Capítulo del libro "Innovación Basada en el Conocimiento Científico"
Editores: B. Santelices, F. Lund, T. Cooper, J.A. Asenjo.
Publicado por la Academia Chilena de Ciencias, año 2013.

7

# ÁREA INFORMÁTICA

## Yahoo Labs Latin America

**Ricardo Baeza-Yates (Coordinador)**,
Yahoo Labs Latin America, Santiago, Chile.
Universidad de Chile, Departamento de Ciencias de la Computación,
Santiago, Chile.

**Mauricio Marín**,
Universidad de Santiago, Departamento de Ingeniería Informática,
Santiago, Chile.

## Resumen

Este laboratorio es único en su tipo en Latinoamérica, siendo fundado en 2006 gracias a un convenio privado con la Fundación para la Transferencia Tecnológica de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. En todos estos años, el laboratorio ha producido innovaciones que impactan los servicios y tecnología de Yahoo y aportan al mundo científico por medio de publicaciones, patentes y otras formas de divulgación. El trabajo de investigación del laboratorio se centra en tecnologías de búsqueda y minería de datos de la Web. Esta labor ha permitido que investigadores nacionales y extranjeros tuvieran acceso a datos únicos en el mundo a través de posdoctorados, investigadores visitantes y asociados y pasantías de estudiantes de posgrado.

El laboratorio también ha contratado ingenieros de software egresados de las universidades nacionales, quienes participan en el desarrollo de optimizaciones y nuevas funcionalidades para los distintos componentes que forman parte de los servicios de Yahoo. Del mismo modo, en el ámbito nacional, el laboratorio colabora con las empresas que basan sus servicios en el uso de la Web. Para ello les transfiere soluciones tecnológicas avanzadas y desarrolla junto con ellas proyectos de investigación.

En este capítulo explicaremos el proceso de innovación que se realiza en el laboratorio día a día a través de dos ejemplos exitosos de algunas

de las nuevas tecnologías de impacto mundial inventadas en Chile. Estas innovaciones son públicas, en el sentido de que pueden usarse gratuitamente a través de los servicios que Yahoo presta en la Web, aunque hayan sido financiadas de forma privada o sean resultados de proyectos Fondef precompetitivos, donde participa el laboratorio junto con empresas y universidades nacionales.

## Introducción

Desde su inicio en marzo del año 2006, el laboratorio de I+D de Yahoo Labs en Chile ha tenido una producción científica fuertemente orientada a la investigación aplicada. Es decir, formulación y prueba de hipótesis a problemas de investigación que surgen de la constante necesidad de crear nuevas funcionalidades y optimizar el funcionamiento de los servicios que Yahoo pone a disposición de sus más de 700 millones de usuarios en todo el mundo.

Una característica relevante del tipo de investigación realizada es que, para lograr soluciones de utilidad práctica, esta debe considerar el comportamiento y las preferencias de los usuarios de Yahoo, lo que obliga a contar con acceso a fuentes de datos de la interacción de personas reales e infraestructura de procesamiento para grandes volúmenes de datos. Gracias a la relación con Yahoo Labs en Estados Unidos varias de las soluciones propuestas por el laboratorio de Chile se han transformado en proyectos de ingeniería destinados a servicios en producción.

Como subproductos del trabajo realizado en la investigación aplicada para Yahoo, se han obtenido más de 70 publicaciones indexadas en Scopus y 25 en ISI. La gran mayoría tiene como coautores a estudiantes de doctorado y magíster de las principales universidades nacionales. Estos trabajos también dieron origen a la presentación de 8 patentes en Estados Unidos, todas ellas ellas de propiedad industrial de Yahoo.

Algunos de estos resultados tuvieron difusión internacional acerca del impacto de la ciencia de la computación en la sociedad, con menciones en medios tan importantes como The Wall Street Journal y Scientific American. Es el caso del análisis de datos de Twitter durante el terremoto del 2010, en el que se demuestra que las informaciones verídicas se transmiten ampliamente, mientras que las falsas son desmentidas con rapidez por las mismas personas que usan esta red social de *microblogging*, que es la más extendida en el mundo. También tuvo difusión a nivel nacional un sistema de análisis en tiempo real de mensajes Twitter para predecir tendencias en las elecciones presidenciales.

Para los estudiantes de posgrado y pregrado, el laboratorio proporciona recursos destinados a investigación aplicada que de otra manera serían imposibles de obtener en universidades tanto en Chile como en el extranjero. Dichos recursos están compuestos por el acceso a diversas bases de datos para investigación, las que contienen datos generados

por usuarios reales de los distintos productos de Yahoo, el contacto con investigadores de Yahoo Labs, y la posibilidad de usar computadores de alto rendimiento para procesar dichos datos y probar las aplicaciones en un ambiente similar al obtenido en producción con usuarios reales. Esto ha permitido la titulación de más de 40 estudiantes de ingeniería en computación e informática de universidades nacionales.

El laboratorio recibió apoyo financiero de parte del programa de atracción de inversiones de alta tecnología de Corfo y de los programas de inserción de doctores y tesis de doctorado del sector productivo de Conicyt. Estos fondos tuvieron como objetivo principal la contratación en el laboratorio de investigadores con doctorados recientes y de quienes realizan posdoctorados durante dos o tres años en Chile. En la mayoría de los casos se trata de jóvenes beneficiados con el programa de Becas Chile para doctorados en Chile y en el extranjero.

Varios de estos investigadores jóvenes han sido posteriormente contratados como académicos en universidades nacionales, lo cual permite que continúen ligados al laboratorio mediante un esquema de proyectos de colaboración que incluye financiamiento para estudiantes, presentación de trabajos en congresos internacionales y asistencia a reuniones semestrales de Yahoo Labs en California ("Science Week"), donde se tiene la oportunidad de adquirir de primera fuente una visión global respecto del estado del arte en la disciplina.

Este esquema ha resultado ser exitoso puesto que ha permitido conservar a un gran número de investigadores y estudiantes relacionados con las temáticas de investigación abordadas en el laboratorio, y establecer y mantener un capital humano avanzado en Chile en el área de desarrollo de tecnologías para la Web, lo que beneficia no solo a Yahoo Labs en Chile, sino también a las propias universidades y empresas nacionales que utilizan la Web para prestar servicios a sus clientes.

El beneficio para las universidades proviene de la formación interna de equipos expertos en investigación aplicada y conocimiento real del proceso que va desde la idea hasta el producto final, lo que les permite aumentar su participación en fondos públicos de I+D y transferir a sus estudiantes conceptos de innovación cimentados en la ciencia, esto último entendido como productos de software basados en tecnologías Web y adoptados por muchos usuarios, de modo que dicho proceso está fuertemente relacionado con la solución a problemas de investigación necesarios para lograr productos exitosos.

