Estudiantes DCC

En esta sección de la Revista estudiantes recientemente graduadxs del Departamento de Ciencias de la Computación (Universidad de Chile) nos cuentan, junto a sus profesores guías, sobre sus trabajos de memoria y/o tesis.

Tesis de doctorado

Similarity-based Web Queries (Consultas por Similitud en la Web)

Estudiante: Sebastián Ferrada
Profesores guías: Benjamín Bustos y Aidan Hogan

Hacer un doctorado fue una continuación natural tras finalizar el magíster. En el magíster creé IMGpedia (imgpedia.dcc.uchile.cl), una base de conocimiento que contiene información visual y semántica sobre las imágenes de Wikipedia. En IMGpedia se pueden realizar consultas, como por ejemplo: "Obtener pinturas del Louvre que sean similares a un autorretrato de Van Gogh".

La información guardada en IMGpedia es estática, es decir, se calcularon descriptores visuales y relaciones de similitud una sola vez y se guardaron. El siguiente paso parecía lógico: utilizar esa información (o cualquier otra) de forma dinámica, calculando relaciones de similitud de forma eficiente y a pedido de un usuario. Este fue el problema central de mi doctorado.

En mi tesis titulada "Similarity-based Web Queries" o "Consultas por Similitud en la Web" hago tres contribuciones. Primero, propongo un nuevo algoritmo para resolver *similarity joins* de forma aproximada. El *similarity join* es una operación que dados dos conjuntos de datos, obtiene pares de objetos, uno de cada conjunto, tales que son similares entre sí bajo algún criterio. Los criterios de similitud son tan variados que podríamos discutir sobre ellos en un tratado de filosofía intentando descifrar la respuesta a ¿cuándo dos cosas son parecidas? En la práctica, "las cosas" suelen ser vectores de números y "el parecido" se mide a través de funciones de distancia, siendo dos vectores más similares mientras más pequeña sea la distancia entre ellos. Nuestro algoritmo, *root-join*, agrupa objetos cercanos entre sí y calcula distancias solo entre elementos de cada grupo, ofreciendo así garantías de bajo tiempo de ejecución, al costo de no obtener una respuesta 100% correcta.

Como segunda contribución, definí e implementé un operador de bases de datos para extender el lenguaje de Web de consulta SPARQL, lo que permite efectivamente realizar *similarity joins* sobre la Web de Datos, dándole la libertad a los usuarios para seleccionar las dimensiones relevantes para ellos y seleccionar una función de distancia acorde a sus necesidades. La tercera contribución son dos aplicaciones. Por un lado, cargamos los datos de IMGpedia en un servidor que soporta nuestro operador de *similarity join*, para permitir que las consultas sobre las imágenes sean ahora dinámicas. Por otro lado, utilizando el conocimiento adquirido al definir un operador nuevo en un lenguaje de consulta, actualizamos una propuesta antigua para poder realizar *clustering* sobre los resultados de una consulta SPARQL, definiendo una sintaxis que permite al usuario seleccionar un algoritmo de *clustering* y proveer los parámetros necesarios.

Como me fue bien durante el magíster, cometí el error de pensar que el doctorado iba a ser fácil y exitoso. Sin embargo, tuve que lidiar con problemas de salud mental justo a la mitad de mi periodo de estudios, lo que por un lado mermó mi motivación para hacer cosas y, por otro, puso en peligro el proyecto de tesis. Soy afortunado de contar con una gran red de apoyo, lo que me ayudó a salir adelante, realizar tratamiento y finalmente concluir el trabajo. En el último año del doctorado, comenzó también la pandemia, lo que me forzó a terminar mi tesis desde casa y a realizar mi defensa a través de videoconferencia. El doctorado fue una experiencia muy difícil, pero ya habiendo terminado, creo que ha valido la pena, pues estoy haciendo el trabajo que siempre quise hacer: investigar y enseñar.

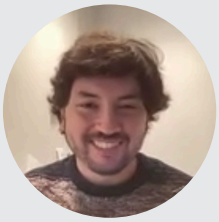
El trabajo con mis profesores guías, Benjamín y Aidan fue muy fructífero. Ellos me daban la independencia necesaria para probar cosas por mi cuenta y luego me ayudaban a discutir conceptos y a resolver dudas que iban surgiendo en el camino. Siempre me sentí muy apoyado por ellos y me enseñaron un montón, no solo sobre temas técnicos de mi tesis, sino también sobre cómo ser un investigador en general. Siempre voy a estar agradecido de haberlos tenido en mi camino.

