Entrevista Nacional:
Ricardo Baeza - Yates

Por Gonzalo Navarro

Es, por lejos, el investigador chileno más citado en ciencia de la computación, y el único investigador del área que pertenece a la Academia de Ciencias de Chile. Fue el creador del Centro de Investigación de la Web (CIW), único Núcleo Milenio en la disciplina, y es responsable del establecimiento del primer laboratorio Yahoo! Research en Chile y Barcelona. Hoy es vicepresidente de Investigación para Europa y Latinoamérica de Yahoo! Research, y supervisa los dos laboratorios ya mencionados más uno reciente en Haifa, Israel. Además es un viajero incansable, de actividad frenética en general, sin que por ello deje algo de tiempo para una buena conversación. Quisimos conocer de primera mano la visión de este profesor del DCC sobre la Web, la ciencia de la computación y Chile en particular.

Se ha dicho que investigar en la Web es algo único, en el sentido de que por un lado presenta desafíos algorítmicos, matemáticos y de sistemas formidables, y por otro puede tener un impacto inmediato en la sociedad. Como alguien que comenzó su carrera en algoritmos, ¿qué es lo que te atrajo, y qué es lo que te atrae ahora, de investigar en la Web? ¿Los temas de investigación han cambiado con el tiempo o son los mismos? ¿Sigue estando el foco en lo algorítmico o se ha movido a lo estadístico? ¿Cuáles son los desafíos más importantes en investigación en la Web para los próximos años?

La Web es actualmente el conjunto de datos más grandes que ha producido la humanidad, pues en la práctica uno puede generar un número infinito de páginas a través de páginas dinámicas. En septiembre de 2008 ya había más de 180 millones de servidores Web y posiblemente más de 20 mil millones de páginas estáticas. Su complejidad se debe a su volumen, continuo cambio y diversidad. Lo que me atrajo inicialmente fue la posibilidad de que las tecnologías de búsqueda que me interesaban pudieran ser usadas por muchas personas, por ejemplo a través del buscador de todo Chile, TodoCL. En otras palabras combinar la teoría y sus aplicaciones. Los temas relacionados con búsqueda de información siguen existiendo pero aparecen otros nuevos y uno de ellos es entender mejor la Web, donde la estadística y la minería de datos son importantes. Pero eso no significa que los temas anteriores dejen de ser importantes. Por eso los desafíos actuales son una mezcla de retos ya conocidos (recolectar la Web, jerarquizar páginas, etc.) y entender mejor la Web para poder aprovechar el conocimiento implícito que las personas generan, ya sea aportando contenido (en particular vía Web 2.0) o usando la Web.

Desde 2002 dirigiste el Núcleo Milenio Centro de Investigación de la Web, el único núcleo Milenio en computación en Chile, y en 2006 abriste el laboratorio de Yahoo! Research en Chile. ¿En qué crees que estas iniciativas han impactado o esperas que impacten en el país. ¿A nivel científico, social, educacional? ¿Ves diferencia entre el estado de la investigación en la Web antes de 2002 y ahora? ¿En qué sentido dirías que se ha avanzado?

A nivel científico el CIW permitió crear el grupo más grande de investigación en computación en un tema específico, sobrepasando la masa crítica. Este grupo ha logrado impacto internacional y por supuesto nacional. Por ejemplo, cuando

Gary Marchionini, conocido investigador de la Universidad de Carolina del Norte, nos visitó hace algunos años, me comentó que él pensaba que teníamos uno de los mejores grupos del mundo. De hecho, la producción de artículos actualmente supera varias veces a la de 2002. Un hito importante para hacer investigación de punta en algunos temas fue el buscador de Chile, TodoCL, que comencé el año 2000,

y permitió tener datos que pocos grupos de investigación académicos tenían. Todos estos hechos fueron fundamentales para convencer a Yahoo! de tener una sede de investigación en Chile, siendo el principal el capital humano.

A nivel social hemos tenido impacto a través de iniciativas que socializan la investigación como la Ventana Digital, los estudios de la Web Chilena y recientemente con el libro Cómo Funciona la Web. A nivel educacional hemos introducido temas nuevos que han permitido formar recursos humanos capacitados en temas diversos como tecnología de buscadores, algoritmos de compresión o Web semántica. A nivel gubernamental el laboratorio de Yahoo! se usa permanentemente como ejemplo para atraer otras iniciativas similares.

