Sistemas de Información

Ma. Andrea Rodríguez.

Ma. Andrea Rodríguez:

INVESTIGACIÓN EN SISTEMAS DE INFORMACIÓN ESPACIAL EN DIICC-UDEC

Departamento de Ingeniería Informática y Ciencias de la Computación, Universidad de Concepción.

Desde mis estudios de posgrado mi investigación ha estado centrada en el manejo de información geográfica (espacial) y espacio-temporal. Sistemas tales como monitoreo ambiental o planificación territorial, buscadores y servidores de mapas en la Web (Google Earth, Google Map), localización automática de vehículos y sistemas de navegación o ruteo, entre otros, son las aplicaciones inmediatas de la investigación que llevo a cabo. Estos sistemas representan objetos que son localizados en un espacio de más de una dimensión y en un instante o intervalo de tiempo. La información espacial es compleja, ya que debe lograr representar la geometría de los objetos y satisfacer las restricciones impuestas por el dominio espacial bajo representación. Así mismo, estos datos deben ser manipulados por un conjunto de operadores que definen características, tales como área, perímetro o largo, y definen relaciones espaciales, tales como adyacencia, inclusión o separación. El manejo de este tipo de información requiere del desarrollo de modelos conceptuales y lógicos, estructuras de datos y algoritmos de procesamiento de información espacial y espacio/temporal.

Mi trabajo de investigación ha estado apoyado por el financiamiento obtenido desde Fondecyt, Centro de Investigación de la Web, ECOS/CONICYT y Fundación Andes. Las temáticas que se han cubierto abarcan distintos aspectos en el manejo de información espacial. A un nivel semántico se propuso funciones de similitud entre conceptos espaciales definidos en una o varias ontologías. A un nivel de estructuras de datos se propuso un meta-índice para resolver consultas espacio-temporales en un ambiente de servidores de datos distribuidos. En el ámbito de buscadores en la Web se implementó una estrategia de georeferenciación de documentos Web y de agrupamiento de noticias con referencia espacial. Mi trabajo más reciente ha abordado la formalización de restricciones de integridad espacial y espacio-temporal y el manejo de inconsistencias espaciales. El estudio considera modelos de bases de datos espaciales que son extensiones al modelo relacional para los cuales define restricciones de integridad que combinan atributos temáticos y geométricos. En una primera etapa fueron consideradas restricciones para información espacial estática y actualmente se están estudiando restricciones de integridad para regiones que evolucionan en el tiempo. Esta formalización ha permitido luego analizar semánticas de reparación de bases de datos inconsistentes y definir medidas de consistencia para caracterizar una base de datos inconsistente. El producto de mi trabajo de investigación se encuentra en publicaciones en revistas, tales como: IEEE Transactions on Knowledge and Data Engineering, Information Systems, IEEE Evolutionary Computation, International Journal of Geographic Information Science, y en conferencias internacionales, tales como: Symposium on Spatial and Temporal Databases SSTD, ACM SIGSPATIAL GIS, Database Systems for Advanced Applications DAFTA, Web Information Systems Engineering WISE, entre otras.

En mi investigación he contado con colaboración a nivel nacional e internacional. En forma cercana he trabajado con Loreto Bravo (Universidad de Concepción) en la formalización de restricciones de integridad, y con Mónica Caniupán (Universidad del Bío-Bío) y Leopoldo Berstossi (University of Carleton, Canadá) en la definición de una semántica de reparación de bases de datos espaciales inconsistentes. Con Mauricio Marín (Yahoo Research! y Universidad de Santiago) he mantenido una constante colaboración en el trabajo de estructuras de datos para objetos en movimiento, lo que se basa en un trabajo previo con Gonzalo Navarro (DCC Universidad de Chile) y con el entonces alumno de doctorado del DCC Gilberto Gutiérrez (Universidad del BíoBío). Junto a Claudio Gutiérrez (DCC Universidad de Chile) he explorado propiedades topológicas de redes y comparto el interés por aplicar conceptos de Web semántica en el contexto de Linked Data y, en particular, Geo-Linked Data. A nivel internacional he trabajado, entre otros, con Max Egenhofer (supervisor de mi tesis doctoral) y con Fred Fonseca (Penn State University) en aspectos de ontologías para información espacial. Actualmente mantengo investigación conjunta con Nieves Brisaboa (Universidad de A Coruña) en cuanto a medidas de inconsistencia y con Christophe Claramunt (Naval Research Institute, France) para la modelación de restricciones de integridad espacio-temporal. No menos importante ha sido la colaboración de estudiantes de pregrado y del Magíster en Ciencias de la Computación de la Universidad de Concepción. En estas temáticas se han graduado diez alumnos de Magíster y más de quince alumnos de la carrera de Ingeniería Civil Informática de esta Universidad.

