CIENCIA DE LA WEB EN CHILE

Panorama de la Investigación sobre la Web en Chile

Mauricio Marín
Sociedad Chilena de Ciencia de la Computación. Yahoo! Research Latin America, Universidad de Chile. PhD en Computer Science, University of Oxford, UK.
mmarin@yahoo-inc.com

Claudio Gutiérrez
Profesor Asociado, DCC, Unviersidad de Chile. Ph.D. en Computer Science de Wesleyan University, Estados Unidos. Investigador Asociado del Centro de Investigación de la Web y el Grupo Khipu de bases de datos.
cgutierr@dcc.uchile.cl

PREHISTORIA

La investigación relacionada a las Tecnologías de la Información (TI) en Chile es bastante joven comparada con otras áreas de la ciencia y la ingeniería. Como muestra [2], el número de artículos relacionados a las TI publicados durante los años 80s, por autores en Chile en revistas internacionales con comites editoriales, promediaba los 70 anuales. Esto da un promedio de 4,85 artículos por habitante, el máximo indice por habitante en Latinoamérica (el promedio en Latinoamérica era en ese entonces de 1,59 artículos por persona).

Con respecto a la investigación sobre la web, una buena fotografía del estado del arte a inicios de los '90s es la siguiente frase de un artículo describiendo los mayores logros de las TI en Chile:

"This has allowed the spread of new technologies such as LANs and WANs (ethernets, bitnet, uucp, internet)" [3]

Es decir, las mayores preocupaciones eran la espina dorsal de Internet y la difusión y escalabilidad de las técnicas para redes.

Probablemente el primer artículo técnico sobre los problemas de Internet fue el escrito por R. Baeza-Yates, J.M. Piquer, y P. Poblete, en el cual discutían los problemas asociados a las conexiones a Internet en Chile [4]. Curiosamente, en esta primera investigación aparecieron los mismos tipos de problemas que asombraron a Darwin cuando visitó Chile durante la primera mitad del siglo XIX; es decir, lo complejo de su geografía:

"Chile has had two international 56Kbps links to the Internet since January 1992.

Probablemente el primer artículo técnico sobre los problemas de Internet fue el escrito por R. Baeza-Yates, J.M. Piquer y P. Poblete, en el cual discutían los problemas asociados a las conexiones a Internet en Chile

This online coAAnnection to the world is having a great impact on academic research, extending state-of-the-art communication technology to our country. The impact will be even greater considering the traditional isolation of the country, surrounded by mountains and sea at the end of the world, and a very particular geography: almost 4350 kilometers long from north to south with an average width of 190 kilometers.

En 1995 la palabra web aún no asomaba en el léxico de los investigadores nacionales. En el conocido artículo "Computing in Chile: The Jaguar of the Pacific Rim?", publicado en el Communications of the ACM, la Web no es mencionada. Los problemas aún eran los de Internet y las conexiones, cuya evolución es presentada en los siguientes hitos [5]:

- 1985: Mail electrónico internacional (uucp), seguido por Bitnet en 1987 (U. de Chile).
- 1987: Implementación de la primera red automática de cajeros dispensadores de plata.
- 1991: Inicio de la primera red de datos y de conectividad a Internet (entre la Universidad de Chile y la Universidad Católica).
- 1994: Puesta en funcionamiento de servicio ISDN experimental.

Probablemente el primer artículo en mencionar a la Web fue *A Model for Visualizing Large Answers in WWW* [1], presentado en la Conferencia Chilena de Ciencia de la Computación. El modelo presentado estaba basado principalmente en técnicas de recuperación de información, en las cuales uno de los autores, Ricardo Baeza-Yates, era un reconocido experto internacional.

El siguiente paso fue el estudio de la web chilena, que usó la información de un motor de búsqueda local, *todoCL.cl*, proyecto liderado por el mismo Baeza-Yates. *TodoCL* comenzó a operar en marzo de 2000 en colaboración con otro proyecto similar en Brasil, Akwan. En particular, *TodoCL* ha sido el único motor de búsqueda local cuyo objetivo es la Web chilena, y ha entregado datos para realizar estudios acerca de la caracterización de esta Web y otros estudios basados en *query logs*. Además, esto datos permiten hacer investigaciones sobre la dinámica de la Web que a menudo son impracticables.