La formación de profesionales e investigadores por la vía de trabajos de tesis en temas relevantes para Yahoo también tiene beneficios para las empresas nacionales, y esto es cada vez más evidente dada la ubicuidad y transversalidad creciente de productos de software basados en la Web y redes sociales. Estos productos, cuando son exitosos, se caracterizan por tener un alto potencial de crecimiento exponencial en cantidad de usuarios y volumen de datos gestionados en tiempo real.

Típicamente, el ciclo de vida respecto de la innovación en estos productos de software consiste en que una primera versión del producto se construye utilizando software existente de dominio público y se pone en producción en servidores de proveedores de servicios de computación en la nube (*cloud computing*). Esta es una estrategia razonable, puesto que en las primeras etapas de desarrollo e inserción en el mercado lo relevante es definir bien el concepto detrás del producto y su estrategia de monetización. El caso típico son los emprendimientos que apoya el programa Start-Up Chile de Corfo.

Sin embargo, cuando dichos productos son ampliamente adoptados por grandes comunidades de usuarios, surgen problemas de escalabilidad en el software que afectan la experiencia de sus usuarios, tanto respecto de la calidad de los resultados como del tiempo de respuesta. Si el software no es capaz de seguir el crecimiento exponencial de usuarios y volumen de datos, las posibilidades de fracaso son muy altas. Precisamente, estos tipos de problemas son los que se abordan en la investigación que se efectúa en Yahoo Labs en Chile y, por lo tanto, la formación de capital humano avanzado que realiza el laboratorio tiene relevancia para las empresas nacionales.

En las siguientes secciones de este capítulo se describen dos casos de estudio que ilustran el proceso de innovación llevado a cabo en el laboratorio. El primero tiene relación con un proyecto de I+D de impacto interno en Yahoo vinculado con optimizaciones a motores de búsqueda de publicidad. El segundo, con el desarrollo de un proyecto Fondef que fue posible gracias a la formación de capacidades de investigación al interior del laboratorio y su posterior difusión en las universidades participantes. En ambos casos se describe la solución encontrada para organizar equipos de I+D que posibiliten la realización de innovación basada en investigación aplicada.

## CASO DE ESTUDIO 1: Optimización de motores de búsqueda vertical

Los motores de búsqueda vertical son sistemas altamente optimizados para responder cientos de miles de consultas por segundo bajo un contexto bien específico, como generar la publicidad digital asociada a una búsqueda o al contenido de una página web. Por lo tanto, sus algoritmos y estructuras de datos son diseñados de acuerdo a requerimientos dictados por el tipo de trabajo que deben realizar. Para el caso de la publicidad se utilizan estructuras de datos diseñadas para recuperar textos pequeños que contienen los avisos publicitarios que mejor calzan con las consultas que llegan al buscador, los cuales deben ser desplegados junto con la respuesta a dichas consultas. Estas tienen la forma de un conjunto de palabras que en el caso de la publicidad tienden a ser numerosas.

La restricción para el tiempo máximo de respuesta para cada consulta es del orden de unas pocas decenas de milisegundos. Esta exigencia no es un tema menor si se considera que cada vez que un usuario del correo de Yahoo selecciona leer un correo, gran parte del texto se envía primero al motor de búsqueda de publicidad para recuperar avisos publicitarios pertinentes al texto y luego se construye la página HTML que es presentada al usuario como respuesta a su petición de lectura del texto del correo, todo lo cual debe ocurrir en una fracción de segundo.

Si se considera que los productos de Yahoo interactúan en todo momento con millones de usuarios, la eficiencia de las soluciones propuestas es un requerimiento de gran importancia. La solución es desplegar el motor de búsqueda en centros de datos con cientos de computadores, cada uno con muchos procesadores, los cuales están dedicados exclusivamente a ejecutar las computaciones asociadas a la solución de las consultas. Para acelerar este proceso se utilizan estructuras de datos distribuidas, llamadas índices invertidos, en los procesadores del centro de datos.

Los índices invertidos están compuestos de listas de documentos en las que, en nuestro caso, cada uno de ellos es un aviso publicitario. Se tienen tantas listas como palabras relevantes existan en la colección de documentos (ver Figura 1). Por lo tanto, el primer paso para resolver una consulta es recuperar la lista de documentos que contengan las palabras que forman la consulta. Luego de esto se aplica un algoritmo para ordenarlos de acuerdo a una métrica de relevancia y se seleccionan aquellos de mayor puntaje como respuesta a la consulta.

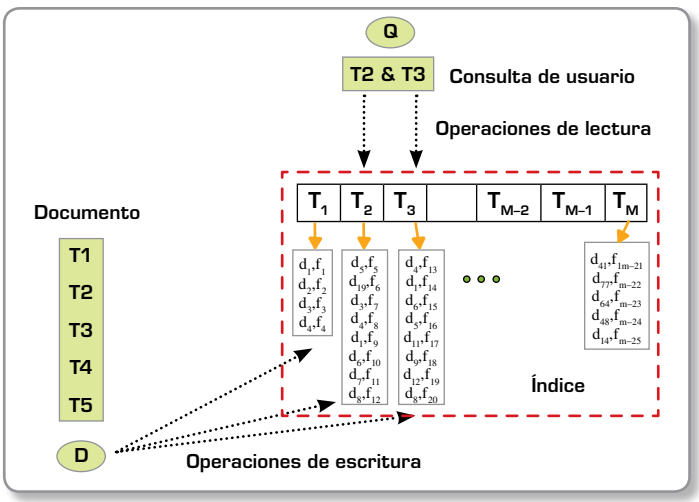


Figura 1. Índice invertido y su relación con la solución de consultas de usuarios y la inserción de nuevos documentos. Para calcular las respuestas a las consultas es necesario leer las listas invertidas de los términos que componen cada consulta. Para insertar un nuevo documento, es necesario actualizar (escribir) las listas invertidas de los términos contenidos en él.

El objetivo de este proyecto fue estudiar optimizaciones a diversos aspectos relacionados con la implementación, despliegue y operación del motor de búsqueda vertical. Se destinaron dos equipos de trabajo al proyecto que estaban constituidos por personas que ya venían trabajando en investigación con anterioridad en el laboratorio de Yahoo, de manera que el problema de la afinidad entre personas y conocimiento profundo de las potencialidades de cada uno ya estaba resuelto de antemano al proyecto.

Como es sabido, cada equipo debe tener un propósito que motive y dé sentido al esfuerzo que debe realizar. Ciertamente que trabajar en torno a un sistema real de gran escala, proporcionado por una empresa de la importancia de Yahoo para la disciplina, presenta desafíos y oportunidades de crecimiento profesional que justifican el esfuerzo. No obstante, en equipos compuestos por universitarios con vocación por la investigación lo anterior no necesariamente representa en sí mismo una motivación que satisfaga de forma completa los intereses personales.

Es decir, los tópicos abordados en el proyecto debían ser lo suficientemente genéricos como para constituirse en una fuente de progreso en la carrera académica que permitiera la generación de publicaciones desarrolladas en torno a líneas de investigación amplias, las cuales debían dar cabida a intereses adicionales, tales como la formación de investigadores a nivel de doctorado, sin entrar en conflicto, por ejemplo, con temas de derechos de propiedad industrial.