Benjamin Bustos:

CONTENT-BASED MULTIMEDIA INFORMATION RETRIEVAL

Departamento de Ciencias de la Computación, Universidad de Chile.

Mis principales áreas de investigación se centran en las áreas de búsqueda por similitud en colecciones de datos multimedia, especialmente colecciones de imágenes en la Web, modelos 3D y secuencias de video, y en el área de algoritmos de indexamiento para información no estructurada, con énfasis en el manejo de colecciones muy grandes de información multimedia.

En particular, he desarrollado algoritmos y técnicas de indexamiento para espacios métricos, no métricos y multimétricos.

Algunos proyectos de investigación recientes en los cuáles he participado son los siguientes:

- (2010) Investigador (contraparte chilena) del Proyecto SCHR 1229/2-1 "German-Chile Research Cooperation on 3D Object Retrieval", financiado por la Fundación Alemana de Ciencia (DFG) dentro del Programa de Cooperación Chileno-Alemana en Investigación.
- (2007-2009) Investigador principal del Proyecto FONDECYT 11070037, "Effective and efficient retrieval in multimedia databases".
- (2007-2008) Investigador Joven en el Núcleo Milenio Centro de Investigación de la Web.

Colaboradores internacionales y nacionales

Colaboro con investigadores nacionales e internacionales en tópicos de investigación como indexamiento en espacios métricos y no métricos, búsqueda por similitud en colecciones de objetos 3D, búsqueda de imágenes en la Web y teoría de indexamiento multimedia.

Grupo PRISMA: Benjamin Bustos, Violeta Chang, José Saavedra, Iván Sipirán y Juan Manuel Barrios.

Mis principales colaboradores en investigación son: Prof. Tomas Skopal, Charles University in Prague, República Checa; Dr. Tobias Schreck, Technische Universitaet Darmstadt, Alemania; Dr. Oscar Pedreira, Universidade da Coruña, España; Dra. Bárbara Poblete, Yahoo! Research Lab; Dr. Nelson Morales, DELPHOS Lab, AMTC, Universidad de Chile.

Alumnos de Posgrado

Actualmente dirijo cuatro estudiantes de Doctorado en Ciencias, mención Computación (Juan Manuel Barrios, José Saavedra, Iván Sipirán, y Violeta Chang, ésta última en conjunto con el profesor Gonzalo Navarro), y un alumno de Magíster en Ciencias mención Computación (Víctor Sepúlveda).

Journals y Conferencias

En los últimos cinco años he publicado siete artículos de revista, 16 artículos en conferencias internacionales y dos capítulos de libro. Principalmente publico en las siguientes revistas y conferencias internacionales: ACM Computing Surveys; IEEE Transactions on Knowledge and Data Engineering; Multimedia Tools and Applications; Eurographics Workshop on 3D Object Retrieval (3DOR); International Conference on Similarity Search and Applications (SISAP).

Grupo de investigación

Soy Director del Grupo de Investigación PRISMA (Pattern Recognition, Similarity Search, and Indexing in Multimedia Archives), perteneciente al DCC de la Universidad de Chile. El objetivo principal del grupo es investigar nuevos algoritmos y técnicas para poder realizar búsquedas en grandes colecciones de datos multimedia en forma eficaz y eficiente.

En la actualidad, el grupo PRISMA trabaja en variados proyectos de investigación, que corresponden principalmente a las tesis de doctorado de los asistentes de investigación del grupo. Algunos de estos proyectos son: búsqueda en colecciones de modelos 3D; búsqueda con medidas de similitud no métricas; detección de copia de videos; búsqueda en imágenes basada en sketchs; búsquedas por similitud usando índices comprimidos.

Desarrollo industrial y transferencia tecnológica

A través del Grupo de Investigación PRISMA, recientemente hemos realizado un exitoso proyecto de cooperación con la empresa chilena Orand, especializada en el desarrollo de software para proyectos de innovación. El proyecto consistió en el desarrollo de algoritmos para el reconocimiento del nombre y endoso en cheques manuscritos.

Esta tecnología se encuentra actualmente implementada en el "Chequemático", una máquina pagadora de cheques del Banco BCI. Actualmente se encuentran otros proyectos en carpeta para ser realizados junto a Orand.

Contacto

E-mail de contacto:
bebustos@dcc.uchile.cl.

Web del Grupo PRISMA:
http://prisma.dcc.uchile.cl.