EL CENTRO DE INVESTIGACIÓN DE LA WEB (CIW)

El año 2002 marca un punto de quiebre en la investigación de la Web en Chile, con la instalación del *Centro de Investigación de la Web* (CIW). Este centro, dirigido en sus primeros años por Baeza-Yates, nació con el financiamiento del gobierno de Chile a través de Mideplan.

El desafío era doble. Por una parte, desarrollar a nuevos niveles la investigación sobre la Web en Chile y, por otro, desarrollar la necesaria sinergia entre investigadores sin mayor interacción científica previa y dedicados con anterioridad a temas de investigación no directamente relacionados con la Web. Al final el proyecto resultó muy exitoso.

Después de tres años, el CIW fue capaz de reunir a 10 investigadores (2 post-docs), más de 55 estudiantes, y producir cerca de 50 artículos cada año. El campo de investigación de este proyecto era bastante amplio, aunque podía ser agrupado en dos áreas principales: (1) Bases de datos y recuperación de información, y (2) Sistemas distribuidos y redes. A continuación describimos los principales temas de investigación del centro basándonos en un informe interno del CIW del año 2004:

Bases de datos y recuperación de información

Esta área cubre recuperación multimedia, información espacial e información semiestructurada. El tema que subyace a estos tres es el de *combinatorial pattern matching*, un área de investigación que estudia desde un punto de vista combinatorial cómo buscar ciertos patrones en estructuras discretas y regulares como secuencias o grafos. Otro tema es cómo agregar información semántica a los contenidos, como metadatos y la web semántica. Incluimos aquí también el tema de minería de datos en la Web. A continuación describimos cada una de estas subáreas:

- Análisis multimedia y técnicas de búsqueda es un área de investigación en la que había trabajado Ricardo Baeza-Yates, Gonzalo Navarro (DCC, Universidad de Chile), Andrea Rodríguez (Informática y Computación, Universidad de Concepción), y Javier Ruiz del Solar (Ing. Eléctrica, Universidad de Chile). Esta área tenía un postdoc y varios alumnos de Ph.D. y master. En general, cubría todos los problemas de búsqueda relacionados con textos y multimedia, con mayor énfasis en algoritmos espaciales de búsqueda y algoritmos de *string matching*.
- Web semántica fue inicida por Carlos Hurtado (hoy en Universidad Adolfo Ibañez) y Claudio Gutiérrez (DCC, Universidad de Chile). Ellos comenzaron formalizando las especificaciones del Consorcio de la Web y justificando en términos de sus bases de datos. Esta área resultó ser muy exitosa. También relacionado con esto estaba el tema de agregar información semántica a los contenidos Web y obtener información desde ellos utilizando minería de datos.

Crawling y técninas de ranking es un área de investigación en la que Ricardo Baeza-Yates, Mauricio Marín (por entonces en la Universidad de Magallanes) y Andrea Rodríguez trabajaban. El principal tema de investigación era cómo crawlear la Web completa, reuniendo páginas para indexarlas y luego construir motores de búsqueda eficientes. Muchos tradeoffs aparecen cuando se trata de construir uno de estos motores, y el trabajo del grupo se centró en proponer nuevas tecnologías para mejorar la eficiencia y exactitud de la búsqueda.

TodoCL ha sido el único motor de búsqueda local cuyo objetivo es la Web chilena, y ha entregado datos para realizar estudios acerca de la caracterización de esta Web y otros estudios basados en query logs. Además, esto datos permiten hacer investigaciones sobre la dinámica de la Web que a menudo son impracticables.

Sistemas distribuidos y redes

Esta área cubre las tecnologías de programación para aplicaciones Web (*Web Agents, Web Services, Distributed Programming*), protocolos de comunicación para nuevos medios (*Multimedia over IP*), y tecnologías para mejorar el rendimiento de las herramientas de la Web (*parallel search, crawling technologies*).

