DOCTORADOS DEL DCC

TERESA BRACAMONTE

Título tesis: **Improving Web Multimedia Information Retrieval Using Social Data**
Profesora guía: **Bárbara Poblete Labra**

Estudié Ingeniería Informática en la Universidad Nacional de Trujillo en Perú, y después de un par de años trabajando entre la industria y la docencia universitaria, decidí que era hora de empezar estudios de postgrado. Mi plan inicial era realizar un magíster, pero las cosas se dieron de forma distinta a la planeada, y en marzo de 2010 estaba empezando el Doctorado en el Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile.

Mi tesis de doctorado "Improving Web Multimedia Information Retrieval using Social Data" se enmarcó en el área de minería de datos, y la investigación se enfocó en mejorar la recuperación de información multimedia en la Web usando datos generados a partir de la interacción entre usuarios y documentos multimedia, por ejemplo, imágenes. El objetivo principal de mi tesis, fue demostrar la relevancia del contexto asociado a documentos multimedia para mejorar el proceso de recuperación de documentos multimedia en la Web. La investigación realizada se orientó en mejorar la recuperación de información multimedia desde dos perspectivas: extrayendo conceptos relevantes a los recursos multimedia; y mejorando las descripciones multimedia con datos generados por el usuario. Para ambos casos, diseñamos algoritmos que funcionan independientemente del tipo de multimedia y del idioma de los datos de entrada.

Detectar conceptos relevantes asociados a documentos multimedia no es una tarea trivial. Por eso nos enfocamos en detectar los conceptos que podían derivarse de una misma consulta en base a los términos asociados a las imágenes que un motor de búsqueda considera relevantes. Por ejemplo: si la consulta fuese "llave" los conceptos visuales relacionados serían la llave con que abrimos puertas, o la llave de agua. Por otro lado, para mejorar la descripción de documentos multimedia, la mejor opción fue utilizar los términos de las consultas con que los buscamos, y luego propagar esta información a imágenes (cuasi) duplicadas. De esta forma, si alguien consulta por imágenes de automóviles usando la palabra "auto", y luego otra persona usa "carro", ambas palabras quedarán registradas como relacionadas con fotos de automóviles visualmente parecidos.

Más allá de la propuesta algorítmica, el desafío más grande fue realizar una evaluación con la que confirmásemos qué tan buenos eran nuestros métodos con respecto a otros. Como no existían conjuntos de datos que nos permitieran hacer evaluaciones automáticas, tuvimos que recurrir a diseñar estudios de usuarios. Afortunadamente, contamos con el apoyo de la comunidad del DCC, y del ahora Instituto Milenio Fundamentos de los Datos, para encontrar participantes voluntarios.

Mi primer contacto con el DCC fue a través de antiguos compañeros de la universidad que eran asesorados por Benjamín Bustos, y lo que me animó a venir definitivamente fue una oferta de trabajo como asistente de investigación que posteó Bárbara Poblete que en ese entonces regresaba a Chile después de terminar su doctorado en España. Trabajar con Bárbara fue una experiencia muy buena, de continuo aprendizaje, y de muchas oportunidades. La primera vez que conversé con ella, yo estaba en Perú y recién había recibido la carta de aceptación al doctorado. Ella estaba trabajando en Yahoo! Research, y aún no era profesora del DCC. Fue por medio de Yahoo! que me vi inmersa en el mundo del Big Data y me encantó. Gracias a Bárbara también tuve la oportunidad de contactarme con investigadores fuera de Chile, y hacer pasantías en el Centrum Wiskunde & Informatica en Amsterdam (Holanda), y en Microsoft Research en Nueva York (Estados Unidos).