Carlos Hurtado, Alberto Mendelzon, asador, Claudio Gutiérrez y Gonzalo Navarro.

Claudio Gutiérrez:

SEMÁNTICA, BASES DE DATOS, WEB

Departamento de Ciencias de la Computación, Universidad de Chile.

Desde hace casi diez años, con diferentes colegas, hemos venido desarrollando en el Departamento de Ciencias de la Computación de la Universidad de Chile, los aspectos semánticos de manejo de datos en la Web.

Expliquemos. Lo que hizo popular a la Web fue la aplicación de técnicas de recuperación de información, tradicionalmente un área completamente disjunta de las de bases de datos. La primera, anclada en técnicas estadísticas; la otra, en la lógica. Una tiene como objetivo recuperar la mayor cantidad (recall) de la mejor (según algún criterio) (precisión) información con poca estructura (lenguaje natural, documentos, etc.). La otra, responder lógicamente a consultas y razonar sobre la información estructurada. No es casualidad que ambas comunidades tengan poco en común.

El punto de partida fue la aplicación de técnicas clásicas de bases de datos (pensadas y motivadas por aplicaciones de negocios y empresariales) al ámbito de la Web. El gran inspirador de este enfoque fue Alberto Mendelzon, quien era uno de principales teóricos de las bases de datos relacionales, un argentino muy latinoamericanista, que trabajaba en la Universidad de Toronto, en Canadá. Nuestro grupo tuvo la oportunidad de interactuar con él. Carlos Hurtado había sido su alumno en Toronto y por medio de él comenzamos a trabajar conjuntamente en estos temas.

Así comenzó a desarrollarse una masa crítica de investigadores y alumnos en torno a estos temas. El punto de partida fue estudiar RDF (Resource Description Framework), el lenguaje para describir recursos en la Web, como un modelo de datos, en la tradición de la disciplina de bases de datos. Partimos trabajando con Carlos Hurtado, con Alberto y luego con un conjunto amplio de colegas y estudiantes: Ernesto Krsulovic (estudiante de Magíster en el DCC de la Universidad de Chile, hoy consultor independiente), Renzo Angles (estudiante de Doctorado del DCC de la Universidad de Chile, hoy en la Universidad de Talca), Marcelo Arenas (Pontificia Universidad Católica, PUC), Jorge Pérez (Universidad de Talca, hoy terminando su doctorado en la PUC), Sergio Muñoz (Facultad de Ciencias de la Universidad de Chile), Alejandro Vaisman (Universidad de Buenos Aires), Andrea Rodríguez (Universidad de Concepción), y muchos alumnos: Marcela Calderón, Cristián Vásquez, Álvaro Graves, Mauro San Martín, Daniel Hernández, etc. A nivel internacional nos acompañaron los profesores Leopoldo Bertossi (Toronto), Peter Wood (UK), Mariano Consens (Toronto), Axel Polleres (Irlanda), Enrico Franconi (Bolzano), Asunción Gómez-Pérez (Madrid), Manolis Koubourakis (Grecia), y varios otros. Y varios estudiantes que han venido del extranjero a visitar nuestro grupo y trabajar con él: Draltan Marín (el primero que especificó formalmente la semántica lógica de RDF), J. Hayes (que desarrolló el formalismo de grafos de RDF), Javier Fernández (que se ha dedicado a desarrollar la escalabilidad del formato RDF), etc.

El grupo desarrolló aspectos teóricos y prácticos de estos temas: Las especificaciones del Consorcio de la Web (W3C) en estas materias: RDF, RDFS, SPARQL; especificaciones para el Gobierno chileno (XML, metadatos, hoy DataGov). A nivel académico interactúa con grupos de Bases de Datos y de Web Semántica. Entre ellos están centros europeos, norteamericanos y latinoamericanos. Podemos señalar la Universidad de Buenos Aires y Bahía Blanca en Argentina, Universidad de la República en Uruguay, Universidad Católica de Arequipa en Perú, Universidad Central

de Venezuela, la PUC de Río de Janeiro. En Europa desarrollamos intercambio con DERI (Irlanda), UPM (España), Bolzano (Italia), TU Vienna (Austria), Oxford (UK), y en Estados Unidos el RPI. En la misma línea publica en conferencias de esa áreas: International Semantic Web Conference, Extended Semantic Web Conference, World Wide Web Conference, PODS, y diversos Workshops del área, y journals de bases de datos y Web Semántica, como TODS, TKDE, JWS, JCSS, etc.

