GWSs. Comúnmente, el descubrimiento de servicios Web es del tipo sintáctico a través de interfaces estándares para una búsqueda basada en taxonomías o palabras claves.

En presencia de diferentes proveedores de información geográfica, sin embargo, es usual que una simple búsqueda sintáctica no permita un descubrimiento apropiado de información. Consideremos, por ejemplo, el caso de querer encontrar mapas que describan parques nacionales en el sur de Chile. Uno debería hacer una búsqueda por *parques nacionales* y por *el sur de Chile* (la relación *en* es generalmente eliminada en los buscadores tradicionales) o por una secuencia fija de caracteres *parques nacionales en el sur de Chile*. Esta búsqueda meramente sintáctica no podría considerar los siguientes aspectos de la semántica de la consulta:

- ¿Qué zonas incluye el sur de Chile?
- ¿Cuál es la semántica de la relación *en*? ¿Qué sucede si un parque nacional se sobrepone a dos zonas geográficas (centro y sur de Chile)?
- ¿Son los parques administrados en la región considerados nacionales o no?

Todos estos aspectos han inducido a nuevos trabajos dentro del contexto de *Web Semántica* denominado *Geo Web Semántica* [1,8]. La idea es crear

representaciones de los recursos de información geográfica, y asociar estos recursos con estructuras de representación formales que están siendo construidas en el contexto de la Web Semántica.

Máquinas de búsqueda Web geográfica

Máquinas de búsqueda Web geográfica (MBGs) surgen como una nueva forma de recuperar información de la Web que explota dos ideas básicas: (1) recursos en la Web tienen una referencia geográfica e (2) información geográficamente cercana es más relevante. La ideas de estas máquinas de búsqueda es que uno pueda preguntar por palabras claves y por una localidad geográfica, de manera que los resultados de una búsqueda sean documentos que hagan referencia a esa localidad o a alguna geográficamente cercana [7]. Así, los resultados pueden no sólo ser presentados como una lista de documentos en orden de relevancia, sino que visualmente como símbolos sobre un mapa cuyas ubicaciones indican la existencia de documentos que las referencian. Una arquitectura para estas máquinas extiende la arquitectura clásica de máquinas tradicionales (figura 7.3).

Tres aspectos importantes en la implementación de una MBG son: (1) cómo extraer las referencias geográficas y asociar un *geo coding* a los documentos Web, (2) cómo realizar el indexamiento de las páginas según su geo referencia y contenido de manera de agilizar su búsqueda por palabras claves y por nombre de lugares, y (3) cómo introducir en la relevancia (i.e. *ranking*) de los resultados el concepto de vecindad espacial o *proximidad* de los documentos respecto a una consulta de un usuario. Mientras los dos últimos aspectos incorporan a los mecanismos de indexamiento y relevancia de buscadores tradicionales nociones referente al manejo de información espacial, geo coding es una tarea particular de MBGs.

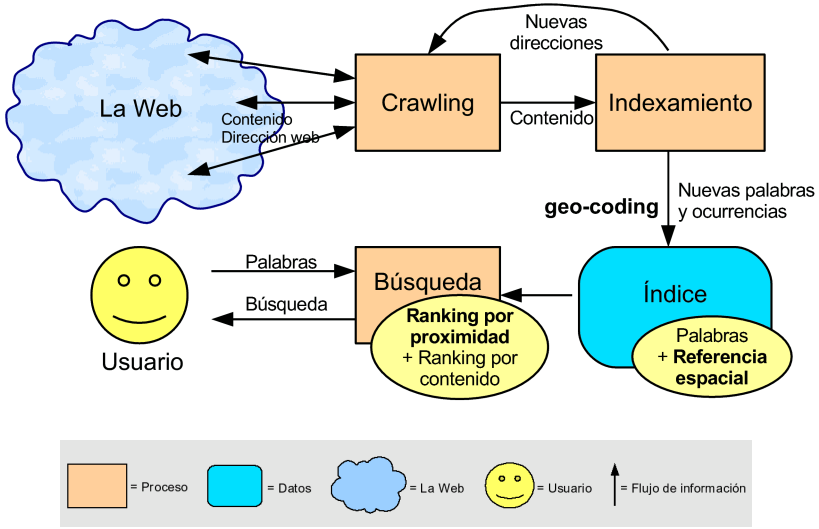


Figura 7.3: Una arquitectura para máquinas de búsqueda de información geográfica compuesta de: (1) crawling para recolección y segmentación de páginas Web en texto y geo referencias, (2) indexamiento para generar un índice que agilice la búsqueda en base a términos y geo referencias, y (3) la búsqueda que recupera y selecciona documentos en base a un ranking por contenido (texto) y proximidad (espacio).

Extraer las geo referencias de una página Web no es una tarea fácil. Una de las formas más simples de geo referenciar una página Web es considerar la localización del servidor que la contiene como lugar de referencia. Otra forma de asociar geo referencias a documentos Web es agregando información de metadatos geo-espaciales (geo tags), denotando que el contenido de la página Web es relevante para cierta localización. Esta localización puede ser descrita usando protocolos estándares basados en XML tales como el GeoRSS [6]. Finalmente, otras técnicas realizan un *parsing* o

segmentación de documentos completos para extraer nombres de lugares, como ciudades o regiones.