Actualmente estoy haciendo un postdoc en el Departamento de Ciencias de la Computación de la Universidad de Linköping, en Suecia.

Tesis de doctorado

Data Structures and Algorithms for Analyzing DNA Sequences in Compressed Space

Estudiante: Diego Díaz
Profesor guía: Gonzalo Navarro



Mi interés por hacer un doctorado surgió a partir de los desafíos técnicos a los que me enfrentaba en mi trabajo. En ese momento era ingeniero de proyectos en el Centro de Modelamiento Matemático de la Universidad de Chile y estaba analizando datos de ADN. Mi problema era que los *datasets* que manejaba eran tan masivos, que los programas que necesitaba para procesarlos solo se podían ejecutar en el clúster de Beauchef (Leftraru).

Buscando investigadores que trabajaran en temas afines a los míos, di con el profesor Gonzalo Navarro. Tuvimos una primera reunión y me explicó cómo sería la dinámica si era su alumno. Sentí que era una buena oportunidad, así que postulé al doctorado del DCC, y afortunadamente quedé.

Mi primer tiempo en el programa fue un poco difícil, ya que mi formación en computación durante pregrado no fue tan rigurosa como la que se da en el DCC. Tuve que nivelarme rápidamente para no fallar en los cursos y eso me generó estrés. A pesar de esto, lo recuerdo como un tiempo muy enriquecedor donde aprendí varias cosas nuevas.

Mi investigación de doctorado tuvo como objetivo desarrollar algoritmos y estructuras de datos eficientes para procesar datos genómicos. Este tipo de colecciones son *strings* que codifican las secuencias de ADN de los genomas. El problema es que estos datos son tan masivos, que operar sobre ellos resulta muy caro computacionalmente. Otra característica de los datos de ADN es que son muy redundantes. Es decir, las mismas secuencias aparecen varias veces en el texto, con pequeñas variaciones. Uno puede sacar ventaja de este hecho para reducir los costos del análisis. La idea general consiste en determinar un pequeño diccionario de frases representativas del texto, realizar cálculos sobre estas frases, y luego extrapolar los resultados a las copias de las frases. Como los datos son repetitivos, el diccionario debería ser mucho más pequeño que el texto original, reduciendo de esta forma el número de operaciones redundantes.

La idea de las frases representativas suena razonable. Sin embargo, tiene sus salvedades. Por ejemplo, si el algoritmo que utilizas para construir el diccionario es más caro que el algoritmo que originalmente querías hacer más eficiente, entonces no tiene mucho sentido. En el doctorado, exploramos técnicas que tuviesen un buen *trade-off* entre nivel de compresión y el uso de recursos computacionales.

Cuando realizas un doctorado, no solo adquieres conocimiento en el área en la que trabajas. También hay otro tipo de aprendizajes que son claves para investigar, como ser capaz de explicar claramente tu trabajo a tus pares, ya sea de forma oral o escrita. Lo que me gustó del DCC es que también se preocupa de que las personas desarrollen este tipo de habilidades. Por ejemplo, el idioma oficial del programa es el inglés. Esto puede parecer una dificultad extra al principio, pero luego que te acostumbras, comunicar tus resultados se va a haciendo cada vez más fácil.

Actualmente estoy realizando un postdoctorado en la Universidad de Helsinki, en el grupo de algoritmos bioinformáticos. Me gustaría continuar trabajando en la academia en el futuro, donde pueda investigar los temas que considero interesantes.

Tesis de doctorado

Towards a Fine - Grained Linking Approach

Estudiante: Henry Rosales
Profesores guías: Aidan Hogan y Bárbara Poblete

Mi paso por el programa doctoral del DCC partió en el semestre de primavera de 2016. Mi primer desafío fue seleccionar el tema de investigación y el/la supervisor/a que se alineara a mis intereses. Si bien ya me había documentado al respecto mirando principalmente las publicaciones de los profesores en DBLP, me ayudó mucho asistir a las charlas que cada *profe* impartió en el ramo de Investigación. Mi línea de investigación hasta ese momento se orientaba hacia el reconocimiento de patrones y minería de datos, por lo que decidí acercarme a la profesora Bárbara Poblete, quien aborda temas afines. Tuve la suerte de que la profesora Bárbara me propusiera una cotutoría con el profesor Aidan Hogan y así poder unir áreas sumamente interesantes: minería de datos y web semántica.