En la actualidad el grupo está enfocado en el desarrollo de estos temas ligados a Linked Data, Open Data y movimientos que tienden a desarrollar los aspectos de razonamiento y escalabilidad en la Web. Entre las principales líneas de trabajo y actividad están:

a) Desarrollo y estudio de estándares W3C: RDF, SPARQL.

b) Participación formal e informal en grupos trabajo de W3C.

c) Desarrollo y estudio de nuevos modelos de datos y lenguajes de consulta para la Web. Particularmente en torno a la especificación RDF: SPARQL, RDB2RDF, HDT.

d) Desarrollo y estudio de aplicaciones de estas técnicas en Gobierno (DataGov) en la región y en Chile.

e) Formación de comunidad, a través de seminarios, workshops y charlas o de investigación dirigidas a la comunidad local.

f) Cursos y Extensión: A través de Educación Continua del DCC, en Cursos internacionales en Escuelas de Verano (Bolzano, UPM, Buenos Aires, Bahía Blanca, Montevideo, Arequipa, etc.).

g) Contacto e intercambio con otras organizaciones y grupos de investigación en diferentes niveles (W3C, ONG locales, Gobierno, KHIPU, Datos-Chile, etc.)

Loreto Bravo:

LIMPIEZA Y CONSISTENCIA DE LOS DATOS

Departamento de Ingeniería Informática y Ciencias de la Computación, Universidad de Concepción.

Desde que comencé mi Doctorado me he centrado en temas de investigación relacionados con Datos Inconsistentes. Durante mis estudios en Canadá me centré, junto con mi supervisor, Leopoldo Bertossi, en el manejo de inconsistencias en bases de datos relacionales, en sistemas de integración de bases de datos y en sistemas P2P. En el contexto de bases de datos relacionales, nos concentramos en el problema de Consistent Query Answering y en la utilización de programas lógicos de reparación para computar las respuestas consistentes. Aplicando ideas de esta investigación, estudiamos además la semántica de sistemas de integración de datos y P2P en la presencia de restricciones de integridad.

Al terminar mi Doctorado realice un Posdoctorado en el grupo de Bases de Datos de la University of Edinburgh, UK. Ahí trabajé con Wenfei Fan, Floris Geerts y Shuai Ma en extensiones a restricciones de integridad, como dependencias funcionales y de inclusión, especialmente diseñadas para la limpieza de datos. Estudiamos los problemas de satisfacibilidad e implicancia para estas restricciones.

Durante mi Posdoc comencé también a trabajar, junto a Irini Fundulaki (ICS-Forth, Grecia) y James Cheney (University of Edinburgh, UK) en control de acceso para bases de datos XML. En particular, nos concentramos en la detección de inconsistencias de las políticas de control de acceso, es decir, en detectar si es posible conseguir por medio de una secuencia de operaciones permitidas una acción que es prohibida. También hemos estudiado el problema de reparar las políticas en forma automática. Esta investigación ahora cuenta con el financiamiento de un proyecto Fondecyt de iniciación.

Loreto Bravo.

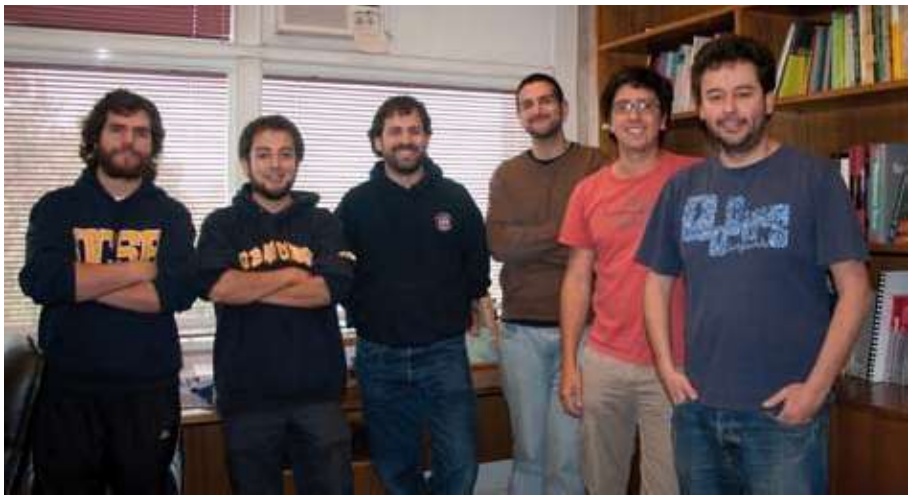
Ya instalada en Chile he comenzado a realizar investigación junto con Andrea Rodríguez (Universidad de Concepción) en la formalización y estudio de propiedades de restricciones de integridad para bases de datos espaciales. También he trabajado con Mónica Caniupán (Universidad del Bío-Bío), Carlos Hurtado (Universidad Adolfo Ibáñez) y Leopoldo Bertossi en consistencia de dimensiones de Data-Warehouses. Finalmente, también en conjunto con Leopoldo Bertossi, hemos continuando con la investigación de bases de datos P2P comenzada durante mi Doctorado y estamos preparando un artículo "Database Repairs and Consistent Query Answering" para la Synthesis Lectures on Data Management de Morgan & Claypool.