Actualmente, el desarrollo de máquinas de búsqueda Web geográficas presenta grandes desafíos para la investigación. Algunos de estos temas son:

- Desarrollo de prototipos de buscadores a gran escala. Esto involucra consideraciones de rendimiento y escalabilidad para la recolección, indexamiento y búsqueda de páginas con geo coding.
- Técnicas que combinen procesamiento de lenguaje natural, data mining, análisis de enlaces entre documentos y estructura de documentos para obtener mejoras en cuanto a geo coding.
- Crawling y ranking que incorporen nociones de localidad. Así, por ejemplo, uno podría analizar la estructura espacial y de conexión entre las páginas Web y detectar páginas que sean más enlazadas globalmente (ej. en el mundo) o localmente (ej. en el país). Luego páginas con referencia local serían más relevantes en el contexto de una búsqueda con referencia a esa localidad.
- Procesamiento de consulta y diseño de interfaces de búsqueda espacial. En este sentido, analizar estrategias de ranking de resultados que permitan una mejor visualización de éstos de manera de reflejar tanto su similitud espacial y de contenido.
- Mining de datos geográficos en la Web. Temas de mining de datos geográficos refieren al análisis (clasificación, reconocimiento de patrones, agrupamiento) de la estructura, uso y contenido de las páginas Web en base a un criterio espacial. Por ejemplo, analizar la distribución geográfica de las personas que se conectan a ciertos sitios y el enlace entre páginas geográficamente distribuidas.

Para saber más

- La presentación "A Spatial Dimension for Searching the World Wide Web" http://www.ciw.cl/recursos/andreaHIS2002.pdf trata temas relacionados con la búsqueda geo-espacial.
- El libro "The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society" trata estos temas. Se puede ver un capítulo de ejemplo en: http://www.geospatialweb.com/
- Wikipedia tiene una entrada para la Geoweb: http://en.wikipedia.org/wiki/Geoweb

Referencias

1. Max Egenhofer (2002). Toward the Semantic Geospatial Web, ACM GIS: Symposium on Advances in Geographic Information Systems, ACM Press, pp. 1-4.
2. ISO/TC. ISO/TC 211 Geographic information/Geomatics. URL: http://www.isotc211.org/
3. OpenGis. OpenGIS' Specifications (Standards). URL: http://www.opengeospatial.org/standards
4. Google Earth. URL: http://earth.google.com/
5. Google Maps. URL: http://maps.google.es/
6. GeoRSS: Geographically encoded Objects for RSS feeds. URL: http://www.georss.org/
7. Andrea Rodríguez (2002). A Spatial Dimension for Searching the World Wide Wb, Hybrid Intelligent Systems, IO Press, pp. 583-592.
8. Frederico Fonseca and Andrea Rodríguez (2007). From Geo-Pragmatics to Derivation Ontologies: New Directions for the GeoSpatial Semantic Web. Transactions in GIS 11(3): 313-316.

Capítulo 8
Multimedia en la Web

Javier Ruiz del Solar

El universo creciente de la información mutimedial en la Web

El mundo digital en el cual estamos inmersos genera un número inmenso y siempre creciente de datos digitales, que cada día es más difícil de administrar. Cámaras fotográficas, cámaras de video digital, audio digital, televisión digital, Internet (mensajes instantáneos, E-mails, etc.), música y videos disponibles en la Web son algunas de las principales fuentes de estos datos digitales. De acuerdo al estudio *How much information?* [1] en el año 2002 cinco exabytes de información fueron creados en el mundo (1 exabyte corresponde a 10^{18} bytes ó 10^{9} gigabytes!), 92% de esta información fue almacenada en discos duros y alrededor de 1,75% se hizo accesible a través de la Web. No se cuenta con datos más recientes, pero se estima que la cantidad de información generada crece a tasas mayores al 30% anual.

De esta forma la Web se está transformando en una base de datos multimedial[4] gigantesca. Sin embargo, esta información almacenada en la Web

4 La palabra multimedia viene de unir las palabras del latín multum (múltiples, muchos) y medium (medios), osea, significa que la información multimedial proviene de múltiples medios como texto, audio, gráficos, fotografías, videos y ani-

es útil siempre y cuando poseamos los mecanismos necesarios para encontrar la información que requerimos, por ejemplo, la canción o la fotografía que necesitamos en un determinado momento. Puede suceder que a pesar de que la imagen o canción se encuentre en la Web, no seamos capaces de encontrarla. Es tal el tamaño de la Web que son necesarias herramientas que nos ayuden en las labores de búsqueda. Estas herramientas deben ser más sofisticadas que las empleadas en los buscadores tradicionales (por ejemplo, Google o Yahoo!), las cuales fueron diseñadas para búsqueda de texto, no de datos multimediales.

Para simplificar el proceso de búsqueda de información multimedial se requiere que en el momento en que la información sea hecha pública en la Web, ésta sea correctamente clasificada o anotada. Es decir, a la información que se almacenará se le debe dar una descripción adecuada, generalmente textual (un nombre o una frase) que permita que la información pueda luego ser recuperada. Cuando la información a ser hecha pública es un objeto multimedial conocido, una canción o una película, la información del título es suficiente para que éste pueda ser encontrado fácilmente. Piense por ejemplo en las canciones disponibles en el sitio `itunes` [2].

Sin embargo, cuando el objeto multimedial no es conocido, no es fácil determinar cuál es el mejor texto que lo describe. ¿Cómo podríamos anotar adecuadamente las miles de fotografías digitales que tenemos almacenadas en los discos duros de nuestros computadores, o las horas de video digital almacenadas en cintas de video, o la información generada por todos los canales de noticias del mundo, en caso que se quisieran dejar disponibles en la Web?

maciones. Usualmente la información de texto puro no se considera contenido multimedial.