Siempre me ha gustado investigar, y cuando empecé el doctorado pensaba que después de terminarlo seguiría relacionada con el mundo académico. Sin embargo, la combinación de diferentes circunstancias personales y profesionales me llevaron a tomar otro rumbo y desde hace casi dos años trabajo como desarrolladora de aplicaciones en el equipo de Data & Analytics de Equifax Inc. Una de mis principales actividades es extender y adaptar algoritmos de aprendizaje de máquinas de acuerdo a las restricciones del negocio, tal que funcionen en plataformas de Big Data (y ahora Cloud). Se puede decir que ahora trabajo haciendo herramientas para que *Data Scientists* puedan hacer su trabajo.

JAVIEL ROJAS

Título tesis: **Matching and Covering with Boxes**
Profesores guías: **Jérémy Barbay - Pablo Pérez Lantero**

Soy graduado de Ciencia de la Computación de la Universidad de La Habana, y recientemente Doctor en Computación por la Universidad de Chile. La computación desde niño fue una de mis grandes pasiones, y siempre tuve vocación por la docencia (probablemente ser hijo de profesores tuvo mucho que ver). Mezclar esas pasiones pasaba de forma natural por un doctorado, y en busca de eso llegué al Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile. En el Departamento trabajé bajo la supervisión de Jérémy Barbay del DCC, y Pablo Pérez-Lantero de la Universidad de Santiago (USACH). El trabajo con ambos fue excelente, ofreciendo ayuda desde el primer momento en que llegué a Chile (a un hemisferio de distancia de casa) y aportando ideas que ayudaron a enfocar la investigación, y a buscar alternativas cuando llegábamos a puntos muertos donde parecía que nada podía hacerse.

En mi investigación de doctorado estudiamos la complejidad computacional de un subconjunto de problemas con cajas multidimensionales con aplicaciones en distintas áreas, incluyendo geometría computacional, bases de datos, teoría de grafos, redes, entre otras. Estos problemas se enfocan en tres materias: el cálculo de emparejamientos de un conjunto de puntos con rectángulos, la detección de redundancias en la región cubierta por un conjunto de cajas, y el cálculo de distintas medidas de dicha región. Probamos que varios de los problemas son difíciles de resolver de forma óptima en un tiempo "aceptable" (tiempo polinomial), pero que sus respuestas pueden ser aproximadas de forma relativamente rápida. Para los que sí se pueden resolver óptimamente se introducen algoritmos adaptativos que mejoran, para grandes clases de instancias, los mejores algoritmos que se conocían.

Por ejemplo, en una aplicación de mapas a medida en donde uno hace zoom, los nombres de los lugares se ocultan o muestran en función del espacio disponible, de manera que éstos nunca se solapen (para que sean legibles). La aplicación debe decidir qué nombres mostrar, maximizando a la vez cierto criterio (por ejemplo, que se muestre la mayor cantidad de nombres, o los nombres más importantes para el usuario). Este problema se puede representar como un problema con rectángulos: se asigna un área rectangular para cada nombre, y cada vez que se hace zoom se debe encontrar un subconjunto de todos los rectángulos que no se solapen y que maximicen el criterio deseado (ver **Figura**). Este problema se conoce como Maximum Independent Set of Rectangles (MISR), y en general es poco probable que se puedan obtener algoritmos que lo resuelvan de forma óptima en un tiempo aceptable (cuando uno hace zoom no desea esperar un año a que las etiquetas se muestren, espera que esto se haga de forma rápida), por lo que se recurre a algoritmos que lo resuelven de forma aproximada, o que resuelven casos especiales del problema de forma óptima. En esta línea, estudiamos dos casos especiales de MISR, que a la vez generalizan otros problemas que ya se habían estudiado. Para el primero mostramos que se puede resolver de forma óptima en tiempo polinomial; y para el segundo mostramos que es poco probable encontrar un algoritmo que lo resuelva de forma óptima en tiempo polinomial, pero que se puede obtener de forma rápida una solución que es a lo sumo cuatro veces peor que el óptimo.