Los resultados de mi investigación han sido publicados en conferencias como VLDB, ICDE, EDBT, LPAR, DBPL, IJCAI y en las revistas Information Systems y Journal of Applied Logic de Elsevier.

Marcelo Arenas:

INTEROPERABILIDAD EN SISTEMAS DE MANEJO DE INFORMACIÓN

Departamento de Ciencia de la Computación, Pontificia Universidad Católica de Chile.



Marcelo Arenas y su grupo PUC Chile.

El área de Bases de Datos, a pesar de ser un tema clásico en Ciencia de la Computación, cobra mucha relevancia hoy en día por los desafíos que imponen las nuevas tecnologías. Siguiendo esta premisa nuestro grupo investiga temas de manejo de información motivados por problemas de interoperabilidad entre aplicaciones muy relevantes por el uso de Internet y la Web.

Dos de nuestras principales áreas de investigación son el intercambio de información y la integración de información. El problema de intercambio de información surge cuando dos aplicaciones (bases de datos, páginas Web, servicios Web, etc.) que trabajan de manera independiente desean compartir información y a la vez mantener su independencia. Por su parte, en integración de datos el problema principal es proveer a un usuario (persona o máquina) de una vista unificada a fuentes de datos dispares e independientes. Ambos problemas están muy relacionados y varios de nuestros artículos han ayudado a formalizarlos y dar solución a algunos de los desafíos que ellos presentan.

Otra de nuestras áreas de investigación es el manejo de información en la Web semántica. Ésta es una iniciativa de la W3C para agregar información a la Web que tenga tanto sentido para las personas como para las máquinas. Nuestro grupo, colabora estrechamente con investigadores de la Universidad de Chile, ha estado en el centro de la definición de las tecnologías básicas de la Web semántica, en particular de los lenguajes de consulta para datos semánticos de la Web.

Una característica definitoria de nuestro grupo es la rigurosidad, tanto en la formalización de los problemas como en la formulación de soluciones. Creemos firmemente que una base matemática sólida es esencial para dar soluciones que puedan ser comprobadamente mejores que las actuales y robustas de implementar. Es así como nuestra investigación tiene un fuerte componente teórico basado en herramientas como lógica computacional, en particular teoría de modelos finitos, complejidad computacional y complejidad descriptiva.

Parte de la calidad e impacto de nuestro trabajo puede ser medido por los premios académicos que estos han obtenido. Nuestro grupo ha obtenido cinco premios al mejor artículo ("Best Paper Award") en las más destacadas conferencias de teoría de bases de datos (PODS'03, PODS'05, ICDT'10) y Web semántica (ISWC'06, ESWC'07).

Colaboradores

Mantenemos una estrecha colaboración con investigadores de la Universidad de Chile, en particular con Pablo Barceló y Claudio Gutiérrez. Se destaca también nuestra colaboración con la industria internacional, en particular con Ron Fagin de IBM Almaden y Phil Bernstein de Microsoft Research. Parte de nuestros miembros han hecho pasantías y estadías cortas en estas empresas.

Adicionalmente colaboramos con investigadores de universidades internacionales entre los que podemos destacar a Leonid Libkin, Juan Reutter, Wenfei Fan y Kousha Etessami, (University of Edinburgh); Juan Sequeda (University of Texas at Austin); Cristian Riveros (Oxford University); Axel Polleres (National University of Ireland); Leopoldo Bertossi (Carleton University); Mariano Consens (University of Toronto); Filip Murlak (University of Warsaw); Alan Nash (Aleph One LLC); Rajeev Alur (University of Pennsylvania); Neil Immerman (University of Massachusetts).

Alumnos vigentes

Jorge Pérez (Doctorado), Martín Ugarte (Doctorado), Carlos Buil-Aranda (Doctorado visitante, Universidad Politécnica de Madrid), Sebastián Conca (Magíster), Andrés Letelier (Magíster) y Alejandro Mallea (pregrado).