En este contexto, el *big picture* que mueve a la comunidad científica del área es que actualmente los centros de datos consumen del orden del 2 % al 5 % de la energía eléctrica mundial. Las estimaciones no son precisas, pero algunos especialistas mencionan que el año 2012 se requerían 30 centrales nucleares para generar la energía eléctrica necesaria destinada a alimentar los centros de datos del mundo. En particular, dado el volumen de recursos utilizados y el alto tráfico de requerimientos de usuarios que deben atender, los centros de datos que alojan los distintos tipos de motores de búsqueda para la Web son responsables de una fracción importante de ese consumo de energía.

Una de las contribuciones del laboratorio en Chile ha sido mostrar que con algoritmos eficientes de procesamiento de consultas es posible reducir la cantidad de recursos de hardware requeridos para atender una determinada carga de trabajo, y por lo tanto, al disminuir la redundancia de recursos, es posible reducir la cantidad de energía consumida. Esto es importante pues el mayor consumidor de energía eléctrica en un centro de datos es el aire acondicionado y este consumo es proporcional al número de computadores en el mismo.

El proyecto consistió en (1) desarrollar algoritmos de compresión de índices para permitir almacenar mayor cantidad de datos y realizar procesamiento eficiente de consultas en cada procesador, y (2) desarrollar estrategias de asignación de recursos que permitan determinar con pre-

cisión la cantidad de procesadores requeridos para atender una determinada carga de trabajo o tráfico de consultas. Cada parte fue abordada por un equipo distinto compuesto de ingenieros e investigadores contratados por el laboratorio, y la participación de tesistas de doctorado y magíster, donde los investigadores realizan la cosupervisión de las tesis en conjunto con los profesores de las respectivas universidades, en este caso la Universidad de Chile, la Universidad de Santiago de Chile y la Universidad Técnica Federico Santa María.

## SUBPROYECTO 1: Compresión de índices invertidos

Una reducción del consumo de memoria del índice invertido puede conducir a una reducción de la cantidad total de procesadores requeridos para dar servicio al tráfico de consultas. No obstante, el tiempo de respuesta de cada consulta no puede degradarse al punto de superar la restricción dada por el tiempo máximo de respuesta. Descomprimir una lista de documentos para luego aplicar un método de *ranking* sobre esos documentos es un proceso más costoso en tiempo de ejecución que trabajar sobre la lista original.

Una vez construidas a partir de la colección de documentos, las listas invertidas contienen secuencias de números enteros que se utilizan para identificar cada documento. La solución propuesta parte de la observación de que en las listas invertidas ocurren secuencias largas de identificadores de documentos cuyos valores difieren en una unidad. Por lo tanto, dichas secuencias pueden ser sustituidas por una representación comprimida del total de identificadores en cada secuencia, lo cual conduce a un ahorro de espacio, puesto que ya no es necesario almacenar los valores de cada uno de los identificadores porque se pueden calcular durante el proceso de *ranking* de documentos. Este simple mecanismo permitió superar por un amplio margen el método de compresión de listas invertidas que se utilizaba en el motor de búsqueda hasta ese momento.

Por supuesto que esta descripción es una sobresimplificación del proceso de I+D que fue necesario realizar, ya que se consideraron varios otros aspectos que van desde la implementación eficiente de los algoritmos hasta su integración y funcionamiento correcto en la arquitectura de software del motor de búsqueda. Respecto de la investigación, se debió estudiar la adaptación de métodos de compresión existentes para considerar las secuencias de identificadores similares y diseñar heurísticas para agrupar documentos relacionados, de modo de aumentar la probabilidad de que dichas secuencias tengan un largo mayor y ganar de esta manera en compresión.

Posteriormente, el esquema de compresión fue extendido para permitir almacenar junto con los identificadores de documentos las posiciones donde la palabra asociada a la lista invertida aparece dentro del documento. Esto permitió extender las funcionalidades del motor de bús-

queda con métodos de *ranking* de documentos que otorguen un mayor puntaje a los documentos que contienen las palabras de búsqueda más cercanas unas con otros dentro del documento. Antes de este trabajo no era posible aplicar ese tipo de *ranking* sobre los documentos indexados por el motor de búsqueda.

Como líneas de investigación para tesistas e investigadores surgidas a partir de este trabajo, podemos mencionar las siguientes:

- [Tesis de magíster] Hasta ahora, el motor de búsqueda ha sido sólo visto como un conjunto de procesadores que mantienen un índice invertido para resolver consultas. Si pensamos en incluir *rankings* de documentos que consideren las posiciones de las palabras de búsqueda dentro de los documentos, entonces, ¿por qué no considerar también los documentos como parte del proceso de solución de consultas? En particular, un trabajo de tesis de magíster [M1] ha sido dedicado a mostrar que en varios casos no es necesario almacenar explícitamente las posiciones de las palabras en los documentos, sino que es suficiente con almacenar cada documento debidamente comprimido y calcular las posiciones solo para un subconjunto de los documentos con mayor probabilidad de estar dentro de los de mejor puntaje mediante la descompresión de estos para encontrar las posiciones. Resulta que para motores de publicidad esta es la mejor opción, puesto que se trata de consultas con gran número de términos y documentos pequeños. Estos resultados dieron lugar a dos publicaciones en la mejor conferencia del área [SIGIR 2012] [SIGIR 2013].
- [Tesis de doctorado] Una tesis doctoral [D4] derivada de este proyecto plantea un cambio radical de enfoque. En lugar de mirar el proceso de *ranking* y despliegue de resultados como dos tareas independientes, es decir, índice invertido y luego recuperación de los documentos seleccionados, la tesis plantea utilizar estructuras de datos llamados Wavelet Trees (WT) para representar textos comprimidos autoindexados. Los WT no han sido aplicados al contexto de motores de búsqueda, pero se anticipa que permiten tratar todos los documentos asignados a un procesador como un solo documento comprimido sobre el cual es posible realizar búsquedas eficientes. La ventaja está en que permiten realizar *rankings* de documentos basados en las posiciones de las palabras en cada uno de estos últimos. La hipótesis es que los WT, al concentrar en un mismo espacio de memoria tanto el texto como el índice de búsqueda, representan una alternativa más eficiente respecto de la cantidad total de procesadores requeridos para servir una determinada carga de trabajo. Esta línea de trabajo ha dado lugar a una publicación de conferencia [SPIRE 2010] y otra en una revista del área [IPM 2012].
- [Tesis de doctorado] El método de *ranking* de documentos utilizado por el motor de búsqueda se ha constituido en un estándar de la industria. Si cada uno de los procesadores involucrados en la solución