José Piquer y Éric Tanter (en ese tiempo alumno de doctorado) trabajaron en Agentes móbiles y Programación distribuida, en una plataforma de programación basada en reflexión para Java (Reflex). Mauricio Marín y Gonzalo Navarro trabajaron en paralelismo. Un cluster de 10 nodos fue instalado en el DCC de la Universidad de Chile, sobre el cual se trabajó en el desarrollo de algortimos paralelos para bases de datos de textos, y para el *scheduling* distribuido de servidores simultáneos en la Internet.

Como se puede ver, la investigación no sólo era amplia con respecto a los temas, sino también con respecto a los participantes: El CIW incluyó investigadores de Santiago, Concepción, Punta Arenas, y estudiantes de muchas partes del país.

En el año 2003, por impulso del CIW, y bajo la tutela del *International World Wide Web Conference Committee* (IW3C2) y la Sociedad Chilena de Ciencia de la Computación (SCCC), se organizó en Santiago el Primer Congreso Latinoamericano de la Web. Este evento continuó además durante los siguientes años, convirtiéndose en una de las referencias de los investigadores de la región trabajando en el tema de la Web.

La expresividad y eficiencia son la clave para el éxito de las nuevas aplicaciones Web.

ACTUALIDAD

El CIW fue un éxito total. Fue además instrumental para atraer a Chile, en 2006, al primer laboratorio Yahoo! en el hemisferio sur. También las escuelas de verano realizadas anualmente y las nuevas fuentes de financiamiento atrajeron al CIW una ola de estudiantes a los temas relacionados con la Web. Hoy en día difícilmente hay un departamento de Computación en Chile dedicado a la investigación que no cubra algo ligado a la Web.

Una nueva etapa se abrió en 2006. Gonzalo Navarro reemplazó a Baeza-Yates como director del CIW, quien pasó a estar a cargo del laboratorio de Yahoo! Durante 2004 nuevos post-docs se incorporaron al proyecto: George Dupret and Benjamin Piwowarski cumplieron un importante papel en el área de minera de datos. También nuevos profesores de universidades locales se unieron al centro: Pablo Barceló, Benjamín Bustos y Éric Tanter del DCC de la Universidad de Chile, y Marcelo Arena, del DCC de la Universidad Católica. El centro además incorporó pos-docs desde fuera del país, así como estudiantes de posgrado de fuera de Santiago y de otros países de la región.

Las áreas de investigación se fueron refinando a través de los años. Para dar un panorama de la investigación actual, describiremos las líneas de investigación vigentes del CIW siguiendo su informe interno del año 2008:

- **Estructuras de datos compactas.** El objetivo es tratar de sacar ventaja del creciente gap que existe entre las velocidades de los niveles consecutivos de la jerarquía de memoria, mediante el diseño de estructuras de datos que operen con poco espacio y que por tanto quepan en memorias más rápidas.
- **Recuperación multimedia.** Las consultas que se hacen a las bases de datos multimedia buscan por similaridad más que por exactitud (como se hace en las bases de datos tradicionales). Esta área intenta buscar modelos de similaridad para objetos multimedia que se correspondan con la noción humana e intuitiva de similaridad, y también estructuras de datos y algortimos que apoyen una eficiente búsqueda por similaridad.
- **Lenguajes de programación y ambientes.** Intenta proveer el apoyo adecuado a softwares complejos que realizan *debugging*, a la programacion de ambientes inteligente, y a los aspectos dinámicos de los lenguajes de programación. Todos estos son importantes para el diseño de software para aplicaciones Web.
- **Estudio de la estructura de la Web.** Se intenta entender la dinámica de la Web, por ejemplo su estructura de conectividad, crecimiento y dinámica de cambio, etc. Esto permite un amplio rango de estudios, desde entender muchos fenómenos sociales hasta poder construir aplicaciones Web más eficientes.
- **Lenguajes para bases de datos.** El objetivo es diseñar lenguajes de consulta apropiados para manipular información en la Web, que es más compleja que en las bases de datos tradicionales. Por ejemplo, XML y RDF presentan nuevos desafíos en términos de expresividad y eficiencia, y son la clave para el éxito de las nuevas aplicaciones Web.