El segundo desafío que enfrenté fue la selección del tema de investigación. Esto llevó varias semanas de debate junto con los tutores. Finalmente seleccionamos la tarea de *entity linking*. Esta tarea tiene como objetivo identificar las entidades en un texto en lenguaje natural y enlazarlas con sus entidades correspondientes en una *knowledge base/graph*. Luego de revisar por algunos meses el estado del arte de *entity linking* decidí enfocarme en la creación de un sistema que tuviera un enfoque multilingüe, ya que había poco trabajo relacionado. Los primeros trabajos fueron propuestos en el 2007 junto a la popularidad de Wikipedia y la creación de *knowledge bases* como Freebase y DBpedia.

Al ser *entity linking* una tarea relativamente reciente, había muchas ideas por desarrollar. Sin embargo, varias preguntas importantes no quedaban claras en la comunidad, por ejemplo: ¿qué entidades se deberían identificar y enlazar?, ¿son claros los mecanismos de evaluación?, ¿las métricas de evaluación actuales son suficientes?, entre otras. Eran pocas las preguntas que tenían respuestas, por tanto tuvimos que responderlas.

Optando por la definición más amplia de entidad, creamos una herramienta de anotación para crear nuestros propios *datasets* multilingües. Esta herramienta sirvió de base para anotar manualmente y semiautomáticamente un *dataset* en cinco idiomas. Con este *dataset* exploramos la factibilidad de adoptar traducción automática obteniendo resultados favorables. Por otra parte, propusimos una métrica basada en *fuzzy sets* para poder evaluar a un mismo sistema en diferentes dominios de aplicación. En resumen, cuando ya estaban las bases para crear el sistema de *entity linking* multilingüe que soñé ya tenía todos los requisitos para defender la tesis.

El doctorado me dejó muchas experiencias no solo relacionadas con la ciencia de la computación. Pude expandir mis horizontes más allá de América y darme cuenta de que: al pedir un Uber en Hong Kong te pueden recoger en un Tesla; en Bolonia (Italia) puede resultar difícil imprimir un póster; en Colombia no se baila la salsa de la misma forma que en Cuba; la espera de un vuelo de conexión en Estados Unidos puede retrasarse un día completo. Como dice una frase que escuché en la Escuela de Verano ISWS'18 en Italia y con la que no puedo estar más de acuerdo, "la investigación debe ser divertida, sino lo es, es porque algo estás haciendo mal".

Actualmente me encuentro realizando un postdoctorado en el Instituto Milenio Fundamentos de los Datos (IMFD). Estoy participando en la construcción de un *knowledge graph* junto a los profesores Aidan Hogan (Unversidad de Chile) y Renzo Angles (Universidad de Talca) sobre varias fuentes de datos. Por otro lado, tuve la oportunidad de dictar el ramo de Programación Avanzada en la Universidad de O'Higgins como profesor de jornada parcial.

Tesis de doctorado

Piecewise Adjacent Contours for Multicellular Structures in Fluorescence Microscopy Images

Estudiante: Jorge Jara
Profesores guías: Nancy Hitschfeld y Steffen Härtel

Mi motivación para hacer el doctorado surgió del interés en geometría e imágenes, aplicadas a células y sus componentes en distintos escenarios experimentales, con especial foco en imágenes tridimensionales de fluorescencia en vivo y de alta resolución. Mi tema de doctorado abordó algoritmos de geometría computacional y procesamiento de imágenes para representar membranas celulares dinámicas capturadas en estas condiciones. El título de mi tesis de doctorado es "Piecewise Adjacent Contours for Multicellular Structures in Fluorescence Microscopy Images", que puede traducirse a "Contornos Adyacentes por Tramos para Estructuras Multicelulares en Microscopía de Fluorescencia".