de una consulta ejecuta el mismo método de *ranking*, entonces una pregunta interesante surge cuando uno considera que no es necesario pedir a todos los procesadores la misma cantidad de respuestas, como ocurre actualmente. Por ejemplo, si la consulta se envía a 100 procesadores y solo estamos interesados en los 10 mejores resultados, entonces a lo más 10 procesadores serán capaces de aportar resultados dentro de los 10 mejores. Enviar la consulta a 100 procesadores y solicitar a cada uno 10 resultados es ciertamente un desperdicio de uso de recursos de cómputo con el respectivo consumo de energía innecesario; el problema es encontrar una heurística que permita predecir con cierta probabilidad de éxito cuántos resultados solicitar a cada procesador. Un primer resultado en esta línea ha sido publicado en [Euro-Par 2013] y es parte de una tesis doctoral [D6]. Las listas invertidas almacenan las frecuencias con que cada palabra aparece referenciada en cada documento. La propuesta de esta tesis consiste en representar la distribución de estas frecuencias con series de Fourier, de manera de utilizar los coeficientes de la serie para clasificar las listas invertidas distribuidas en los procesadores. Una aplicación de esta idea consiste en utilizar los coeficientes para determinar el número de resultados que es necesario solicitar a cada procesador durante el procesamiento de una consulta.

- [Tesis de magíster] Los procesadores sobre los cuales es desplegado el motor de búsqueda habitualmente poseen muchos procesadores que permiten la ejecución eficiente de muchas hebras (*threads*) de ejecución. Estas hebras pueden ser utilizadas para aumentar el nivel de concurrencia en el procesamiento de consultas. Aquí surge un problema de planificación de tareas donde dos o más hebras pueden ser destinadas a acelerar el procesamiento de consultas y reducir de esta manera el tiempo de procesamiento de cada una. Esto requiere del diseño de una estrategia de asignación dinámica de hebras a base de una predicción del tiempo de ejecución de la consulta. Esta predicción debe ser realizada en línea, a medida que se recepcionan las consultas en el procesador, para lo cual se exploran métodos basados en aprendizaje de máquina. Este trabajo es parte de una tesis de magíster [M4] y surge de publicaciones anteriores de los investigadores del proyecto [CIKM 2010] [Euro-Par 2008].

Dos de los tesistas participaron desde el comienzo en el proyecto actuando como ingenieros de apoyo y recopilación del estado del arte en métodos de compresión de índices invertidos, lo que les permitió adquirir conocimiento de primera línea en el área de trabajo de sus respectivas tesis. En general, los trabajos de tesis mencionados dan cuenta de las líneas de investigación derivadas del proyecto que se desarrollan actualmente en el laboratorio de Yahoo.

## SUBPROYECTO 2: Planeación de capacidad de motores de búsqueda

Otro problema de uso eficiente de los recursos computacionales desplegados en el centro de datos tiene que ver con la determinación de cuántos recursos son realmente necesarios para una determinada carga de trabajo, y cuál es la cantidad requerida para una cierta predicción del tráfico de consultas en el mediano plazo.

La arquitectura del motor de búsqueda estudiado en el proyecto está formada por un conjunto de servicios. Los servicios principales son tres: (1) "Front-Service", el cual es un conjunto de procesadores encargado de recepcionar las consultas, enviarlas a los otros dos servicios para obtener la solución y remitir la respuesta al usuario; (2) "Caching-Service", el cual contiene respuestas calculadas anteriormente para las consultas más frecuentes, y (3) "Index-Service", el cual contiene el índice invertido para calcular la respuesta a la consulta si esta no es encontrada en el "Caching-Service". La Figura 2 muestra un esquema general de la arquitectura.

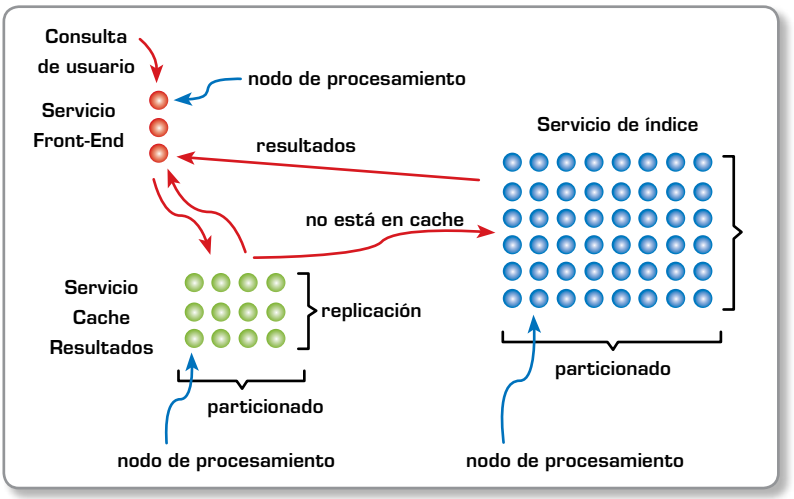


Figura 2. Componentes principales del motor de búsqueda de publicidad. Cada círculo indica un procesador y las flechas indican el camino seguido por las consultas de los usuarios del motor de búsqueda. Las consultas llegan a uno de los procesadores del "Front Service", luego se envían a los procesadores del "Results Cache Service", y si la respuesta no es encontrada en el cache, la consulta se envía a los procesadores del "Index Service".

Cada servicio es particionado en un conjunto de procesadores y el contenido de cada procesador es replicado en procesadores adicionales para mejorar el rendimiento y para tolerar fallas de procesadores. Por lo tanto, poder dimensionar adecuadamente la cantidad total de procesadores requeridos para procesar las consultas sin llegar a sobredimensionar excesivamente los recursos desplegados en el centro de datos es un

tema relevante desde el punto de vista de la operación económica del motor de búsqueda.

El problema planteado es del tipo de búsqueda del óptimo en un espacio combinatorial donde cada punto del espacio está dado por una cantidad específica de particiones y réplicas para cada servicio. El desafío consistió en modelar un sistema complejo, como el motor de búsqueda, para poder determinar la configuración óptima de los servicios.

La solución encontrada consistió en tres pasos. Primero se formuló un modelo del sistema basado en teoría de colas con sus respectivas fórmulas de rendimiento, suponiendo que las consultas circulan unas de otras por la red en forma independiente. Esta simplificación permite reducir significativamente el espacio de búsqueda a un 5 % de configuraciones posibles de contener el óptimo. Para las configuraciones restantes se recurre a una simulación discreta del rendimiento del motor de búsqueda. Para esto se creó una metodología de modelación y simulación basada en una serie de programas de *benchmark* diseñados para medir el costo de las operaciones relevantes en cada servicio.

Además, el hardware fue representado utilizando modelos de computación paralela diseñados para recrear las características relevantes del costo de computación y comunicación del proceso de solución de múltiples consultas concurrentes en el motor de búsqueda. También se contempla el uso de redes de Petri para facilitar la especificación y verificación del modelo de simulación. Una vez encontrado el óptimo mediante simulación, como último paso se aplica una heurística de particionamiento de un grafo en un conjunto de grupos de procesadores. En este caso, los nodos del grafo representan procesadores y los arcos entre pares de nodos indican el volumen de comunicación entre procesadores.