El problema no es solamente temporal, es decir, de tener el tiempo suficiente para realizar las anotaciones, sino de cómo describir el contenido de una cierta fotografía o video. Distintas personas generarán distintas descripciones en distintos instantes de tiempo. Por ejemplo, una persona de nombre Juan puede sacarse una fotografía durante sus vacaciones en Isla de Pascua. En la fotografía aparece Juan con su amiga María, una playa y una embarcación de nombre Anakena. ¿Cuál es la mejor descripción o anotación para esta fotografía? ¿Juan? ¿Juan en vacaciones?, ¿Juan y María? ¿Anakena? ¿pareja en la playa? ¿vacaciones en Isla de Pascua? ¿playa? ¿mar? ¿arena? ¿embarcación en la playa? Obviamente todas estas anotaciones podrían utilizarse, sin embargo, al momento de publicar la fotografía no puede saberse cuál es la mejor descripción. La mejor dependerá de qué se esté buscando y de quién realice la búsqueda. El problema obvio es que la anotación y la búsqueda de la fotografía suceden en distintos instantes, por lo que al anotar la imagen no se conoce los requerimientos de sus futuras operaciones de búsqueda.

Una segunda alternativa para anotar objetos multimediales consiste en usar categorías fijas como las que por ejemplo usa el sitio YouTube [3] para clasificar sus videos. El problema en este caso es que a la fotografía o al video a ser publicado se le debe asignar una cierta categoría fija. Los problemas son obvios: el objeto multimedial a ser clasificado puede caer en más de una categoría y la determinación de la categoría depende del ser humano que realice la categorización.

Una tercera alternativa para realizar las anotaciones es permitir que un sistema computacional puedan realizar las anotaciones en forma automatizada (sin intervención humana) y por lo tanto objetiva. En este caso pediremos al sistema computacional que genere una descripción del objeto multimedial a ser anotado. Esta descripción se usará posteriormente como

un índice, por medio del cual se podrá buscar al objeto en cuestión. Al usar índices, la tarea de anotar se denomina indexación[5]. Se le agrega el adjetivo automatizada para enfatizar el hecho de que esta labor se realiza sin intervención humana.

Indexación automatizada de la información multimedial

La indexación automatizada de información mutimedial trae consigo múltiples beneficios como ahorro en tiempos de búsqueda y estandarización en las anotaciones de las imágenes. Como fue mencionado anteriormente, los seres humanos realizan anotaciones o clasifican las imágenes de acuerdo a criterios propios. Además, cuando un humano anota una imagen se preocupa obviamente de las necesidades de búsqueda presentes y no piensa en las futuras. Esto provoca que la información que no es anotada en el presente, no pueda ser buscada en el futuro. Sin embargo, la indexación automatizada permite anotar en forma objetiva la mayor cantidad de características posibles y de esta forma anticiparse a las necesidades futuras de búsqueda del usuario.

Existen indexaciones de tipo texto (cada imagen recibe una descripción textual explícita, i.e. una frase que la representa), por atributos (cada imagen es descrita mediante una especificación de ciertos atributos que contiene, ej. texturas) o por contenido (la forma, el color o alguna otra característica de los objetos que contiene la imagen es utilizada en forma implícita para su indexación). En cada uno de estos caso el objeto multimedial se almacena junto a su descriptor en la base de datos.

5 Se utilizan también los términos indización (de índice) e indexamiento (del inglés index).

En la figura 8.1 se muestra la fotografía de una mujer sentada en un jardín, y dos posibles tipos de indexación, usando descriptores de texto y color. En el caso de los descriptores de texto, la entrada al sistema es una descripción textual entregada por un humano, y el módulo de indexación automatizada (en rigor semi automatizada en este caso) determina una versión reducida del descriptor de texto usando un algoritmo de stemming, que filtra algunas palabras (ejemplo: artículos, conjunciones) y reduce otras a su raíz (verbos y sustantivos). En el caso de los descriptores de color se muestra el uso de histogramas de color RGB, que permiten calcular estadísticas del contenido cromático de la imagen en los canales rojo (R: red), verde (G: green) y azul (B: blue).

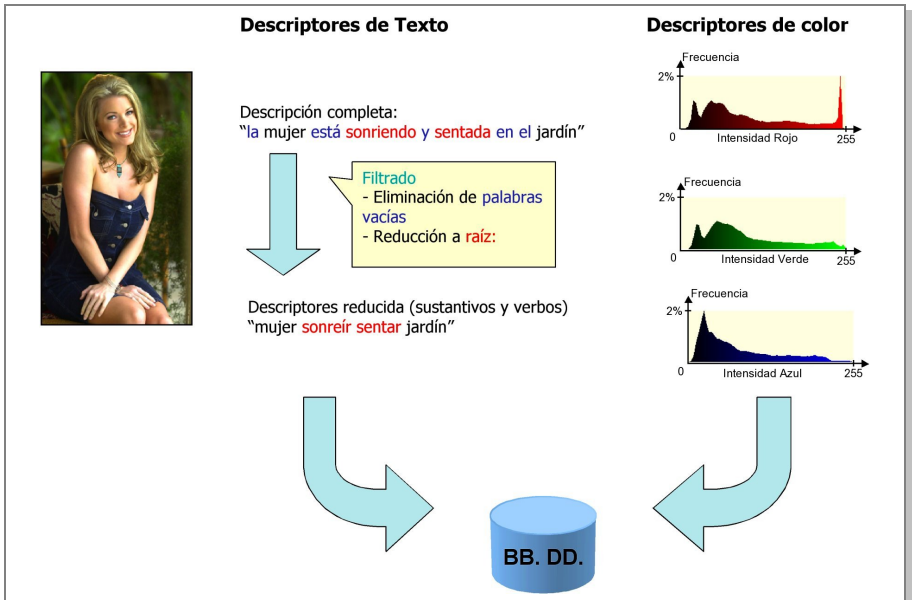


Figura 8.1: Ejemplo de cálculo de descriptores de texto y de color sobre una imagen.