Mis áreas de interés son la geometría computacional, el análisis adaptativo de algoritmos, y el diseño y análisis de estructuras de datos compactas. En la actualidad realizo un postdoc en el Instituto Milenio Fundamentos de los Datos, donde estudio la existencia de algoritmos para evaluar eficientemente consultas sobre una base de datos, aprovechando distintas representaciones geométricas de éstos.

JOSÉ MIGUEL HERRERA

Título tesis: **Learning to Rank Social Knowledge for Question Answering in Streaming Platforms**
Profesores guías: **Bárbara Poblete Labra - Denis Parra Santander**

Soy Ingeniero Civil Informático con un Magíster en Ciencias de la Computación obtenido en la Universidad Técnica Federico Santa María. En este mismo lugar, hice docencia y también labores administrativas. También trabajé dos años en la industria.

Hacer el doctorado no fue una tarea fácil. En un principio, el hecho de haber trabajado en la industria, hizo que perdiera el ritmo de estudio. A eso hay que sumarle la alta exigencia del DCC durante el primer año relativo a tomar ramos, elegir un área de investigación y además preparar el examen de candidatura para fines de ese año (mientras escribo estas líneas pienso: "ufff... ya pasó"). La elección de un tema de investigación fue algo no menor puesto que venía de otra área de computación. Sin embargo, fue ahí cuando en una charla de la profesora Bárbara Poblete me di cuenta que analizar y buscar patrones en redes sociales era lo que quería hacer.

En particular, con la profesora trabajamos en un problema de investigación llamado *Question Answering* (QA) que consiste en extraer conocimiento de plataformas especializadas de preguntas y respuestas como Yahoo! Answers, Quora o Stackoverflow. Sin embargo, nuestro estudio se centró en interacciones QA producidas en microblogs, como Twitter. Si bien los microblogs no están diseñados para interacción de preguntas y respuestas, hay estudios que indican que alrededor del 10% de los mensajes diarios corresponden a preguntas. Por lo tanto, usando Twitter para todos los propósitos, el objetivo de nuestra investigación era que dada una pregunta, pudiésemos extraer un conjunto de respuestas candidatas y las ordenáramos de acuerdo a su relevancia (si responden a las preguntas).

Una de las principales novedades de esta investigación es que, además de extraer tweets, también extrajimos hilos de conversación dado que brindan mayor cantidad de información. De esta manera, construimos un modelo de ranking basado en más de 60 características extraídas de hilos de conversación de Twitter tales como aspectos sociales, categorías de las palabras (sustantivos, adjetivos, adverbios, etc.), tiempos de llegada de los mensajes, contenido semántico de las preguntas y respuestas, entre otras.

Las contribuciones de esta investigación fueron las siguientes: a) demostramos la posibilidad de utilizar microblogs como una fuente valiosa de información de preguntas y respuestas, b) identificamos las características más relevantes para preguntas que tienen solo una respuesta (*factoid questions*), c) creamos un modelo de ranking para determinar respuestas relevantes en la tarea de preguntas que tienen solo una respuesta (*factoid questions*) y, d) demostramos que el modelo se puede aplicar a preguntas más complejas (*non-factoid questions*) producidas en microblogs.

Respecto a mi experiencia de cursar el doctorado en el DCC, fue una decisión acertada. Mi visión de la ciencia e investigación cambió de manera radical. En el DCC se respira ciencia y, en general, los académicos son destacados investigadores en sus áreas a nivel nacional e internacional. Aprendí a investigar y a ser más crítico. Conocí un montón de personas extraordinarias de Chile y sobretodo extranjeros que nos compartieron su cultura y contagiaron con su buena onda. Eso sí, hubo momentos en los que no fue fácil compatibilizar los estudios con la vida familiar. Durante el doctorado me casé y tuve una hija (y un hijo cuatro días después de mi examen de grado), pero todo esto no fue un impedimento para continuar con mis estudios. Una buena organización de mi tiempo permitió que pudiera terminar exitosamente este desafío y cuando hubo dificultades, pude conversarlo abiertamente con las personas indicadas para llegar a una solución.