El método propuesto representa una innovación en el estado del arte, puesto que no se conocen métodos efectivos de planeación de capacidad para motores de búsqueda. La bibliografía solo reporta métodos teóricos y poco realistas para la complejidad de un sistema como el abordado en el proyecto. La clave está en formular modelos de simulación lo suficientemente precisos para obtener resultados cercanos a la realidad, pero a la vez lo suficientemente livianos para permitir ejecuciones del respectivo programa de simulación en tiempos tolerables para los administradores del centro de datos. El nivel de detalle del modelamiento del costo de las computaciones y el costo del hardware tienen un impacto directo en el tiempo de ejecución de los programas de simulación.

Al igual que en el subproyecto 1, el equipo de trabajo estuvo apoyado por tesistas de magíster y doctorado como las siguientes:

- [Tesis de magíster] Parte del trabajo realizado fue abordado mediante una tesis de magíster [M3] destinada a estudiar la utilización de redes de Petri (*timed colored Petri nets*) como método formal para modelar el costo del procesamiento de consultas en el motor de búsqueda y verificar modelos. La tesis fue finalizada durante el transcurso del

proyecto, y dio lugar a una publicación en una conferencia [ICATPN 2012] y posteriormente una versión extendida fue aceptada en una revista [FINF 2013]. El tesista actualmente está cursando un doctorado bajo la supervisión el mismo equipo del proyecto.

- [Tesis de doctorado] Adicionalmente, un tesista de doctorado [D5] está estudiando extensiones a la metodología de planeación de capacidad mediante la incorporación de un formalismo de modelación y simulación de gran poder de expresividad para sistemas de mayor complejidad. Para esto, el estudiante realizó una estadía de 4 meses con un grupo especializado en este tema en la Universidad de Carleton en Canadá. También realizó mejoras a la predicción del costo de comunicación entre procesadores, lo cual fue publicado en [PDP 2013]. Por otra parte, para que la metodología de planeación de capacidad de motores de búsqueda sea de utilidad práctica en un centro de datos, es relevante reducir los tiempos de ejecución del conjunto de simulaciones requeridas para encontrar el número óptimo de procesadores y determinar el mapa de despliegue de servicios en estos procesadores. Los tiempos de ejecución de estos simuladores superan la media hora y, por lo tanto, el tesista tuvo la tarea de evaluar distintas estrategias de paralelización de estos simuladores con el objetivo de ejecutarlos sobre varios procesadores y reducir de esta manera el tiempo de ejecución de cada simulación. La paralelización involucra la solución a un problema complejo de sincronización de eventos de simulación que ocurren en paralelo en distintos procesadores, para el cual se han propuesto varios protocolos de sincronización de eventos. La metodología de modelación y simulación debe ser extendida para incluir el protocolo de sincronización de eventos de mejor rendimiento.
- [Posdoctorando] El proyecto también contó con la participación de un investigador joven que ingresó al laboratorio a realizar un posdoctorado. Su trabajo consistió en validar experimentalmente una hipótesis que el equipo de trabajo había planteado como solución al problema de la paralelización de las simulaciones. Esta consistió en adoptar un enfoque de sincronización aproximada de eventos paralelos, donde la calidad de la aproximación es controlada mediante un método de ajuste automático del avance en el tiempo de simulación en cada procesador. Este método fue desarrollado por el equipo del proyecto y la hipótesis era que esta estrategia de simulación paralela aproximada iba a permitir la obtención rápida de resultados aproximados muy similares a los de la simulación secuencial o paralela exacta del mismo sistema. Es decir, resultados con una precisión de sobre el 95 % con tiempos de ejecución mucho menores que la simulación paralela exacta. El estudio experimental realizado por el posdoctorando permitió validar esta hipótesis y los resultados fueron publicados en [PADS 2013].

Los trabajos previos que sirvieron de base para formar un equipo especializado en este tema tienen relación con una tesis doctoral [D1] y proyectos iniciales que derivaron en publicaciones donde, para validar hipótesis, era necesario formular modelos de predicción del rendimiento de motores de búsqueda para la Web [Euro-Par 2011] [HPDC 2010] [PARCO 2010] [ECIR 2010].

## CASO DE ESTUDIO 2: Proyecto Fondef "Observatorios de la Web en tiempo real"

Las capacidades formadas en recursos humanos y colaboración en investigación al interior del laboratorio permitieron la formulación y adjudicación de un proyecto Fondef (D09I1185) relacionado con la construcción de tecnología escalable para observar tendencias en la Web y redes sociales en tiempo real. Fue necesario identificar y establecer alianzas con empresas nacionales de base tecnológica que tuvieran experiencia en el mercado para este tipo de tecnología. Se contó con el patrocinio de la Universidad de Chile, la Universidad de Santiago y la Universidad de Concepción.

La innovación principal introducida en el proyecto tiene relación con la recolección e indexación en el espacio y tiempo de objetos en tiempo real, y aplicar sobre estos objetos un conjunto de operadores espacio-temporales. A partir de esto se propone construir productos de software con mayor valor que los disponibles en el mercado. En este contexto, el concepto de "objeto" es genérico, pudiendo representar a personas o entidades mencionadas en la Web y redes sociales, y el espacio puede ser geográfico o virtual. Por ejemplo, una aplicación de esta idea puede ser la situación representada en la Figura 3, donde el espacio está constituido por medios de prensa agrupados por afinidad editorial, los objetos representan personas, y los operadores además de responder a consultas como la descrita en la figura pueden incluir cosas adicionales, tales como el análisis de los textos que referencian al objeto (discurso positivo, negativo o neutral), y con los cuales otros objetos similares al analizado han tenido coocurrencia en el tiempo y el espacio.

En general, el proyecto plantea la construcción de los siguientes componentes: (1) Clasificador de documentos, (2) detector de tópicos emergentes, (3) detector de comunidades emergentes, (4) analizador de opiniones, (5) detector de sentimientos, (6) etiquetador social colaborativo, (7) seguidor de series de tiempo, (8) identificador de entidades, (9) operadores espacio-temporales, y (10) recolector, indexador y buscador espacio-temporal.

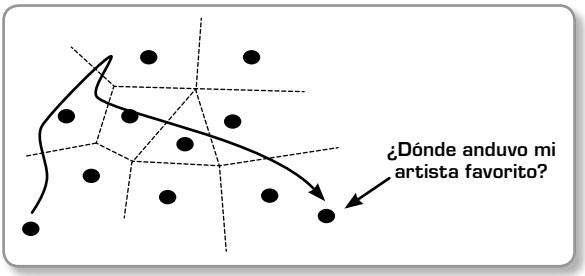


Figura 3. Seguimiento a entidades (personas) distribuidas en el espacio (medios de prensa) a lo largo del tiempo.

