Grandes Datos y Algoritmos Eficientes para Búsquedas de Escala Web

Gabriel Tolosa[1], Santiago Banchero[1], Esteban Ríssola[1], Tomás Delvechio[1], Pablo Lavallén[1] y Esteban Feuerstein[2]
{tolosoft, sbanchero, earissola, tdelvechio, plavallen}@unlu.edu.ar; efeuerst@dc.uba.ar
[1]Departamento de Ciencias Básicas, Universidad Nacional de Luján
[2] Departamento de Computación, FCEyN, Universidad de Buenos Aires

Introducción

La variedad, cantidad y crecimiento de la información disponible en la web actual abre constantemente interrogantes sobre nuevas técnicas eficientes para su almacenamiento y acceso. Los motores de búsqueda son el enfoque más general y popular para acceder a grandes volúmenes de datos por lo cual las cuestiones relacionadas con su eficiencia son temas de activa investigación.

En este escenario, áreas como el procesamiento de datos masivos, que reciben aportes de diversos disciplinas (optimización, *machine learning*, algoritmos aproximados, entre otras) demandan soluciones complejas, involucrando computación y almacenamiento distribuido, algoritmos eficientes y arquitecturas escalables.

Todo el contexto presenta oportunidades únicas 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 cuatro 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. Un Tesista de la maestría "*Exploración de Datos y Descubrimiento de Conocimiento*", DC-FCEyN, UBA.

Líneas de Investigación

Estructuras de Datos
└Distribuidas

índice invertido | sharding | 2D/3D

Feuerstein, Costa, Marín, Tolosa & Baeza-Yates. **3d inverted index with cache sharing for web search engines.**
└Euro-Par, 2012

└Escalables

real-time | índice invertido | pruning

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

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

└Algoritmos Eficientes

caching | optimización | cost-aware | redes sociales

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

Grandes Datos en Aplicaciones Web

logs | machine-learning | optimización | redes sociales

Tolosa & Feuerstein. **Using big data analysis to improve cache performance in search engines**.
└AGRANDA, JAIIO 44, 2015

Plataformas para Grandes Datos

índice invertido | big data | hadoop | spark

Redes sociales y Comunidades

redes sociales | real-time | grafos

Ricci & Tolosa. **Efecto de los trending topics en el volumen de consultas a los motores de búsqueda.**
└CACIC, 2013

Proponemos la construcción de estructuras de datos eficientes orientadas a mejorar el rendimiento de soluciones a problemas de datos masivos.

Planteamos el diseño de arquitecturas para aplicaciones de búsquedas para problemas concretos, donde una solución de propósito general no es la más óptima.

Diseñamos algoritmos para el tratamiento de datos provenientes de redes sociales, interactuando con motores de búsqueda.

Buscamos definir nuevas técnicas de *caching*, incorporando la información del análisis de redes sociales a las políticas de admisión y reemplazo.

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.*

Fang, Macdonald, Ounis & Habel. **Examining the Coherence of the Top Ranked Tweet Topics**. SIGIR 2016.

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.

Wang, Wang, Yu & Zhang. **Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework**. VLDB 2015.