Quisiera aprovechar esta instancia para agradecer al DCC, la Universidad de Chile, Conicyt, al Instituto Milenio Fundamentos de los Datos y al Centro de Investigación de la Web Semántica (CIWS) por permitirme realizar mis estudios de doctorado. En particular, agradecer a mis mentores Bárbara Poblete y a Denis Parra por la paciencia, dedicación, tiempo, enseñanzas, simpatía y sobre todo la calidad humana. También agradecer a todas las personas que conocí en el DCC: compañeros del doctorado, estudiantes de magíster y de pregrado, y en general, a todo el staff del DCC.

Desde enero de 2019 soy científico de datos en el área de innovación y transferencia tecnológica del Instituto Milenio Fundamentos de los Datos.

VANESSA P. ARAYA

Título tesis: **Spatio-Temporal Historical Event Visual Exploration Through Social Media-Based Models**
Profesora guía: **Bárbara Poblete Labra**

Después de terminar mi pregrado, en 2012, estaba un poco perdida en términos de qué área seguir trabajando. Sabía que me gustaba la visualización de datos y que quería trabajar en conjunto con otras disciplinas. Después de darme muchas vueltas decidí que quería trabajar con datos geotemporales, y en particular me interesaba visualizar cómo evolucionaban las noticias en términos de impacto. La idea inicial era bastante básica: si una noticia pasaba en algún lugar, se iba a mostrar como una burbuja sobre ese lugar y su tamaño aumentaría o disminuiría en relación a la cantidad de gente que hablara de ella. Con esa idea comencé a trabajar con la profesora Bárbara Poblete, quien había trabajado anteriormente con datos de Twitter. La idea inicial evolucionó a un proyecto de tres partes: (1) modelo de datos, (2) herramienta de análisis y (3) visualización geotemporal.

Gran parte de la primera etapa trabajé en colaboración con Mauricio Quezada, también alumno de doctorado de Bárbara. Ésta se trató del procesamiento y modelamiento de noticias obtenidas de Twitter. Los datos utilizados fueron extraídos anteriormente, utilizando una metodología de un trabajo previo de Bárbara y Mauricio en conjunto con otros investigadores[1]. Durante parte importante del preprocesamiento nos dedicamos a extraer información geográfica de las noticias, incluyendo tanto lugares que estaban involucrados en el mundo real como aquellos lugares hacia donde se propagaban a través de las redes sociales. Analizando los datos obtenidos como resultado, nos dimos cuenta que muchas noticias no solo referenciaban a un único lugar, sino que muchas veces relacionaban a más de un país a la vez. Esto significa que no solo podíamos saber "dónde" pasaba una noticia, sino también las relaciones entre países que se generaban como consecuencia de eventos noticiosos y que era posible saberlo con datos de Twitter. Un ejemplo de este tipo de análisis fue la crisis de Crimea, en donde observamos cómo partía como un evento pequeño y local en Ucrania, para después evolucionar a un evento con impacto internacional que involucró tanto a Rusia como Estados Unidos.

La segunda parte de mi tesis se trató de la realización y evaluación de la herramienta de visualización que motivó el proyecto en un principio. La llamamos Galean y su objetivo principal fue permitir la búsqueda y exploración de los eventos noticiosos modelados en la etapa anterior. En particular, la idea era que dada una noticia recién ocurrida pudiéramos rastrear noticias pasadas que nos permitieran entenderla. La versión chilena fue desarrollada por Jazmine Maldonado a partir del prototipo que utilizamos para la investigación. Esta versión se mejoró como proyecto en colaboración de la Biblioteca Nacional y está actualmente disponible en www.galean.cl.

Al usar y evaluar la herramienta nos dimos cuenta de que hacer seguimiento de las relaciones entre países a lo largo del tiempo no era una tarea fácil. Esto se complicaba aún más si al mismo tiempo queríamos visualizar a qué lugares una noticia se propagaba. Esto nos llevó a buscar otras maneras de representar los datos. Así surgió la idea de Cartoglifos, una representación más abstracta del mundo que permitía ver más de una variable geográfica al mismo tiempo en un espacio más reducido.