- **Consultas complejas para objetos en movimiento.** Las bases de datos de objetos en movimiento son una solución factible al problema de escalabilidad de un sistema centralizado de bases de datos. El problema que se ha tratado en el CIW es el de encontrar métodos de indexamiento distribuido usando un meta índice también distribuido.

El CIW ha ganado en el último tiempo un amplio reconocimiento en la comunidad de invesigación sobre la Web. Indicadores de esto son los 3 premios al mejor artículo obtenidos por investigadores del centro en conferencias relacionadas a la Web, y la apertura del laboratorio de investigación de Yahoo! en Chile.

Yahoo! Research Latin America es un nuevo laboratorio de investigación, localizado en la Escuela de Ingeniería de la Universidad de Chile. Bajo la dirección de Baeza-Yates, este laboratorio se concentra en las áreas de investigación de la Web y minería de datos. A continuación detallamos algunas de las áreas investigadas en el centro:

- **Búsqueda social.** El laboratorio ha estudiado la información obtenida a través de la búsqueda realizada por los usuarios mediante clicks. Básicamente, esta es la información que un usuario entrega al interactuar con su motor de búsqueda, por medio de escribir una consulta, luego cliquear en ciertos documentos, y eventualmente reescribir la consulta original.
- **Motores de búsqueda Sync/Async.** Los motores de búsqueda deben saber lidiar eficientemente con el tráfico de consultas generado por los usuarios. Redundancia de hardware puede ser reducida utilizando estrategias de procesamiento de consultas, en el caso en que un gran número de consultas pueden ser resueltas concurrentemente.
- **Bulk-synchronous crawling.** Los centros de datos de gran escala para los crawlers son capaces de manetener un gran número de conexiones HTTP activas, para poder bajar lo más rápido posible el enorme número de páginas Web desde una sección de la Web especificada. Esto genera un continuo flujo de nuevas URLs de documentos. El laboratorio investiga como se puede manejar eficientemente este problema mediante paralelización.

CONCLUSIONES

La información presentada en este artículo muestra el crucial rol que el CIW ha jugado en crear, desarrollar y consolidar la investigación acerca de la Web en Chile. En este punto es importante remarcar que el CIW no sólo ha sido importante en términos de investigación, sino también en términos de alcance de un público más amplio. En relación a esto podemos mencionar los estudios acerca de la Web chilena, la Ventana Digital que conectó a las ciudades de Arica y Santiago, un concurso para estudiantes acerca de la Web, y finalmente el libro ¿Cómo funciona la Web? que se ha vuelto bastante popular entre los alumnos y profesores de enseñanza media.

Con respecto a la investigación, presentamos un panorama cualitativo del estado del arte de la investigación sobre la Web en Chile. Por cierto, esto no representa toda la investigación sobre la Web realizada en Chile, pues nos hemos concentrado en aquella realizada por el CIW y Yahoo!, que son actualmente las dos mayores fuentes de investigación acerca del tema en el país.

Creemos que un estudio también cuantitativo acerca del tema se hace bastante necesario. Los datos y los investigadores están todos presentes, una oportunidad que no debería ser desperdiciada. BITS

REFERENCIAS

[1] O. Alonso and R. Baeza-Yates, A Model for Visualizing Large Answers in WWW. In *XVIII Int. Conf. of the Chilean CS Society*, 1998.

[2] R. Baeza-Yates, D. Fuller, J. Pino. Innovation as a critical success factor for the development of an information technology industry in Chile. In *12th IFIP World Computer Congress*, 1992.

[3] R. Baeza-Yates, D. Fuller, J. Pino. IT landmarks in less-developed countries: The Chilean case. In *21st CAIS/ACSI Annual Conference*, 1993.

[4] R. Baeza-Yates, J.M. Piquer, P. Poblete. The Chilean internet connection or I never promised you a rose garden, In *INET O93*, 1993.

[5] R. Baeza-Yates, D. Fuller, J. Pino, S. Goodman. Computing in Chile: The jaguar of the Pacific rim?, *Communications of ACM*, 38. 1995.