Búsqueda o Recuperación de información multimedial

Las operaciones de búsqueda de información multimedial, también conocidas como recuperación[6] de información multimedial, se realizan utilizando los descriptores almacenados en la base de datos junto a los objetos multimediales[7]. El sistema ideal de búsqueda debiera, a partir de una descripción textual en lenguaje natural del contenido de una imagen, encontrar todas aquellas imágenes que corresponden a dicho contenido, sin importar como éstas fueron anotadas. Por ejemplo, "imágenes con perros", "imágenes del hundimiento de un barco", "imágenes de Isla de Pascua", "fotos de mi mamá", etc. Sin embargo, este sistema ideal no es realizable en la actualidad. Si las imágenes fueron anotadas usando descriptores textuales es muy poco probable que las personas que anotaron dichas imágenes hayan utilizado los mismos descriptores textuales usados en la operación de búsqueda. Si las imágenes fueron anotadas utilizando un sistema de indexación automatizada en base a su contenido de bajo nivel (color, bordes, textura, etc.) es difícil encontrar la adecuada correspondencia entre los descripciones textuales de alto nivel utilizados por los seres humanos, para describir el contenido de las imágenes, y las descripciones de bajo nivel utilizadas por los sistemas computacionales de recuperación de imágenes. Este problema se conoce como el gap semántico existente entre las descripciones de bajo y alto nivel [4].

Debido al no resuelto problema del gap semántico, en la actualidad debe utilizarse el mismo tipo de descriptores tanto para anotar como para

6 En la literatura científica de habla inglesa esta operación se conoce como *retrieval*.

7 Recordemos que en nuestro caso estamos interesados en bases de datos accesibles a través de la Web.

recuperar las imágenes. De esta forma existen dos tipos de sistemas de indexación principales: aquellos basados en descriptores de texto y los basados en descriptores de contenido de bajo nivel, extraídos de imágenes de ejemplo.

Sistemas de búsqueda en base a anotaciones textuales. Se utilizan descriptores de texto, y el problema de búsqueda o recuperación de un objeto multimedial se reduce a la comparación entre el descriptor de texto que define la operación de búsqueda y los descriptores de texto almacenados en la base de datos (figura 8.2). Como fue anteriormente explicado, los problemas con estos métodos son: (i) distintos seres humanos realizan distintas descripciones (anotaciones) de una misma imagen. (ii) Las anotaciones de una imagen están relacionadas a la relevancia de los objetos y personas que se encuentren en ella. Pero la relevancia del contenido puede cambiar en el tiempo. Por ejemplo, previo al escándalo Lewinski, en la imágenes de video almacenadas en la Casa Blanca nadie hubiera anotado la presencia de Lewinski. Luego del escándalo sus imágenes se hicieron relevantes.

Sistemas de búsqueda por contenido en base a ejemplos. Dada una imagen de ejemplo el sistema de búsqueda retorna imágenes parecidas en contenido al ejemplo. Como paso intermedio el sistema extrae en forma automatizada un descriptor del contenido de bajo nivel de la imagen, el cual es comparado con los descriptores de bajo nivel almacenados en la base de datos (figura 8.2). Mediante este tipo de sistema, las imágenes que retornan son parecidas a las del ejemplo. De esta forma, si la imagen del ejemplo contiene una puesta de sol, el sistema retorna imágenes de puestas de sol; y si la imagen del ejemplo contiene árboles, el sistema retorna imágenes de árboles. Como este tipo de sistemas tiene por objetivo la recuperación de imágenes parecidas a la de ejemplo, la comparación entre los descriptores se traduce en la determinación de la similitud de estos. Algunas medidas de similitud comúnmente

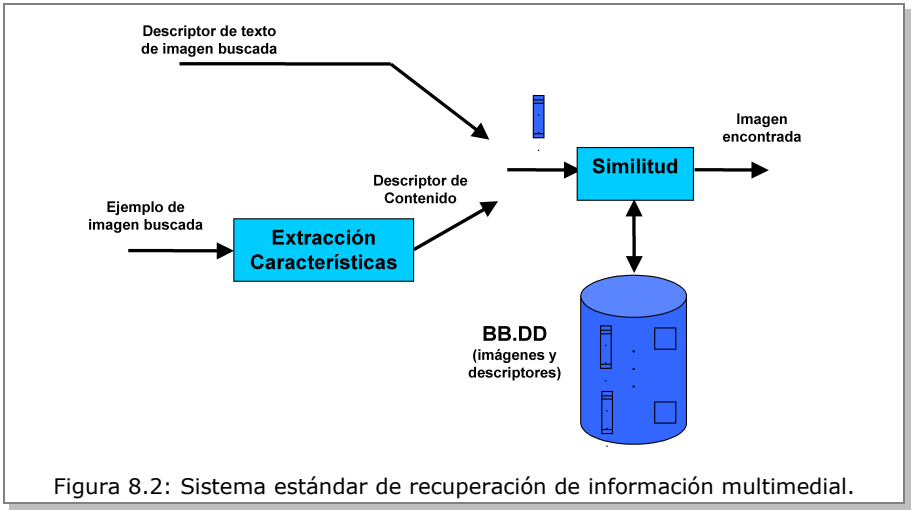


Figura 8.2: Sistema estándar de recuperación de información multimedial.

usadas son de distancia vectorial (Euclidiana, de Mahalanobis, etc.) y correlación. Entre los descriptores de bajo nivel más utilizados se encuentran aquellos basados en información de color (e.g. histogramas de color), texturas (e.g. matrices de co-ocurrencia) y bordes (e.g. histogramas de bordes direccionales).