La microscopía de fluorescencia es una técnica óptica que permite observar células y organismos con imágenes 3D y, mediante el uso de múltiples marcadores fluorescentes, distinguir estructuras en distintos colores. Más aún, es posible realizar capturas en el tiempo y producir secuencias de video, ya que las muestras en observación pueden vivir durante varios días. La desventaja de esta técnica es que su resolución es limitada debido a fenómenos de dispersión de la luz, lo que impide distinguir objetos demasiado pequeños (en comparación con la microscopía electrónica, la mejor existente, pero casi imposible de usar con muestras vivas). Además, el comportamiento natural de los especímenes vivientes involucra movimientos y deformaciones muy rápidos que pueden hacer que las imágenes adquiridas pierdan aún más definición, viéndose borrosas. Aun así, la microscopía de fluorescencia ha sido clave para una colorida revolución de imágenes biomédicas en los últimas dos décadas, ya que ha hecho posible observar fenómenos 3D en tiempo real, como el desarrollo del cerebro de mosca, ratón, o pez cebra fluorescente. Este último fue utilizado como modelo para mi trabajo de tesis.

Desarrollé ALPACA, que es una abreviatura del nombre en inglés "ALgorithm for Piecewise Adjacent Contour Adjustment", traducido como "algoritmo para ajuste de contornos adyacentes por tramos", publicado el 2020 en la revista Journal of Microscopy (https://doi.org/10.1111/jmi.12887). ALPACA detecta y optimiza contornos de estructuras celulares adyacentes capturadas en imágenes de microscopía de fluorescencia. ALPACA se enfoca en secciones de contorno adyacentes que son cuantificadas para análisis de forma y organización, para aplicaciones a nivel subcelular, celular y supracelular. Combinamos ALPACA con algoritmos de la familia de contornos activos para interpolar y optimizar la forma de cada contorno y de sus secciones, y lo evaluamos con imágenes sintéticas y reales capturadas por microscopía confocal de disco rotatorio 3D *in vivo*, con el apoyo de usuarios con distintos niveles de experiencia en el sistema biológico modelo y microscopía de fluorescencia.

Tesis de magíster

Un método interpretable para clasificación general usando teoría de Dempster-Shafer

Estudiante: Sergio Peñafiel
Profesor guía: Nelson Baloian

En esta tesis abordamos el problema conocido en el campo de inteligencia artificial como "clasificación", esto es, dada una muestra caracterizada por un conjunto de parámetros, determinar a qué clase pertenece, de un conjunto predefinido. Las aplicaciones de este problema son tan variadas, que pueden ir desde identificar el tipo de subespecie de una flor dadas sus características, hasta determinar si un individuo tiene riesgo de contraer una enfermedad dado su historial médico. Para medir la eficacia de un método o modelo de clasificación normalmente se usa la precisión que tiene en clasificar una muestra en la clase que le corresponde. Hoy en día ha tomado fuerza la idea de que un modelo de clasificación debe ser también interpretable, es decir, el modelo debe inherentemente explicar por qué tomó tal o cual decisión.

En general los modelos más precisos son poco interpretables y viceversa, por lo que el desarrollo de uno que sacrifique algo de precisión pero que sea interpretable es una tarea interesante de resolver. Para la tesis, entonces, desarrollamos un modelo totalmente nuevo, basado en la teoría de la plausibilidad de Dempster-Shafer, que en precisión es comparable a los mejores que se conocen, pero es ampliamente interpretable. La motivación para el desarrollo de esta tesis fue una invitación de una empresa japonesa dedicada a la informática médica para desarrollar un modelo que permita predecir el riesgo de un individuo de sufrir un ataque al corazón, por lo que tuvimos la oportunidad de viajar a Japón para probarlo con datos recopilados por el Ministerio de Salud japonés, para una cierta área (prefectura) obteniendo muy buenos resultados. Los resultados se publicaron en dos revistas de alto impacto (IEEE Access y Artificial Intelligence with Applications).

Tesis de magíster

Implementando un reporte seguro de acoso sexual

Estudiante: Ilana Mergudich
Profesor guía: Alejandro Hevia

El trabajo de tesis de Magíster en Ciencias mención Computación de Ilana fue motivado por la necesidad de obtener un sistema eficiente para reportar situaciones de acoso y abuso sexual en forma responsable pero anónima, esto es, donde las acusaciones anónimas tuvieran la garantía de ser reveladas bajo condiciones que colectivamente se consideren adecuadas. Una instancia concreta de esta idea fundamental es un sistema informático, implementado como sistema distribuido, donde las identidades de todas las personas participantes (tanto acusadores/as como acusados/as) se mantienen anónimas mientras una condición específica no se alcance, por ejemplo, no exista un número mínimo (digamos dos) de acusaciones contra la misma persona. De ocurrir, se revelan las identidades y se puede proceder a una investigación formal.