Conferencias internacionales

Publicamos en los últimos cinco años en las principales conferencias de bases de datos: ACM Symposium on Principles of Database Systems (PODS), International Conference on Database Theory (ICDT), e International Conference on Very Large Data Bases (VDLB). Publicamos también en las conferencias más importantes de Web semántica: International Semantic Web Conference (ISWC) y European Semantic Web Conference (ESWC). Parte

de nuestra investigación ha sido publicada en conferencias de lógica y autómatas como: Annual IEEE Symposium on Logic in Computer Science (LICS), y el International Colloquium on Automata, Languages and Programming (ICALP).

Revistas

En los últimos cinco años hemos publicado en las revistas: Journal of the ACM (JACM), SIAM Journal on Computing (SICOMP), ACM Transactions on Databases Systems (TODS), IEEE Transactions on Knowledge and Data Engineering (TKDE), SIGMOD Record, Annals of Pure and Applied Logic (APAL), Theory of Computing Systems (TOCS), Journal of Web Semantics (JWS), Logical Methods in Computer Science (LMCS), Journal of Computer and System Sciences (JCSS).

Número de artículos publicados en los últimos cinco años:

- Revistas: 14
- Conferencias internacionales: 11
- Libros: 1
- Capítulos de libros: 2
- Workshops internacionales: 3

Mónica Caniupán:

CONSISTENCIA DE DATOS SOBRE DIFERENTES MODELOS

Departamento de Ingeniería Civil Informática, Universidad del Bío-Bío.

Obtuve el grado de PhD in Computer Science en Carleton University (Ottawa, Canadá) el año 2007 bajo la supervisión del Dr. Leopoldo Bertossi. Mis intereses en investigación están centrados en: (a) Teoría de Bases de Datos, (b) Integridad de Bases de Datos, (c) Calidad de Datos, (d) Representación del Conocimiento y (e) Programación Lógica.

Durante mi Doctorado me dediqué a estudiar y definir optimizaciones para programas en lógica de manera de ser utilizados en el cómputo de información consistente desde bases de datos inconsistentes (bases de datos que no satisfacen sus restricciones de integridad). La tesis se tituló "Optimizing and Implementing Repair Programs for Consistent Query Answering in Databases". Los resultados de esta investigación, fueron publicados parcialmente en: (a) In Current Trends in Database Technology, LNCS 3268 (2004), (b) Conferencia Internacional de la Sociedad Chilena de Computación" (2005), (c) "The Scalable Uncertainty Management Conference (SUM'07), LNCC 4772" (2007). Finalmente se publicó un artículo en Data and Knowledge Engineering Journal el año 2010 (69(6):545-572).

En 2007 me adjudiqué el proyecto Fondecyt de iniciación en investigación "Semantically Correct Answers to Queries in Inconsistent Multidimensional Databases". El objetivo de este proyecto fue definir una semántica de reparación para bases de datos multidimensionales que no satisfacen sus restricciones de integridad. Esto permitiría responder adecuadamente (consistentemente) a consultas de agregación. Demostramos en este trabajo que la teoría definida para bases de datos relacionales no puede ser aplicada a bases de datos multidimensionales; definimos una nueva semántica de reparación y una solución basada en programas en lógica para obtener las reparaciones minimales de dimensiones en Data Warehouses. En este proyecto colaboraron: Loreto Bravo (Universidad de Concepción), Carlos Hurtado (Universidad Adolfo Ibáñez) y Leopoldo Bertossi (Carleton University, Universidad de Concepción). El principal resultado de esta investigación fue enviado al "Data and Knowledge Engineering Journal" y en estos momentos se encuentra en proceso de revisión. También hemos publicado dos artículos en distintas versiones del

Mónica Caniupán.

"Alberto Mendelzon International Workshop on Foundations of Data Management" (2009,2010). Actualmente estoy trabajando junto con Alejandro Vaisman (Universidad de Buenos Aires) en la implementación de soluciones algorítmicas (no basadas en programación lógica) para obtener reparaciones de dimensiones en Data Warehouses. En esta investigación también participa la alumna Noemí Castillo del Magíster en Ciencias de la Computación de la Universidad del Bío-Bío, cuyo título de tesis es "Algoritmos para Computar Reparaciones de Dimensiones en Data WareHouses".

Además, he trabajado en manejo de inconsistencias en Bases de Datos Espaciales con Andrea Rodríguez (Universidad de Concepción) y Leopoldo Bertossi. Resultados parciales de esta investigación fueron publicados en "The 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)" (2008). También enviamos un artículo a "Information System Journal" el cual se encuentra en proceso de revisión.