Finalmente, cabe señalar que también existen sistemas de búsqueda en base a categorías. En este caso el usuario selecciona una categoría y el sistema de búsqueda retorna objetos multimediales correspondientes a esa categoría (“seres humanos”, “animales”, “Chile”, “África”, “deportes”, etc.). Tal como fue señalado anteriormente, la principal limitación de estos sistemas es la rigidez del sistema de categorización, y que el usuario debe navegar por un sinnúmero de categorías y sub-categorías hasta encontrar el objeto multimedial requerido.

Para saber más

- Alejandro Jaimes, Javier Ruiz-del-Solar, R. Verschae, Dinko Yaksic, Ricardo Baeza-Yates, Emilio Davis, Carlos Castillo. "Búsqueda por Contenido Visual: TREC 2003 y la Web Chilena." Presentación. CIW/DCC/DIE - Universidad de Chile. http://www.ciw.cl/recursos/uchile_talk_june26.pdf

Referencias

1. How Much Information? 2003 Project Web site. Berkely University. Disponible el 23 de abril de 2007 en http://www.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm
2. itunes Web site. Disponible el 23 de abril de 2007 en: http://www.apple.com/itunes/
3. YouTube Web site. Disponible el 23 de abril de 2007 en: http://www.youtube.com/
4. Intervalo Semántico. Wikipedia la Enciclopedia Libre. http://es.wikipedia.org/wiki/Intervalo_semántico

Capítulo 9
Redes Sociales

Javier Velasco

Vivimos en sociedad; dependemos de otras personas para gran parte de lo que hacemos diariamente. Durante nuestra vida formamos relaciones con personas que a su vez se relacionan con otros a quiénes no conocemos. Estos vínculos van formando nuestra red social.

Podemos ver ejemplos de redes sociales en espacios tan cercanos como nuestras familias, tan organizados como nuestros trabajos, así como redes formales e informales generadas en torno a intereses comunes como deportes, colecciones, o religiones. Otra fuente de redes sociales son los lugares de estudio y los espacios de colaboración entre investigadores y artistas, así como el barrio donde vivimos.

El valor de una red social radica en que se construye sobre la base de la confianza; nos permite llegar a personas que de otra manera no podríamos contactar, dado que una recomendación personal de un conocido mutuo genera confianza.

Análisis de Redes Sociales

El *Análisis de Redes Sociales* se ha venido desarrollando como una especialidad desde los años 60's desde diferentes disciplinas de las Ciencias Sociales, con el apoyo de una rama de las matemáticas llamada *Teoría de*

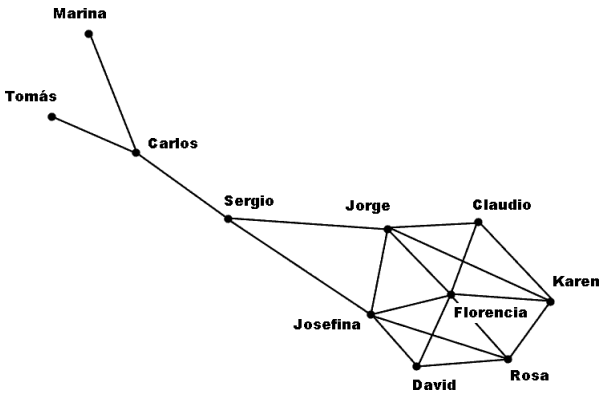


Figura 9.1: Ejemplo de Grafo para una Red Social compuesta por 11 personas

Grafos. Esta última permite cuantificar los vínculos entre las personas que pertenecen a una red social y analizar la estructura de dicha red. En base a la teoría de grafos, el análisis de redes sociales define a las personas como *nodos*, y las relaciones entre éstas como *aristas*.

El análisis de redes sociales se basa en la idea de que *la relación entre las personas* es más importante que *sus características individuales*, es por esto que su estudio se ha desarrollado en términos matemáticos abstractos y representa un enfoque alternativo al estudio tradicional de organizaciones sociales, donde las características individuales son lo primordial.

El análisis de redes sociales se enfoca en la estructura de estas redes, y su unidad de análisis es *la relación entre dos personas*. Las relaciones fuertes entre personas, por ejemplo un matrimonio, conforman *aristas fuertes*. Las *aristas débiles* muchas veces tienen mayor importancia que las *fuertes*, ya que proveen un atajo entre personas que de otra forma no estarían conectadas, generando así mayores oportunidades de exposición a nuevas ideas y gru-

pos de *influencia* [4]. Las asociaciones profesionales son un buen ejemplo donde las aristas débiles pueden resultar sumamente útiles para el desarrollo profesional de una persona.

Parte del análisis en la estructura de estas redes implica determinar la importancia de determinado nodo para el conjunto. Las medidas más comunes para determinar esta importancia son [2]:

Centralidad (Degree Centrality): Dependerá de la cantidad de aristas que conectan a una persona en el conjunto. Los nodos más conectados son más centrales. En el ejemplo de la Figura 9.1, si consideramos un grafo dado por el subconjunto de cinco personas formado por: Jorge, Karen, Josefina, Rosa y Florencia, *Florencia* sería la persona más central.

Cercanía (Closeness Centrality): Depende de la longitud de suma de las aristas que conectan a una persona con todas las demás. Aproxima su "peso"; su capacidad para llegar en pocos pasos a cualquiera. En el grafo del ejemplo (figura 9.1), *Jorge* y *Josefina* tienen el mayor grado de cercanía.

Intermediación (Betweenness Centrality). Es una medida del número de veces que un nodo aparece en el camino más corto entre otros dos nodos. La intermediación nos da una aproximación al peso como conector (como *hub*) del nodo, su importancia para que la red se mantenga unida. En el grafo del ejemplo (figura 9.1), *Sergio* tiene el mayor grado de intermediación.