La estrategia de investigación y desarrollo aplicada se describe a continuación. Para facilitar la independencia del trabajo de los equipos de I+D, se establecieron subproyectos con cada una de las cuatro empresas participantes en el proyecto. Al igual que en el caso de estudio anterior, en cada subproyecto se formaron equipos integrados por investigadores, ingenieros de software y estudiantes tesistas de magíster y doctorado [D2] [D3] [D7] [D8] [M2]. La diferencia está en la interacción con cada empresa socia del proyecto. A cada empresa la ubicamos entre el equipo de I+D y los clientes de la tecnología desarrollada, donde es la propia empresa la que posee una relación de confianza con un cliente determinado, el cual es utilizado para probar el prototipo desarrollado y estudiar el mercado para el producto. La Figura 4 describe el proceso iterativo empleado en cada subproyecto (Plug-In). El proceso es conducido por el emprendedor detrás de la empresa, quien es el primer interesado en llevar pronto los resultados al mercado.

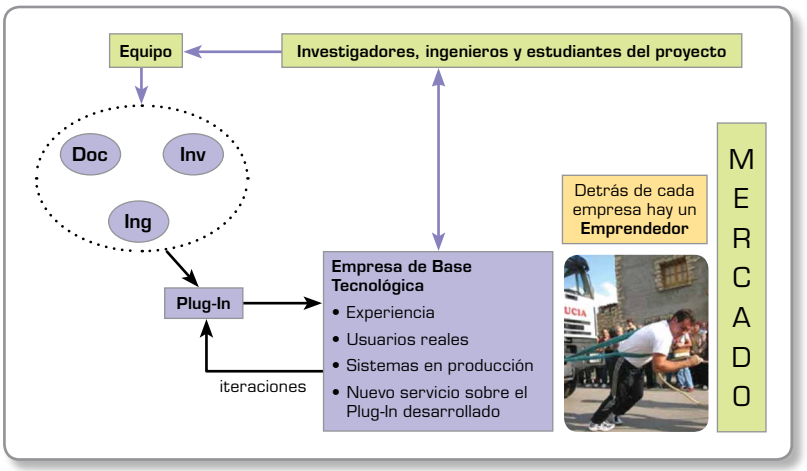


Figura 4. Estrategia para la colaboración entre un grupo de investigación y una empresa de base tecnológica.

Otro aspecto importante del proyecto fue su organización en los llamados Plug-Ins (componentes), cada uno asociado a un subproyecto de I+D con equipos distintos, los cuales son construidos sobre una arquitectura de software común tal que sea posible su integración para construir distintos productos que permitan observar la Web en tiempo real. De esta manera, los equipos de I+D pueden trabajar en paralelo en cada subproyecto, pero los resultados finales pueden ser integrados en distintos productos o ser utilizados de modo independiente. La Figura 5 presenta una vista general de los productos del proyecto y la estrategia de monetización de los Plug-Ins desarrollados.

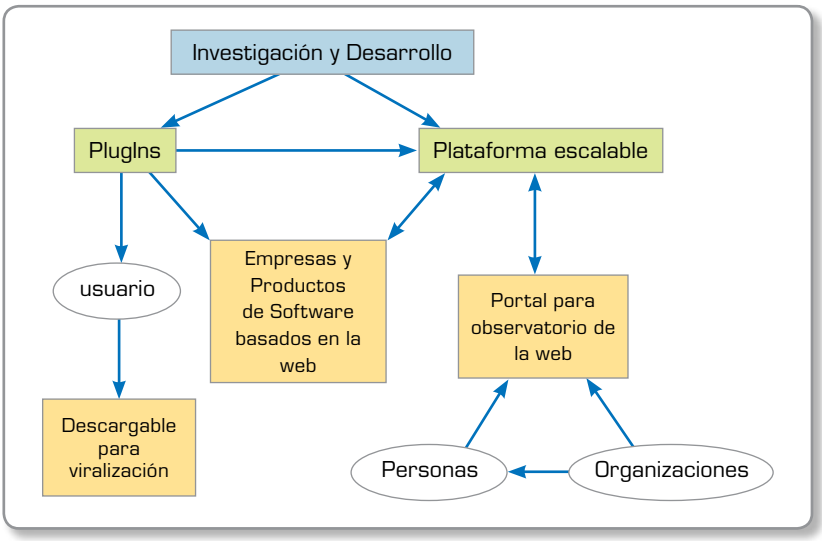


Figura 5. Organización de productos para la estrategia de monetización del proyecto Fondef.

Uno de estos subproyectos dio lugar a un emprendimiento en Silicon Valley por parte de los socios de la empresa. Se construyó un producto de software para el seguimiento y gestión de los contactos que sus usuarios mantienen en diversas plataformas sociales y correo. El objetivo de la empresa es utilizar el sistema desarrollado para crear una plataforma de comercio electrónico basado en recomendaciones realizadas por personas conocidas (contactos) de los usuarios. El sistema desarrollado por el proyecto Fondef representa la tecnología base sobre la cual está construido el producto. Se trata de un motor de gestión de contactos que permite mantener relaciones de afinidades en espacio y tiempo entre usuarios, y permite unificar contactos provenientes de distintas fuentes. Por ejemplo, una persona determinada podría estar representada con distinta información en redes sociales diferentes. La extensión del producto para gestionar proveedores de servicios ha sido apoyada por inversionistas de capital de riesgo y por el programa de innovación empresarial de Corfo.

Otro producto de software originado a partir de uno de los Plug-Ins desarrollados en el proyecto Fondef tiene que ver con un producto para el análisis de correos electrónicos que una de las empresas socias del proyecto estuvo dispuesta a impulsar en el sector de servicios bancarios. La motivación detrás del producto es la siguiente: Se sabe que el correo electrónico es un medio ubicuo de comunicación entre personas en una amplia variedad de tipos de empresas y organizaciones. Diariamente, las personas deben enviar y responder decenas (si no cientos) de correos electrónicos a clientes, subordinados, pares y superiores dentro de la estructura jerárquica de cada organización en la empresa o entre organizaciones dentro de la misma empresa o entre empresas. En este contexto, al cabo de un día laboral, semana, mes o incluso año de intercambio intensivo de correos electrónicos con distintas personas, a muchos usuarios les debería ser atractivo contar con una herramienta de software que les permita analizar el texto de sus correos electrónicos y entregar como resultado distintos perfiles de la comunicación sostenida con los destinatarios.

Este subproyecto fue posteriormente apoyado por financiamiento de Innova Corfo L2, donde la propuesta es mejorar la calidad de los perfiles de correos mediante la incorporación de la cultura y jerga del país. Un objetivo importante de este segundo proyecto I+D es lograr una metodología que permita generar versiones del producto que sean pertinentes a la cultura y jerga de otros países de la región iberoamericana, y ampliar de esta manera el mercado para el producto.