En los últimos cinco años he publicado cinco artículos en conferencias, un artículo en revista ISI y esperamos respuesta de dos artículos enviados a revistas ISI.

Grupo de investigación en Computación de Alto Rendimiento para la Web.

Mauricio Marín:

TEORÍA Y PRÁCTICA EN COMPUTACIÓN ESCALABLE PARA LA WEB

Departamento de Ingeniería Informática, Universidad de Santiago de Chile.

El grupo de investigación en Computación de Alto Rendimiento para la Web se especializa en el estudio de problemas ubicados en la intersección entre recuperación de información, minería de datos, y procesamiento paralelo y distribuido de la información. Su objetivo principal es desarrollar estrategias que le permitan a las aplicaciones de la Web escalar a millones de usuarios de manera eficiente en términos de uso de recursos de hardware y software.

El financiamiento para investigadores y tesistas de posgrado proviene de Yahoo! Research Latin America, el cual es un laboratorio de investigación alojado en la Fundación para la Transferencia Tecnológica (UNTEC) de la Universidad de Chile. A este laboratorio también concurren investigadores de varias universidades nacionales y uno de sus objetivos principales es convertirse en un referente en investigación aplicada trabajando en estrecha colaboración con los programas de posgrado nacionales. Otra línea de financiamiento proviene de Fondef en proyectos tales como el denominado "Observatorios Escalables de la Web en Tiempo Real", en el cual participan académicos y estudiantes de la Universidad de Santiago, Universidad de Concepción, Universidad de Chile y Universidad Técnica Federico Santa María. También existen proyectos de inserción de capital humano avanzado de Corfo y Conicyt, los cuales posibilitan la inclusión de posdoctorandos en las líneas de investigación del grupo.

Algunos de los problemas de investigación estudiados tienen la siguiente motivación: se estima que actualmente los centros de datos contienen del orden de los 60 millones de computadores, los cuales consumen al menos el 2% de la energía a nivel mundial que se utiliza para generar electricidad. Por otra parte, la Web duplica su tamaño cada seis u ocho meses y aún faltan grandes sectores de la población mundial por incorporarse como usuarios de las diversas aplicaciones de la Web. Es, por tanto, relevante desarrollar estrategias que permitan a los centros de datos administrar a centenas de miles de usuarios concurrentes por segundo y a la vez sean eficientes en consumo de energía.

Típicamente los centros de datos operan sus computadores en régimen permanente a una utilización que está entre un 20% y un 40% de su capacidad total. La razón es que estos sistemas deben estar preparados para enfrentar subidas bruscas en el tráfico de peticiones de servicio de usuarios tales como consultas frente a eventos globales que capten el interés de cientos de miles de usuarios concurrentes por segundo. Una línea de investigación desarrollada por el grupo tiene relación con el desarrollo de estrategias de procesamiento de consultas que sean capaces de reducir la cantidad de computadores desplegados en el centro de datos y hacerlos operar a una utilización mayor, pero incluir en ellos técnicas que les permitan absorber eficientemente subidas bruscas en el tráfico de consultas. Las técnicas desarrolladas tienen que ver con estrategias de caching e indexación distribuida, procesamiento paralelo de consultas tanto en sistemas de memoria distribuida como memoria compartida, y selección automática de nodos procesadores basada en aprendizaje de máquina.

El contacto con investigación aplicada real para sistemas Web de gran escala proviene por la vía de proyectos de investigación orientados al estudio de optimizaciones de productos de Yahoo! operando en producción. Actualmente se trabaja en dos proyectos relacionados con motores de búsqueda verticales. El primero, tiene relación con planeación de capacidad en el centro de datos, lo cual requiere el desarrollo de simuladores tanto a nivel macroscópico, es decir, simulación de *clusters* de nodos procesadores, como a nivel microscópico, esto es, simulación de procesadores multicore. Sobre estos simuladores que modelan el hardware, se desarrollan simuladores del software que componen los distintos servicios del motor de búsqueda vertical. Los desafíos en investigación están en la formulación de modelos pertinentes y su combinación con la aplicación de técnicas de optimización metaheurística orientadas a planificar el despliegue de servicios en los nodos procesadores del centro de datos. El segundo proyecto tiene relación con el empleo de técnicas de compresión de índices invertidos y multithreading, para hacer que los nodos procesadores que resuelven consultas enviadas al motor vertical, tengan capacidades de actualización en tiempo real de los documentos indexados en cada nodo. Los desafíos en investigación están en el desarrollo de técnicas de eficientes de indexación y gestión de threads para posibilitar la ejecución concurrente de transacciones de lectura y escritura sobre el índice comprimido.