¿Qué fortalezas y debilidades tiene montar un laboratorio como el de Yahoo! Chile? ¿Cómo lo compararías con la situación en Barcelona o en EE.UU.? ¿Cuáles son los desafíos para la investigación en computación en Chile, qué hay que perseguir para estar al nivel del primer mundo? ¿La falta de masa crítica es un problema serio?

Chile está lejano del resto del mundo y es difícil traer investigadores de otros países. Tampoco es un país grande que pueda tener una fuente de talento local de gran tamaño. Este es el problema principal y la falta de masa crítica es siempre un problema importante. Los desafíos para hacer una buena investigación son similares a los de un país desarrollado, con el agravante que a veces la infraestructura y otros recursos no están disponibles a los niveles adecuados. Sin embargo estos obstáculos se pueden vencer con esfuerzo y tesón, algo que el DCC ha hecho a lo largo de más de 25 años.

¿Existe alguna particularidad especial en una Web regional, como la chilena o la latinoamericana, que valga la pena intentar explotar o comprender? En el caso de los buscadores, ¿se puede hacer mucho mejor que lo que haría un buen buscador horizontal? ¿De qué manera el estudio de una Web puede revelar información sobre la sociedad subyacente?

Las Web regionales tienen algunas características que son locales, como el idioma y el contenido relacionado con la cultura local. Incluso en el tema de los buscadores se pueden hacer algunas cosas mejor como recorrer la Web en forma más completa y más rápido al estar más cerca y hacer mejor ranking de los resultados al ser la colección más homogénea.

Los estudios de la Web local, además de servir para ver su evolución, pueden ser usados para muchos fines, pues la Web es un reflejo de la sociedad y tiene información social y económica de un país. Con herramientas de extracción de información uno podría mejorar la Web (por ejemplo creando nuevas páginas en forma automática en base a contenido existente). También mediante minería de uso en un buscador regional uno podría recolectar el conocimiento implícito que existe en ese uso, lo que hoy se llama sabiduría de la gente. En particular aprovechar la Web 2.0 (anotaciones, opiniones, etc.) para extender la Web, que sería para mi el inicio de la Web 3.0.

¿Cómo defines la Web 3.0?

Hoy hay diferentes definiciones según el autor, muchas de ellas disponibles en Wikipedia. En mi opinión, la Web 3.0 va a suponer el aprovechamiento del contenido y el uso que está generando la Web 2.0 para distintos objetivos, desde extracción de conocimiento hasta generación automática de contenidos. Todo esto será posible gracias a la minería web, que permite capturar la experiencia y el conocimiento de la gente y ponerlo al servicio de todos, consiguiendo al final una Web mayor y mejor. BITS

Entrevista Internacional

Peter Buneman

Por Pablo Barceló

Peter Buneman es Professor of Database Systems en la School of Informatics de la University of Edinburgh. Su trabajo en ciencia de la computación se ha enfocado principalmente en bases de datos y lenguajes de programación; más específicamente en bases de datos activas, semántica de bases de datos, información aproximada, lenguajes de consulta, tipos para bases de datos, integración de datos, bioinformática e información semisestructurada. Últimamente Buneman ha trabajado en problemas asociados a bases de datos científicas tales como procedencia de datos, archivaciones y anotaciones. También ha participado en numerosos comités de programa y ha sido el chair de ACM SIGMOD, ACM PODS y ICDT. Es además fellow de la Royal Society de Edinburgo, fellow de la ACM, y ganador del Royal Society Wolfson Merit Award. Actualmente se desempeña como director de investigación del UK Digital Curation Centre.

Peter, tu realizaste uno de los primeros trabajos acerca de los fundamentos teóricos de XML y, en particular, acerca del diseño de lenguajes de consulta para XML, durante la segunda mitad de los años '90s. ¿Qué fue lo que te hizo trabajar en XML en ese momento?