La realización del proyecto Fondef también se diferencia del caso de estudio anterior, en que el nivel de riesgo es mucho mayor. Al comienzo del proyecto las ideas para posibles productos estaban muy poco definidas y a lo largo de su desarrollo fueron las empresas socias las que dictaron los requerimientos. El proyecto también contó con una participación mucho mayor de tesistas de las universidades participantes. Esto, si bien tiene la ventaja de que permite abordar los temas más indefinidos y de mayor riesgo al inicio del proyecto, tiene el problema de los derechos de propiedad industrial, para el cual cada institución presentaba distintas soluciones. Por otra parte, algunas de las soluciones desarrolladas por los tesistas no encontraron aplicación inmediata en los productos principales del proyecto, puesto que en varios casos la evolución del producto siguió los intereses de la empresa asociada al producto, los cuales no necesariamente calzaron con los intereses del profesor guía de tesis y del estudiante.

Se optó por crear un observatorio de acceso público como una plataforma para probar las distintas soluciones encontradas por los tesistas para implementar los Plug-Ins del proyecto, y darle de esta manera mayor visibilidad nacional al proyecto y a sus resultados. Esta plataforma adoptó la forma de un "Observatorio Político" dedicado a observar la mensajería de Twitter y mostrar tendencias en las elecciones primarias

y presidenciales del 2013. Actualmente, este sistema está funcionando a modo de prototipo y se están incluyendo en él varias de las soluciones de tesistas respecto de operadores avanzados, tales como detección de usuarios líderes de opinión, comunidades emergentes, detectores de tópicos emergentes, etc.

Por otra parte, con el fin de resolver las limitaciones de las instituciones participantes respecto de participación en empresas destinadas a comercializar productos de software, se optó por un modelo de negocios de monetización de la tecnología basada en licenciamiento y un modelo de escalabilidad del negocio basado en sublicencias para empresas con influencia en otros mercados geográficos. La Figura 6 presenta una vista general del modelo de negocios.

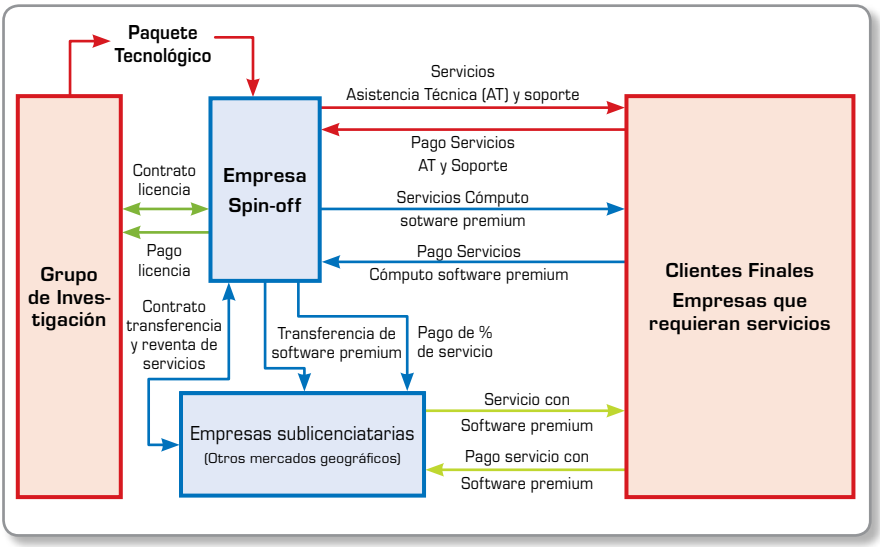


Figura 6. Modelo de negocios para la tecnología generada en cada producto del proyecto Fondef.

## Conclusiones

En los dos casos de estudio sobre la experiencia de Yahoo Labs en Chile se puede observar que es posible realizar investigación aplicada conducente a lograr un impacto en la industria y a realizar actividades académicas de formación y publicaciones científicas. Sin bien en varios aspectos la experiencia Yahoo Labs en Chile ha sido exitosa, también existen aspectos que es posible mejorar respecto a la relación con el medio nacional.

Nuestra experiencia indica que en las universidades se tiende a confundir investigación en aplicaciones de ciencia de la computación con investigación aplicada. La primera es una actividad de corte más académico, habitualmente bien valorada en Fondecyt porque genera numerosas publicaciones en revistas indexadas por el ISI Web of Science, mientras

que la segunda es una actividad fuertemente ligada a un proceso real, el cual va dictando continuamente los requerimientos de las soluciones y va planteando nuevos desafíos en investigación. La presencia de Yahoo Labs en Chile es un aporte en esta segunda línea, y nuestra impresión es que este tipo de investigación no se reconoce adecuadamente. Esto se refleja en la poca valoración que reciben las publicaciones en congresos y patentes de los posdoctorandos que han sido parte de Yahoo en el programa Fondecyt de Iniciación. Esto tiene un impacto negativo en la carrera de los investigadores jóvenes dedicados a realizar investigación aplicada en el área. Sin embargo, hemos visto claramente que estos son capaces de generar valor al país en el ámbito de innovación basada en investigación aplicada.

En particular, en las áreas de investigación abordadas por Yahoo en Chile, la publicación en determinados congresos tiene mayor impacto que en las revistas del área indexadas por el ISI Web of Science. Varios estudios bibliométricos dan cuenta de esta realidad, la cual se ha ido acrecentando en los últimos años, donde se observa que la comunidad del área concentra sus publicaciones principalmente en algunas conferencias de gran impacto.

Otro aspecto relevante tiene que ver con la propiedad industrial de las tesis que realizan los estudiantes en el contexto de proyectos donde están involucradas empresas que exigen la propiedad de todos los derechos. Cada universidad tiene sus propias normas respecto de tesis de estudiantes. La solución que ha encontrado el laboratorio es que en la etapa inicial de los tesistas; estos realizan labores de ingenieros de software y estudian el estado del arte en sus temas específicos, todo lo cual les permite participar productivamente en un equipo de investigación para llegar a conocer en profundidad la problemática abordada. Posteriormente, se separan del equipo y formulan sus propias soluciones como parte de sus proyectos de tesis. Sin embargo, el riesgo de esto es el posible distanciamiento del proceso real y la respectiva propuesta de soluciones de menor utilidad práctica.

Como epílogo, queremos destacar que la labor de un laboratorio de investigación industrial tiene varios roles. Primero, debe avanzar el estado del arte en tecnologías específicas que son relevantes a una empresa para generar beneficios económicos o de calidad de servicio. Segundo, debe transferir las tecnologías relevantes ya existentes a grupos de ingeniería para que resuelvan los problemas que van encontrado. Finalmente, también debe actuar como consultor estratégico en todos los niveles de la empresa, desde el nivel ejecutivo al nivel de ingeniería. En resumen, estos laboratorios deben verse como el puente entre los centros de investigación básicos como las universidades y las unidades empresariales que desarrollan el negocio de una empresa.

# Literatura citada

*Tesis de posgrado*

[D1] Carlos Gómez-Pantoja, tesis de doctorado, Departamento de Ciencias de la Computación, Universidad de Chile. 2014

[D2] Teresa Bracomente Nole, tesis de doctorado, Departamento de Ciencias de la Computación, Universidad de Chile, 2014.