Anualmente uno de los indicadores principales de desempeño del grupo de investigación tiene relación con la publicación de artículos en las conferencias más relevantes del área de recuperación de información y computación paralela y distribuida, como lo son las conferencias con acrónimos SIGIR, WWW, CIKM, ECIR, SPIRE, HPDC, ICPP, IPDPS y Euro-Par. También es relevante generar patentes en los Estados Unidos de América.

Pablo Barceló.

Pablo Barceló:

MODELOS EMERGENTES DE DATOS

Departamento de Ciencias de la Computación, Universidad de Chile.

Desde los años de mi Doctorado, realizado entre 2002 y 2006 en el Departamento de Ciencia de la Computación de la Universidad de Toronto, Canadá, vengo realizando investigación en modelos emergentes de representación y consulta de datos. Esto se refiere principalmente a dos cosas:

(1) El estudio de nuevos formatos para el manejo de información, que van más allá del tradicional modelo relacional, y que han sido impuestos por la aparición en los últimos 15 años de aplicaciones centradas en datos tan importantes como la Web, las bases de datos científicas, las redes sociales, entre varias otras. Estos nuevos formatos de datos destacan por permitir mayor flexibilidad de representación que el modelo relacional, manteniéndose al mismo tiempo la posibilidad de entregar cierta estructura a partir de elementos semánticos y jerarquías. Por esta razón se han llamado "semiestructurados" a este tipo de datos.

En particular, mi investigación se ha centrado en torno a dos modelos de datos semiestructurados: (a) XML (Extensible Markup Language), que es un metalenguaje que permite describir información a alto nivel, y que se ha convertido en el estándar para integrar e intercambiar información en la Web; y (b) las bases de datos de grafos, que es un modelo abstracto que se utiliza para describir aplicaciones centradas en los datos en las que la que la topología de estos es tan importante como los datos mismos (por ejemplo, redes sociales, bases de datos científicas, Web semántica, etc.) En ambos modelos de datos mi investigación se ha centrado en torno a el diseño y análisis de lenguajes de consulta (por ejemplo, entender la expresividad y complejidad de evaluación de estos) y la representación y estudio de la información incompleta e incierta (que aparece ubicuamente en escenarios distribuidos como la Web, donde la información está fragmentada y podría presentar altos grados de incertidumbre).

(2) El estudio de problemas dinámicos asociados a los datos, que aparecen en espacios en donde la información fluye constantemente como la Web, y que no corresponden a la línea más tradicional de estudio en bases de datos donde estos son considerados estáticos. Me he enfocado, en particular, en estudiar problemas dinámicos asociados a la integración y el intercambio de la información. Mi investigación se ha centrado en entender la complejidad computacional de los distintos problemas relacionados con estos dos temas, así como en la potencial aplicación de los lenguajes de consulta tradicionales –por ejemplo, SQL– en este escenario más complejo.

Aunque mi formación de pregrado no es en Ciencia de la Computación –soy Ingeniero Electricista de la Universidad Católica de Chile– siempre me han atraído los temas de la Computación, en particular aquellos que tienen que ver con la teoría y los algoritmos. En particular, los temas de bases de datos concitaron desde un primer momento mi atención porque combinan, de forma bastante equilibrada, dos de mis intereses:

(1) La posibilidad de realizar modelos abstractos de los datos, que no dependieran de una aplicación en particular, sino que más bien aglutinaran las características esenciales que definen a una familia de aplicaciones. Estos modelos abstractos se prestan naturalmente al análisis lógico/matemático de alto nivel, combinando de forma muy interesante herramientas que van desde teoría de autómatas, pasando por expresividad de lenguajes lógicos, hasta llegar a la teoría de complejidad. Muchas veces el trabajo matemático que se hace en bases de datos es de alta dificultad, no teniéndole nada que envidiar al análisis que se hace en otras ramas más teóricas de la computación.

(2) La posibilidad de que dicho estudio teórico sea de impacto para la comunidad más aplicada. Es decir, las bases de datos son un interesante espacio de problemas para el teórico, pero a la vez proveen el espacio para descubrir, mediante dicho estudio, propiedades fundamentales de los modelos de datos que pueden ayudar a la comunidad más aplicada a desarrollar aplicaciones más robustas y eficientes.

Colaboradores

Como es usual en Ciencia de la Computación, nuestro trabajo se ha desarrollado en cercana colaboración con investigadores a lo largo del mundo. En Chile mantengo cercano contacto con Marcelo Arenas (Pontificia Universidad Católica). Mis más cercanos grupos de investigación en la actualidad son