Si bien el sistema propuesto por Ilana no es el primero en resolver este problema con un algoritmo distribuido criptográfico (dicho sistema existente se denomina WhoToo), su contribución principal fue mejorar WhoToo sustancialmente en aspectos internos claves. Para ello, rediseñó varios de los algoritmos criptográficos utilizados por otros igualmente seguros pero significativamente más eficientes. Para comprobarlo más allá de la teoría, Ilana incluso implementó un prototipo factible de ser usado en un contexto real, por ejemplo en una universidad. La fortaleza del trabajo de Ilana fue combinar nuevos algoritmos criptográficos con un diseño orientado a obtener resultados prácticos y eficientes.

El trabajo realizado en esta tesis fue publicado en el artículo "Implementing Secure Reporting of Sexual Misconduct – Revisiting WhoToo" (con Alejandro Hevia como coautor) publicado en octubre de 2021 en una importante conferencia internacional de criptografía (Latincrypt 2021), la cual se realiza en cooperación con la Asociación Internacional de Investigación Criptográfica.

Memoria de pregrado

Estructuras compactas dinámicas más eficientes para bases de datos de grafos con atributos

Estudiante: Dania de la Puente
Profesores guías: Gonzalo Navarro y Diego Arroyuelo

La memoria de Dania de la Puente, "Estructuras compactas dinámicas más eficientes para bases de datos de grafos con atributos", coguiada por Gonzalo Navarro con Diego Arroyuelo (Universidad Técnica Federico Santa María), aborda el problema de representar bases de datos de grafos en forma más eficiente en términos de espacio. Este tema es muy relevante en la actualidad porque las bases de datos de grafos de distintos tipos están apareciendo en todo tipo de aplicaciones como una forma más flexible que las relacionales para almacenar y buscar información, y están creciendo en tamaño. Una representación más compacta permite manejar mayores grafos en memoria principal (donde el procesamiento es mucho más rápido) o en dispositivos de memoria limitada como celulares.

Si bien existen ya algunas representaciones compactas para bases de datos de grafos, la memoria de Dania se centra en el caso menos estudiado de representaciones dinámicas, es decir, que permiten modificar las aristas del grafo al mismo tiempo que este se consulta. También se centra en un tipo de representación de grafos basada en *tries*, que ven la matriz de adyacencia del grafo como una grilla bidimensional y almacenan las aristas como las coordenadas de los puntos correspondientes, intercalando convenientemente sus dígitos para formar las secuencias que el *trie* almacena, del mismo modo que lo haría un *quadtree*.

La memoria tiene tres contribuciones importantes. La primera es completar una estructura que habían desarrollado los profesores guías, basada en lo que se llama un k^2-*tree* dinámico, para agregarle la operación de borrar aristas. El resultado muestra que se puede borrar tan eficientemente como se insertaban aristas en esta estructura. La segunda contribución es comparar la estructura resultante con varias representaciones alternativas de *tries* de coordenadas, algunas ya usadas antes con este propósito y otras aplicadas por primera vez a este problema. El resultado muestra que la nueva estructura adaptada por Dania resulta ser la más conveniente cuando se consideran todos los aspectos de la funcionalidad requerida. La tercera contribución fue integrar la nueva estructura en un sistema de bases de datos existente, que usaba una representación dinámica de k^2-*tree* anterior. Los resultados muestran que la nueva estructura es varias veces más rápida que la anterior, al precio de usar una módica cantidad de espacio extra.