El estudio en la forma de las redes sociales permite determinar la utilidad de éstas para los individuos que las conforman, así como su dinámica. Por ejemplo, el flujo de influencia dentro de una compañía más allá de los roles de trabajo. Este enfoque ha permitido importantes desarrollos en el estudio de las redes sociales que integramos en la vida diaria, y ha sido muy valioso en el estudio de la difusión de enfermedades contagiosas [4].

Redes Sociales y Software

Las redes computacionales surgen como una forma de potenciar la comunicación en redes humanas de trabajo. El desarrollo de Internet y el aumento en la capacidad computacional de los servidores ha permitido el desarrollo de diversos formatos para redes tecnológicas que soportan el funcionamiento de redes sociales y la construcción de nuevas redes sociales.

Las redes sociales computacionales están permitiendo a las personas crear nuevas dinámicas de comunicación más potentes que las anteriores. Diferentes estudios han descubierto que estas redes de software permiten a las personas tanto fortalecer sus redes sociales actuales como formar nuevas redes de manera efectiva [1]. Hoy en día podemos analizar prácticamente cualquier servicio o sistema de Internet en términos de redes sociales.

Todas estas redes tecnológicas serían inútiles si no se nutrieran de la comunicación entre las personas; incluso hay quienes han expresado esto en términos matemáticos. Bob Metcalfe, uno de los inventores de Ethernet, determinó, en relación con las redes de telecomunicaciones, que *la utilidad de la red crece en relación al cuadrado de la cantidad de usuarios conectados* (Ley de Metcalfe). Posteriormente, David Reed descubrió que esta fórmula quedaba corta para describir Internet, ya que además de permitir comunicaciones persona a persona permite la creación de grupos. La ley de Reed postula que *la utilidad de una red, en particular las redes sociales, crece en forma exponencial a la cantidad de personas que la integran* [6].

Aplicaciones de Redes Sociales

A continuación describimos algunos de los ejemplos más comunes de redes sociales mediatizadas por software:

1. Correo electrónico (E-mail): Es una herramienta omnipresente que permite comunicarnos tanto con nuestra familia, amigos, compañeros de trabajo, como con los líderes de opinión y autoridades que antes resultaban inalcanzables. Un análisis de nuestro uso del correo y nuestra libreta de direcciones permitiría generar una imagen de nuestras redes sociales: la frecuencia de los mensajes revelaría la fuerza de las aristas en nuestra red. Las listas de correo electrónico también conforman redes sociales organizadas en torno a temas particulares.

2. Mensajería Instantánea (IM): La comunicación sincrónica que estos sistemas permiten implica una fuerza todavía mayor en las relaciones de las que supone el correo electrónico. La estructura básica de estos sistemas es la lista de contactos, donde organizamos a las personas más relevantes de nuestra red social con quienes queremos (o debemos) estar comunicados de manera permanente. Este grupo comprende herramientas de mensajería por texto (ICQ, MSN, AIM, Y!IM) así como las más recientes de voz y video (Skype, gTalk).

3. La Web Mundial (WWW): Los sitios que componen esta red pertenecen a personas individuales, o bien a organizaciones de éstas. Un análisis de los links entre sitios Web nos podría dar señales interesantes acerca de las relaciones entre las personas o instituciones que los publican.

Existen sitios dedicados a organizar vínculos entre personas relacionadas a determinados temas. Por ejemplo, el portal de Yahoo! ha sido desde sus primeros días un *hub* de conexión hacia diferentes puntos de la WWW.

4. Comunidades En-línea: Algunos Sitios Web permiten crear un perfil personal e ir agregando una lista de contactos para participar con ellos en diferentes formas. Linkedin es un sitio de conexiones profesionales que permite generar recomendaciones laborales de gran credibilidad. E-conozco

es una aplicación similar en Español. Orkut permite organizar redes de contactos y grupos para la participación en foros. Fotolog permite a millones de personas en el mundo publicar sus fotografías y seguir las fotos de sus amigos.

4.1 Facebook: Es una comunidad online cerrada, donde los perfiles pueden editarse para ser visibles sólo a tus amigos. Se presenta como una plataforma sobre la que se pueden montar aplicaciones para que las personas se comuniquen en variados formatos. Facebook goza de gran popularidad y crecimiento, y ha tenido una importante penetración en Chile a partir de fines de 2007. Facebook permite, a través de la opción ver amigos comunes, explorar el fenómeno del *mundo pequeño* en tus redes sociales. Ver Cap. 2.

5. Blogs: Estas bitácoras personales cuentan con diferentes tipos de conexión hacia otras personas. Un blog cita a otro como fuente de información, puede opinar acerca de lo que otro ha publicado, y muchos blogs muestran explícitamente una lista de sus blogs relacionados o amigos. Son tres formas de describir relaciones entre los blogs que revelan y crean redes sociales [5]. El análisis de estos links y su frecuencia dará cuenta de la red social entre sus autores.

5.1 La coctelera: Es un sistema de blogs con sede en España que cuenta con varias de las herramientas para fomentar la interacción de los usuarios, lo que la transforman en un buen ejemplo de red social.

En el perfil de usuario de La Coctelera (figura 9.2) podemos ver una variedad de elementos:

Reseña del autor: permite incluir tu nombre, ubicación (ciudad y país), una fotografía y una pequeña descripción. Esta reseña permitirá a los nuevos visitantes conocer la información básica del autor.