Esa es una muy buena pregunta. Antes de trabajar con XML estuve investigando lenguajes de consulta para objetos complejos. Y claro, muchos de estos lenguajes de consulta parecían ser apropiados para formatos de datos científicos en los cuales nosotros estábamos particularmente interesados. Teníamos muchos de estos formatos científicos, y cada uno tenía sus particularidades y eran todos realmente muy interesantes. Desarrollamos entonces algunas álgebras para objetos complejos que se comportaban muy bien como lenguajes de consultas para estos formatos. Pero había uno de estos formatos, en particular un formato para datos biológicos llamado SDBH, que no podía ser tratado con nuestras álgebras. Los datos en este formato se estructuraban como un árbol con algunas propiedades interesantes: el documento contenía muchos nulos (información desconocida o faltante), tenía cierta estructura pero no tenía esquema, entre otras. Esto muestra que antes de interesarme en XML, incluso antes de saber que XML existía, me interesé en el tema de la información semiestructurada. De hecho, personalmente no me considero una persona que haya aportado demasiado al desarrollo de XML, pero sí me gustaría pensar que mi trabajo en información semiestructurada motivó la posterior investigación en lenguajes de consulta para XML.

En el último tiempo hemos vivido una proliferación de diferentes modelos de datos: desde sólo tener el modelo tradicional relacional, hemos pasado en poco tiempo a codearnos con XML, RDF, datos biológicos, etc. Es más, hace unos días me comentabas que los datos utilizados por los lingüistas no se ajustan a ninguno de estos modelos y que, en realidad, corresponden a otro modelo aun inexplorado. ¿Crees que en el futuro veremos una aún mayor proliferación de modelos de datos, o en algún momento se producirá una estabilización en la que unos pocos modelos de datos serán los que predominen?

Creo que de alguna forma este proceso se estabilizará. Primero que todo, aún la gente que trabaja en XML lo hace bajo ciertas simplificaciones del modelo. Por ejemplo, es usual que los DTDs en la práctica sean bastante más simples que los que estudiamos en teoría, es decir, una simplificación usual es que no contengan recursión en vez de ser gramáticas libres de contexto arbitrarias. Y cuando veo estas simplificaciones me da cada vez más la impresión de que XML se parece mucho a algunos modelos de datos muy simples como son las listas, las tuplas, etc. Y en realidad, bajo la mayor parte de estas simplificaciones, la información ya no se puede considerar semiestructurada; de hecho corresponde a una descripción perfectamente estructurada. Pero por otro lado, siempre hay algún elemento semiestructurado en XML: Uno puede agregar arcos al documento, agregar atributos, etc. Y en esa dirección creo que deberíamos empezar a estudiar la relación entre bases de datos y ontologías ¡aunque odio decir esta última palabra!. Creo que habrá un interesante desarrollo ahí. Entonces lo que va a reaparecer - aunque ya están reapareciendo implícitamente - son los modelos más tradicionales de datos, es decir, modelos estructurados, pero esta vez en conexión con las ontologías, que son mucho más libres y representan al elemento semiestructurado.

¿Crees que nuestra función, como "científicos de los datos o de la información," es tratar de entender cada uno de estos modelos de datos por separado, o crees que quizás necesitamos de una teoría de los datos más

general, que englobe a los diferentes modelos de datos que conocemos, es decir, una teoría general sobre los modelos de datos?

Es una muy buena pregunta a la cual creo no tener respuesta. Tú sabes que la gente ha descubierto las muy hermosas relaciones que existen, por ejemplo, entre el modelo relacional y la lógica de primer orden, o entre la información semiestructurada y el área de autómatas. Pero por otra parte existen otros modelos de datos, como los arreglos, que no calzan en este tipo de caracterizaciones. También los streams, donde hay muy importantes conexiones con el área de lenguajes de programación, pero a los cuales prácticamente no hemos estudiado desde el punto de vista de bases de datos.

Pero respondiendo a tu pregunta, creo que nunca tendremos una especie de gran teoría unificadora de qué son los datos. Pero creo que sí vamos a ser capaces de establecer cada vez más conexiones entre los diferentes modelos, y construir mejores lenguajes de consulta para éstos.

Hace unos días me comentabas que una de las cosas que más te gustaba del área de bases de datos es que aún la teoría y la práctica se mantenían relativamente cercanas. Sería bueno si pudieras explicarme un poco más acerca de esa idea.

Bueno, esto es lo que a mí me gusta de la ciencia de la computación, y muy particularmente del área de bases de datos. Siempre me ha gustado mirar ambos lados, teoría y práctica. Y creo que la mayor parte del tiempo la teoría y la práctica de las bases de datos colaboran bastante bien. En ese sentido el estudio de las bases de datos es un tema muy interesante, pues una buena idea teórica puede llegar a tener un impacto práctico rápidamente.