Memoria de pregrado

Análisis del COVID-19 y sus correlaciones a nivel internacional

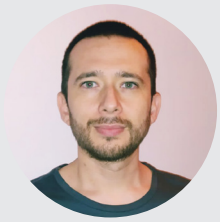
Estudiante: Tamara Novoa - Rodríguez
Profesor guía: Aidan Hogan

Al día de la fecha, aún no se entiende muy bien por qué hay tanta variación entre los diferentes países con respecto a los números reportados, per cápita, de casos y muertes asociadas al COVID-19. Se han planteado varias hipótesis, relacionadas, por ejemplo, con el clima, con la prevalencia de comorbilidades (como la obesidad, la diabetes, etc.), con la edad media de la población, con las políticas públicas del país, etc. Aunque algunas iniciativas ya han analizado esta situación, han estado enfocadas siempre en una hipótesis particular.

En el trabajo de título de Tamara Novoa (alumna del Departamento de Ingeniería Electrica), guiado por Aidan Hogan (profesor del DCC), generamos un "cubo" de datos multidimensional para intentar correlacionar diversas variables sobre países. Habiendo recolectado datos de varias fuentes, separamos las variables en dos categorías: aquellas referentes al COVID-19 (por ejemplo, número de casos, número de muertes, etc.), y aquellas no referentes al COVID-19 (por ejemplo, el PIB per cápita, emisión de dióxido de carbono, saldos promedios, uso de Internet, etc.). Luego buscamos variables correlacionadas entre ambas categorías. Los resultados aportaron nuevas perspectivas de la pandemia y fueron inesperados en algunos casos. Por ejemplo, los países que tienen mayor porcentaje de la población con acceso a instalaciones para lavarse las manos tendían a tener *más* casos de COVID-19. De manera más general observamos un posible factor latente de que los países más desarrollados tendían a tener más contagios y defunciones reportados.

Publicamos un artículo corto en la International Semantic Web Conference (ISWC) y el trabajo fue uno de los tres nominados para el premio "Best Demo".

Memoria de pregrado



Herramienta hidroinformática V.GeoLinkage: automatización de modelos hidrológicos integrados

Estudiante: Felipe Troncoso
Profesor guía: Pedro Sanzana y Nancy Hitschfeld

En el marco de un proyecto multidisciplinario cuyo objetivo fue llevar a cabo el diagnóstico para la gestión de explotación del acuífero Valle de Azapa, se desarrolló la herramienta V.GeoLinkage que permite trabajar con un modelo que integra en una sola visión la perspectiva superficial y subterránea del flujo de agua. Es común en este tipo de proyectos de recursos hídricos que exista una parte del equipo especializada en cada perspectiva (superficial y subterránea), contando con dos modelos para describir la misma red de drenaje. Este tipo de modelación integrada requiere previamente contar con la relación entre los dominios y topologías de ambas representaciones, las cuales generalmente tienen una discretización espacial diferente. El dominio superficial es típicamente segmentado en unidades hídricas: subcuencas, bandas de elevación o unidades de respuesta hidrológica, las cuales son representadas por triángulos o polígonos irregulares simples, mediante arcos y nodos. Por otro lado, el dominio subterráneo está organizado en unidades hidrogeológicas, que son representadas por grillas, triangulaciones, *quadtrees* o diagramas de Voronoi.

La herramienta desarrollada, V.GeoLinkage, encuentra cómo se relacionan estas representaciones para un modelo superficial de la plataforma WEAP (weap21.org) y un modelo subterráneo en la plataforma MODFLOW (www.usgs.gov). Desarrollamos una interfaz en Python que se integra dentro del sistema de información geográfica GRASS-GIS y extrae como mapas vectoriales ambas representaciones. Estos mapas son procesados y relacionados usando la integración con las herramientas que posee GRASS-GIS. Finalmente, se genera un archivo que almacena esta relación entre las discretizaciones espaciales, el cual es usado por WEAP para ejecutar el modelo acoplado.

La automatización lograda en el preprocesamiento permitió pasar de días a minutos, haciendo factible trabajar con escenarios que enriquecieron el diagnóstico con un mayor número de hipótesis, por ejemplo, un cambio en uso de suelo o una expansión espacial de zonas agrícolas, durante el periodo de estudio. En el portal de la Dirección General de Aguas del Ministerio de Obras Públicas se puede acceder al informe completo (snia.mop.gob.cl/sad/SUB5917v1.pdf). Adicionalmente el trabajo lo presentamos en la sesión de hidroinformática de la conferencia internacional European Geosciences Union (EGU), al cual se puede acceder en https://doi.org/10.5194/egusphere-egu21-13909.