Amigos, Ídolos y Fans: La Coctelera clasifica a los contactos en base a la direccionalidad de las referencias. Si pones a una persona en tu lista de contactos y éste no te corresponde, pasa a ser un *Ídolo*. Si la referencia es recíproca se categoriza como *Amigo*. Las personas que te han señalado entre sus contactos sin que tú los incluyas pasan a ser tus *Fans*.

Últimos Post: El perfil incluye un resumen de tus últimos posts, señalando sus horarios y comentarios.

Últimos Comentarios: Hace un seguimiento a la actividad de los comentarios en el blog.

Lo Más Comentado: Lleva una estadística de los posts con mayor cantidad de comentarios.

Habla de: Hace un seguimiento a los *tags* más frecuentes de cada blog (el uso de tags se explica más adelante). Estos tags permiten navegar a los posts asociados a estos términos tanto del mismo autor, el conjunto de usuarios de La Coctelera, y una búsqueda general en Technorati, un buscador especializado de blogs.

6. Clasificación Social (Folksonomies): Las bibliotecas utilizan palabras clave como un elemento crítico en sus sistemas de clasificación de documentos. Algunos sitios Web permiten a cualquier usuario agregar una palabra clave o descriptor a determinado objeto en su colección. De esta manera, son los mismos usuarios quienes organizan los elementos del sistema, tanto de manera individual como colectiva. El sistema genera automáticamente links para todos estos tags, lo que permite a los usurios navegar el sistema con gran flexibilidad. Ver Cap. 10.

Algunos de los ejemplos más relevantes de sitios Web con clasificación social son:

Figura 9.2: Ejemplo de Perfil de Usuario de La Coctelera

6.1. Del.icio.us: Es un sistema de favoritos sociales, en el que un usuario crea su cuenta y comienza a marcar sus páginas favoritas, añadiendo tags descriptores a cada recurso. Estos tags permitirán la navegación por entre los favoritos del total de usuarios en el sistema, generando un sistema de organización colectivo de los recursos. Del.icio.us también permite recopilar una lista de usuarios relacionados (o contactos), y explorar los favoritos de éstos.

6.2. Flickr: Este sitio es básicamente un fotolog con múltiples herramientas de interacción social. Al publicar una foto, el autor puede asignar tags a ésta para la exploración del espacio colectivo tal como en del.icio.us. También ofrece la administración de una lista de contactos con diferentes grados de relación: contacto, amigo y familia. Flickr permite la creación y participación en torno a grupos de interés de acuerdo a reglas fijadas por un moderador. Los usuarios de flickr pueden añadir tags a las fotografías que visitan, marcarlas como favoritos y dejar comentarios. Flickr soporta múltiples formatos de interacción entre las personas, y da pie a diferentes tipos de estudio a la red social que se ha ido formando entre sus usuarios que actualmente superan los dos millones.

7. Filtros Colaborativos: El análisis de actividad en una web mediante minería de datos revela patrones de comportamiento y hace posible generar sistemas de recomendaciones personalizadas que se ajustan a las preferencias particulares de una persona de acuerdo a la actividad del universo de usuarios del sistema. Amazon fue una de las primeras webs en explotar esta técnica para recomendar libros. Cuando uno visita la ficha de un libro en Amazon.com, el sistema sugiere recomendaciones personalizadas de acuerdo a las características del libro y el historial del usuario en el sitio, lo que considera compras y revisiones anteriores, en un análisis cruzado con la actividad de otros usuarios. El sistema va detectando personas con intereses

comunes para generar de esta manera las recomendaciones personalizadas. Mientras mayor sea el historial de tu perfil, mayor será la precisión de las recomendaciones. Ver Cap. 10.

8. Redes P2P: *Peer-to-peer* (inglés) significa conexión entre pares. Estas redes son descentralizadas, no cuentan con un servidor central en su distribución, sino con un conjunto de nodos de igual relevancia [4]. En la práctica las redes P2P se presentan en diferentes formas: unas son puramente distribuidas y otras se apoyan en servidores centrales para realizar sus funciones.

El uso más popular de las redes P2P ha sido el intercambio de música y videos entre personas particulares. Estos intercambios de música y películas muchas veces implica la violación a derechos de autor de dicho material, lo que ha significado problemas legales para las los creadores del software que han diseñado estas redes. Actualmente, las redes P2P han encontrado la forma de desligarse de la responsabilidad legal por tales violaciones. Algunos ejemplos populares de estas redes son: Napster, Kazaa, Gnutella, BitTorrent.

Sitios y Aplicaciones Mencionados

- AIM (http://www.aim.com/)
- Amazon (http://www.amazon.com/) tienda gigante de comercio electrónico, la que comenzó como una librería pronto se transformó en la tienda más grande del mundo. Actualmente vende toda clase de productos.
- BitTorrent (http://www.bittorrent.com/) es un protocolo P2P para el intercambio de archivos.

- Del.icio.us (http://del.icio.us/) sistema de bookmarks (o favoritos) sociales que utiliza clasificación social mediante tags.

- E-conozco (http://www.xing.com/econozco) es una red social laboral en castellano, que ahora es parte de Xing, una red social laboral internacional.

- Facebook (http://www.facebook.com/) comunidad online con múltiples formatos de comunicación, permite el desarrollo de aplicaciones que se montan sobre ésta.

- Flickr (http://www.flickr.com/) un fotolog con múltiples herramientas de interacción social. Actualmente es propiedad de Yahoo!.

- Fotolog (http://www.fotolog.com/) aplicación social de fotos, muy popular en Chile.

- gTalk (http://www.google.com/talk/) mensajería instantánea y voz de Google.

- Gnutella es una red P2P totalmente distribuida, que permite el intercambio de archivos.

- ICQ (http://www.icq.com/) fue uno de los primeros sistemas de mensajería instantánea (IM).