[D3] Jheser Guzmán, tesis de doctorado, Departamento de Ciencias de la Computación, Universidad de Chile, 2014.

[D4] Mauricio Oyarzún, tesis de doctorado, Departamento de Ingeniería Informática, Universidad de Santiago de Chile, 2014.

[D5] Alonso Inostrosa-Psijas, tesis de doctorado, Departamento de Ingeniería Informática, Universidad de Santiago de Chile, 2014.

[D6] Óscar Rojas, tesis de doctorado, Departamento de Ingeniería Informática, Universidad de Santiago de Chile, 2014.

[D7] Juan Zamora, tesis de doctorado, Departamento de Informática, Universidad Federico Santa María, Chile, 2014.

[D8] Diego Caro, tesis de doctorado, Departamento de Ingeniería Informática y Ciencias de la Computación, Universidad de Concepción, Chile, 2014

[M1] Senén González, tesis de magíster, Departamento de Ciencias de la Computación, Universidad de Chile, 2013.

[M2] Felipe Bravo, tesis de magíster, Departamento de Ciencias de la Computación, Universidad de Chile, 2013.

[M3] Jair Lobos, tesis de magíster, Departamento de Ingeniería Informática, Universidad de Santiago de Chile, 2013.

[M4] Danilo Bustos, tesis de magíster, Departamento de Ingeniería Informática, Universidad de Santiago de Chile, 2014.

*Publicaciones indexadas en Scopus*

[SIGIR 2013] Diego Arroyuelo, Senén González, Mauricio Oyarzún, Víctor Sepúlveda, "Document identifier reassignment and run-length-compressed inverted indexes for improved search performance", in proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013), Dublin, Ireland, 2013.

[SPIRE 2013] Carolina Bonacic and Mauricio Marín, "Simulation Study of Multi-threading in Web Search Engine Processors", in proceedings of the 20th Symposium on String Processing and Information Retrieval (SPIRE 2013), Lecture Notes in Computer Science 8214, Oct. 2013.

[Euro-Par 2013] Oscar Rojas, Verónica Gil-Costa, and Mauricio Marín, "Efficient Parallel Block-Max WAND Algorithm", in proceedings of the 19th International European Conference on Parallel and Distributed Computing (Euro-Par 2013), Aachen, Germany, Lecture Notes in Computer Science 8097, August 2013.

[PADS 2013] MAURICIO MARÍN, VERÓNICA GIL-COSTA, CAROLINA BONACIC and ROBERTO SOLAR, "Approximate Parallel Simulation of Web Search Engines", in proceedings of the ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS 2013), Montreal, Canada, May 2013.

[PDP 2013] VERÓNICA GIL-COSTA, ALONSO INOSTROSA-PSIJAS, MAURICIO MARÍN and ESTEBAN FEUERSTAIN, "Service Deployment Algorithms for Vertical Search Engines", in proceedings of the 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP 2013), Northern Ireland, Feb. 2013.

[SIGIR 2012] DIEGO ARROYUELO, TORSTEN SUEL, SENÉN GONZÁLEZ, MAURICIO OYARZÚN and MAURICIO MARÍN, "To Index or Not to Index: Time-Space Trade-offs in Search Engines with Positional Ranking Functions", in proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), Portland, Oregon, US, Aug. 12-16, 2012.

[ICATPN 2012] VERÓNICA GIL-COSTA, JAIR LOBOS, ALONSO INOSTROSA-PSIJAS and MAURICIO MARÍN, "Capacity Planning for Vertical Search Engines: An approach based on Coloured Petri Nets", in proceedings of the 33rd International Conference on Application and Theory of Petri Nets and Concurrency (ICATPN 2012), Hamburg, Germany, June 25-29, 2012.

[Euro-Par 2011] CARLOS GÓMEZ-PANTOJA, MAURICIO MARÍN, VERÓNICA GIL-COSTA, and CAROLINA BONACIC, "An Evaluation of Fault-Tolerant Query Processing for Web Search Engines", in proceedings of the 17th International European Conference on Parallel and Distributed Computing (Euro-Par 2011), Bordeaux, France, Lecture Notes in Computer Science 6852, pp. 393-404, Springer, Aug., 2011.

[CIKM 2010] CAROLINA BONACIC, CARLOS GARCÍA, MAURICIO MARÍN, MANUEL PRIETO-MATIAS and FRANCISCO TIRADO, "Building Efficient Multi-Threaded Search Nodes", in proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada, Oct 26-30, 2010.

[HPDC 2010] MAURICIO MARÍN, VERÓNICA GIL-COSTA, CARLOS GÓMEZ-PANTOJA, "New Caching Techniques for Web Search Engines", in proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC 2010), Chicago, Illinois, June 20-25, 2010.

[ECIR 2010] MARCELO MENDOZA, MAURICIO MARÍN, FLAVIO FERRAROTI, BÁRBARA POBLETE, "Learning to Distribute Queries onto Web Search Nodes", in proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010), Milton Keynes, UK, March. 2010, Lecture Notes in Computer Science 5993, pp. 281-292, Springer.

[SPIRE 2010] DIEGO ARROYUELO, SENÉN GONZÁLEZ, MAURICIO OYARZÚN, "Compressed Self-indices Supporting Conjunctive Queries on Document Collections", in proceedings of the 17th Symposium on String Processing and Information Retrieval (SPIRE 2010), Lecture Notes in Computer Science, pp. 43-54, 2010.

[Euro-Par 2008] CAROLINA BONACIC, MAURICIO MARÍN, CARLOS GARCÍA, MANUEL PRIETO-MATÍAS and FRANCISCO TIRADO, "Exploiting Hybrid Parallelism in Web Search Engines", In 14th European International Conference on Parallel Processing (Euro-Par 2008), Gran Canaria, Aug. 26-29, Spain, Lecture Notes in Computer Science 5168, pp. 414-423, Springer, 2008.

*Publicaciones en revistas indexadas en ISI y Scopus*

[FINF 2013] Verónica Gil-Costa, Carolina Bonacic, Alonso Inostrosa, Jair Lobos, Mauricio Marín, "Modelling Search Engines Performance using Coloured Petri Nets", accepted in Fundamenta Informaticae, 2013.

[IPM 2012] Diego Arroyuelo, Verónica Gil-Costa, Senén González, Mauricio Marín and Mauricio Oyarzún, "Distributed Search Based on Self-Indexed Compressed Text", Information Processing and Management 48 (5), pp. 819-827, Sept. 2012, (Elsevier).

[PARCO 2010] Mauricio Marín, Verónica Gil-Costa, Carolina Bonacic, Ricardo Baeza-Yates, Isaac D. Scherson, "Sync/Async Parallel Search for the Efficient Design and Construction of Web Search Engines", In Parallel Computing 36(4): 153-168, 2010 (Elsevier).