Cuando empecé a hacer el doctorado no tenía mucha idea de lo que se iba a tratar, pero creo que fue una experiencia súper interesante, aunque muchas veces difícil y frustrante. En términos académicos me permitió explorar ideas que me llevaron a descubrir qué cosas me apasionan y en qué dirección quisiera seguir trabajando. Pude trabajar con profesores súper buenos y dedicados, además de comprometidos con hacer cambios en la sociedad más allá de la universidad. Por otra parte, más en términos personales, me permitió viajar y conocer gente de otros países, ver otras realidades y aprender cosas que hubiese sido difícil hacer en otras circunstancias.

Después de terminar mi doctorado, comencé mi postdoctorado en Francia en diciembre del año pasado, en el equipo ILDA en Saclay. Mi interés aún está en la visualización de datos geotemporales, tratando de buscar colaboración con otras disciplinas que tengan problemas relacionados con este tema.

1. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0166694

CARLOS OCHOA

Título tesis: **Synergistic (Analysis Of) Algorithms and Data Structures**
Profesora guía: **Jérémy Barbay**

El Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile tiene un claustro de profesores excelentes. En particular, en el área de algoritmos y estructuras de datos, tener a Gonzalo Navarro y a Jérémy Barbay es un privilegio para el DCC. Desde el punto de vista profesional, gracias a estos profesores tuve la oportunidad de trabajar en algunos de los problemas más desafiantes del área y adquirí las técnicas necesarias para resolverlos. Lo más complicado del doctorado desde el punto de vista no académico, fue adaptarse a un país nuevo con una cultura distinta a la mía.

En mi tesis propongo una nueva clasificación de técnicas algorítmicas: algoritmos cuya complejidad depende del orden en que los datos son datos, algoritmos cuya complejidad depende de la estructura de los datos (es decir, que son independientes de su orden), y algoritmos sinérgicos que se aprovechan del orden y de la estructura de los datos de forma sinérgica. En esta última categoría propusimos algoritmos sinérgicos para ordenar un multiconjunto, para calcular la eficiencia de Pareto y la envoltura convexa, los dos últimos de un conjunto de puntos en el plano. En un gran número de instancias, estos algoritmos son mejores que todos los algoritmos previos que solo se aprovechan del orden o de la estructura de los datos.

El análisis de estos algoritmos sinérgicos involucra un gran número de parámetros. Estos parámetros capturan cómo se relacionan las medidas de complejidad que dependen del orden y de la estructura de los datos. Obtener estos análisis demostraron ser las tareas más desafiantes, pero a la vez las más satisfactorias de la tesis.

La tesis transita por el camino de ir más allá del análisis del peor caso, que es el paradigma que ha dominado el análisis de algoritmos por siempre. Pero dada la gran cantidad de datos que el mundo está recolectando, en muchos dominios en donde el análisis del peor caso es muy pesimista, ha surgido la necesidad de hacer análisis más a la medida, en función de las instancias que aparecen más frecuentemente en la práctica. En este sentido, la tesis propone nuevos algoritmos y análisis en donde se aprovechan varias medidas de la entrada, sentando las bases para hacer análisis más a la medida.

Ahora mismo estoy haciendo un postdoctorado con el profesor Gonzalo Navarro en el Centro de Biotecnología y Bioingeniería (CeBiB). Todavía no he decido si seguiré en la academia o iré a la industria una vez finalizado el postdoctorado. ■

Integrantes de la comisión de grado (de izquierda a derecha): Rajeev Raman (Universidad de Leicester, Reino Unido), Travis Gagie (Universidad Diego Portales), Jérémy Barbay (Universidad de Chile), Carlos Ochoa, Diego Arroyuelo (Universidad Técnica Federico Santa María) y Claudio Gutiérrez. (Universidad de Chile).