Estructuras de Datos y Algoritmos Eficientes para Búsquedas Web y Procesamiento de Grandes Datos

Gabriel Tolosa[1], Esteban Ríssola[1,3], Agustín Marrone[1], Francisco Tonin Monzon[1], Federico Radeljak[1], Tomás Delvechio[1], Pablo Lavallén[1] y Esteban Feuerstein[2]
{tolosoft,tdelvechio,plavallen,eamarrone,ftonin,fradeljak}@unlu.edu.ar
efeuerst@dc.uba.ar; esteban.andres.rissola@usi.ch
[1]Departamento de Ciencias Básicas, Universidad Nacional de Luján
[2] Departamento de Computación, FCEyN, Universidad de Buenos Aires
[3] Facoltà di Scienze Informatiche, Università della Svizzera Italiana

Introducción

Hace ya varias décadas que la información digital crece de forma exponencial. Este fenómeno se complementa con el también creciente numero de usuarios de este mundo digital. Se conforma así un ecosistema donde surgen oportunidades para explotar la masividad de los datos.

Este escenario genera nuevas necesidades de almacenamiento, procesamiento y búsqueda que expanden los limites del trabajo en un único equipo y unos pocos algoritmos. Conceptos como "Big Data" y *machine learning* han ido ganando espacio en los últimos años. Muchos problemas además exigen nuevas escalas en periodos de tiempo cada vez mas acotados o directamente en tiempo real.

Constantemente se presentan nuevas oportunidades en este contexto para avances científico/tecnológicos en áreas como algoritmos, estructuras de datos, sistemas distribuidos y procesamiento de datos a gran escala.

Contexto y Formación de RRHH

Proyectos de Investigación "*Algoritmos Eficientes y Minería Web para Recuperación de Información a Gran Escala*" del Depto. De Ciencias Básicas (**UNLu**) y "*Modelos y herramientas algorítmicas avanzadas para redes y datos masivos*" del Departamento de Computación de la Facultad de Ciencias Exactas y Naturales (**UBA**).

Dirección de tres trabajos finales correspondientes a la Lic. en Sistemas de Información de la UNLu. Dos pasantes alumnos y un becario CIN realizan tareas dentro del proyecto principal. Se finalizo una tesis de la maestría "*Exploración de Datos y Descubrimiento de Conocimiento*", DC-FCEyN, UBA y tres de la Licenciatura en Sistemas de Información (UNLu).

Líneas de Investigación

Estructuras de Datos y Algoritmos

Procesamiento de Top-k

consultas · top-k · search engines · performance

Marrone. **Acelerando MaxScore con Múltiples Upper Bounds.** Tesis en curso, 2018.

Caching

streaming · índice invertido · search engines · caching

Feuerstein & Tolosa. **Cost-aware Intersection Caching and Processing Strategies for In-memory Inverted Indexes.** LSDS-IR, 2014.

Tolosa, Feuerstein, Becchetti & Marchetti-Spaccamela.**Performance improvements for search systems using an integrated cache of lists + intersections**. Information Retrieval Journal, 2017

Tonin Monzón. **Políticas de Admisión a Cache Gestionadas Mediante Árboles de Decisión Adaptivos.** Tesis finalizada, 2018.

Estructuras Escalables y Comprimidas

real-time · índice invertido · performance · prunning

Ríssola & Tolosa. **Improving real time search performance using inverted index entries invalidation strategies** JCS&T, 2016

Procesamiento y Análisis de Grandes Datos

commodity · índice invertido · streaming

Delvechio & Tolosa. **Indexación distribuida con restricción de recursos.** AGRANDA, 2017.

Lavallén. **Procesamiento Distribuido de Flujos de Video sobre Plataformas de Big Data.** Tesis finalizada, 2018.

Proponemos definir y evaluar estructuras de datos híbridas que ahorren espacio y mantengan la performance, amortiguando el impacto del crecimiento de la información que se debe manejar.

Planteamos mejorar técnicas de caching específicas en motores de búsqueda, tanto políticas de reemplazo como de admisión. En especial, se incorpora el análisis del flujo de consultas en streaming.

Diseñamos y evaluamos versiones optimizadas de algoritmos de procesamientos de consultas que permitan mejorar las prestaciones de los servicios de búsqueda.

Construimos arquitecturas para aplicaciones de Big Data, orientadas a la indexación masiva distribuida y al procesamiento de flujos de datos en streaming.

Referencias

Busch, Gade, Larson, Lok, Luckenbill & Lin. **Earlybird: Real-time search at twitter**. ICDE 2012.

Cambazoglu & Baeza-Yates. ***Scalability and Efficiency Challenges in Large-Scale Web Search Engines****. SIGIR 2016.*

Ottaviano, Tonellotto & Venturini. ***Optimal space-time tradeoffs for inverted indexes.*** *WSDM 2015.*

Munro, Navarro & Nekrich. **Space-Efficient Construction of Compressed Indexes in Deterministic Linear Time**. SODA 2017.

Nepomnyachiy & Suel. **Efficient index updates for mixed update and query loads**. BigData 2016.