- Kazaa (http://www.kazaa.com/) es una aplicación P2P semidistribuida que permite el intercambio de música, videos, software y todo tipo de archivos.

- La Coctelera (http://www.lacoctelera.com/) es un sistema de blogs con múltiples herramientas de interacción social.

- Linkedin (http://www.linkedin.com/) comunidad social de perfiles laborales.

- MSN Messenger (http://im.live.com/Messenger/IM/Home/)

- Napster (http://free.napster.com/) uno de los más populares sistemas P2P.

- Orkut (http://www.orkut.com/) es una comunidad online, propiedad de Google.

- Skype (http://www.skype.com/)

- Technorati (http://www.technorati.com/) buscador especializado en blogs.

- Y!IM (http://messenger.yahoo.com/) sistema de mensajería instantánea de Yahoo!

Para saber más

- Tutoriales de redes sociales en castellano, de Steve Borgatti: http://www.analytictech.com/networks/en_castellano.htm
- Linked: How Everything Is Connected to Everything Else and What It Means, por Albert-Laszlo Barabasi, Plume, 2003
- Six Degrees: The Science of a Connected Age, por Duncan J. Watts, W. W. Norton & Company, 2004
- The Wisdom of Crowds, por James Surowiecki, Anchor, 2005.

Referencias

1. Noor Ali-Hasan y Lada Adamic. Expressing Social Relationships on the Blog through Links and Comments. International Conference on Weblogs and Social Media. Boulder, Colorado. 2007.
2. Diego De Ugarte. Teoría de Redes Sociales. Contextos, 2006. http://www.deugarte.com/wiki/contextos/Teoría_de_redes_sociales
3. Steven Johnson. Emergence: The Connected Lives of Ants, Brains, Cities, and Software. Scribner, 2002.
4. Mark Levene. An Introduction to Search Engines and Web Navigation. Addison Wesley, 2005.
5. Cameron Marlow. Audience, structure and authority in the weblog community. MIT Media Laboratory, 2004.
6. David Weinberger. Small Pieces Loosely Joined, a unified theory of the web Perseus Books, 2002.

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 videologs, linklogs y fotologs 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 un práctica cada día más extendida en la Web: compañías de música sindican información sobre discografía que luego es publicada por sitios de comercio electrónico; bolsas de comercio sindican 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 Ramanhatan 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 sindicar 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 comodity) 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.

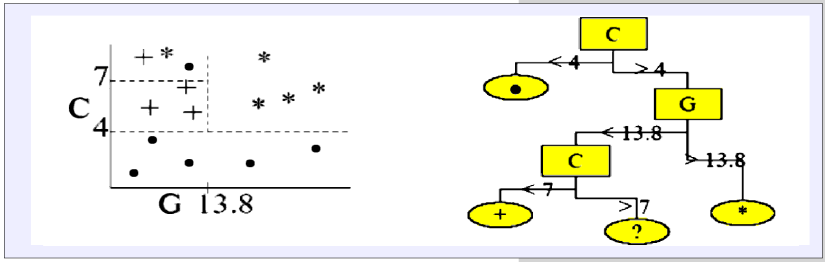
Hoy en día existen cientos de técnicas para desarrollar filtros de información, algunas de las cuales han alcanzado tasas de error menores a un 10% en diversos experimentos. Entre estas están los árboles de decisión, máquinas de soporte vectorial, redes neuronales, redes bayesianas, discriminantes lineales,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


Imagen tomada de www.morgefile.com



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 variable y condición se refleja en la primera "decisión" del árbol (ver figura arriba). 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.

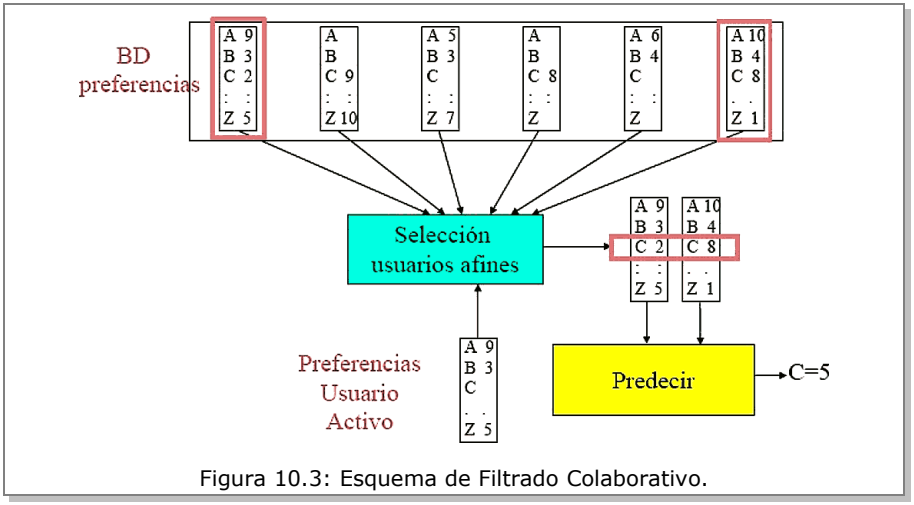


Figura 10.3: Esquema de Filtrado Colaborativo.

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 folcsonomías (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

Figura 10.4: (izquierda) Nube de tags generada por Orbitando. (derecha) Nube de tags generada por Technorati.

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
- KDNuggets 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 Pubdmoz lishers, 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

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 Computacion de la Universidad de Chile, dedicado a la invesgación básica en Ciencia de la computación, particularmente enfocado a la investigación en la Web.

Sus principales áreas de invesgitación son la recuperación de información, la minería en la Web, los aspectos logicos y semánticos de la Web, y el procesamiento de informacion 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.