En ese sentido me gustaría mencionar nuestro trabajo acerca del problema de la procedencia de los datos (provenance). Este es un problema que apareció de nuestro estudio, a través de varios años, de qué significaba para los administradores de muchos tipos de datos diferentes entender de dónde provenía su información. Nos dimos cuenta que este problema necesitaba un modelo teórico simple, que permitiera después trabajar con él. Y esto es lo que más me gusta de ese modelo.

Me preocupa un poco si la computación podrá mantener esta dualidad teoría/práctica indefinidamente. Creo que podría llegar a pasar que nuestra disciplina se escindiera en dos diferentes áreas: Por un lado la ingeniería computacional – como un símil de lo que es hoy por hoy la ingeniería mecánica – y por otro lado, totalmente separada, la ciencia de la computación – de la misma forma que la física teórica está totalmente separada hoy de la ingeniería mecánica.

¿Crees por tanto que la ciencia de la computación, en general, y la teoría de bases de datos, en particular, todavía pueden ser "útiles" ?

Claro que sí. El proceso de tratar de "entender" algo es siempre muy interesante y puede llegar a tener impacto en el mundo real. Creo que esto es cierto respecto a mucha de la investigación teórica en ciencia de la computación. Y si no llega a tener un impacto teórico al menos ayuda a dar una visión más completa del problema, o a dirigir futuras investigaciones que sí podrían tener un impacto.

La verdad es que nunca se sabe que tendrá impacto o que no. Y algunas cosas tienen impacto muchos años después. Por ejemplo, yo entré a la comunidad de bases de datos un poco después de la invención de las bases de datos relacionales. Y lo que la gente decía por ese entonces era algo así como "sí, el modelo relacional es muy elegante; pero la verdad es que nunca tendrá impacto en la práctica!" Y la verdad es que la teoría en este caso fue muy importante para poder poner el modelo relacional en práctica. Este tipo de ejemplos ha sucedido también en muchas otras áreas de la computación como por ejemplo en lenguajes de programación, donde ciertas ideas, como la teoría de tipos, han probado ser aplicables muchos años después de su invención.

Por lo demás, la teoría de la computación es barata. Lo que invertimos en ella es ínfimo con respecto a lo que invertimos en los grandes proyectos de software, que además raramente tienen impacto. Además la teoría tiene sus propios medios de auto-regularse, por lo que creo que deberíamos apoyarla. Por supuesto que existirá un efecto colateral, el que encontraremos también en investigación que tiene muy poca posibilidad de ser aplicada y que más bien tiene valor matemático. Pero personalmente no veo que este sea una buena argumentación en contra de este tipo de investigación.

Y por último, ¿podrías contarnos cuáles son a tu parecer las cuatro o cinco contribuciones más importantes de la ciencia de la computación?

Es una pregunta bastante difícil ... Lo que está de alguna forma mas cerca de mi corazón son todas aquellas ideas desarrolladas con relación a los lenguajes de programación: ideas acerca de tipos, de concurrencia, etc. Todas ellas muy hermosas y que han llegado a tener impacto en la práctica, aunque probablemente no en la forma en que se pensó al inicio. Y lo mismo acerca de las bases de datos: Estoy pensando aquí en todos esas elegantes conexiones que se han establecido entre la lógica y los lenguajes de consulta y que han tenido un profundo impacto en las aplicaciones. Es sólo cosa de ponerse a pensar un poco en cómo esto ha influenciado nuestra manera de almacenar y manipular datos. Y puede ser que estas relaciones teóricas y sus aplicaciones aún parezcan un poco difíciles para el usuario, pero es increíble al menos ver como se han simplificado las cosas con respecto a 20 años atrás.

Otras áreas que encuentro fascinantes son la criptografía y la teoría de complejidad computacional. Estas son áreas de las que sé bastante poco, pero que han tenido real impacto en nuestra manera de entender la computación.

Y por último, con respecto a cuáles serán los mayores desafíos en el futuro, me parece que el principal es que la Ley de Moore no podrá ser validada en un sólo procesador para siempre. Creo que, por tanto, nuestro campo tendrá que abrirse a estudiar en profundidad los modelos de computación paralela. Y mi visión es que habrá desarrollos muy interesantes relacionados con esto. BITS