Modelos y Algoritmos para Problemas de Procesamiento en Entornos de Big Data

Gabriel Tolosa[1], Tomás Delvechio[1], Pablo Lavallén[1],
Agustín Marrone[1], Andrés Giordano[1] y Esteban Ríssola[2]
{tolosoft, plavallen, tdelvechio, eamarrone, agiordano}@unlu.edu.ar; esteban.andres.rissola@usi.ch

[1] Departamento de Ciencias Básicas, Universidad Nacional de Luján
[2] Facoltà di Scienze Informatiche, Università della Svizzera Italiana

Introducción

La asimilación de las TICs por parte de las sociedades genera interés sobre algoritmos para almacenar y procesar grandes volúmenes de datos. Además, se observa la maduración de tecnologías de cómputo en la nube, que ofrecen acceso a cómputo de forma rápida y con esfuerzo mínimo. La convergencia de ambos procesos contribuyó al surgimiento del área de Big Data que ofrece esquemas de computo distribuido priorizando la escalabilidad horizontal y la tolerancia a fallas.

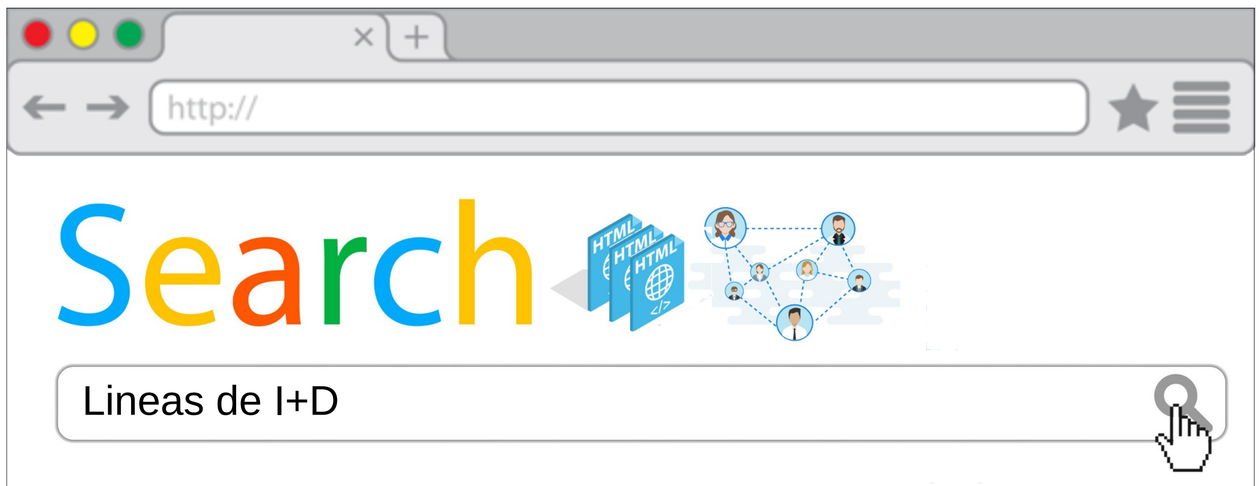
También surge la necesidad de nuevos modelos para resolver los problemas de forma eficiente como, por ejemplo, en motores de búsqueda web, que procesan grandes cantidades de documentos y deben dar respuesta con restricciones de tiempo. Otro ejemplo es el procesamiento de grafos masivos provenientes de diversas fuentes o servicios de streaming, la genómica y la meteorología.

Estos problemas requieren procesamiento distribuido, paralelo y algoritmos altamente eficientes. En la mayoría de los casos, la partición del problema y la distribución de la carga de trabajo son aspectos de las estrategias que requieren ser optimizados.

Contexto y Formación de RRHH

Proyecto de Investigación "*Algoritmos Eficientes y Minería Web para Recuperación de Información a Gran Escala*" del Depto. De Ciencias Básicas (**UNLu**).

Dirección de dos tesis correspondientes a la Lic. en Sistemas de Información de la UNLu. Dos pasantes alumnos, una estancia de Investigación (UNLu) y un becario CIN realizan tareas dentro del proyecto principal.



Algoritmos para *Top-k*

Resolución de consultas en milisegundos requieren técnicas de computo eficiente para calcular los primeros *k* documentos relevantes.

MaxScore | DAAT | Query | Ínverted Index

Búsquedas Selectivas

Partición de colecciones de documentos con alto dinamismo para búsquedas selectivas en cuasi tiempo real.

Searching | Sharding | Caching | Redes Sociales

Cálculo Distribuido en Grafos Evolutivos

Técnicas de procesamiento distribuido sobre grafos masivos evolutivos en entornos escalables.

Grafos Masivos | Escalabilidad | Big Data | Shortest-Path

Estimación de Distancia Geodésica entre Nodos

Cálculos de nodos *landmarks* para reducción de error de estimación de distancias en grafos masivos.

Grafos Masivos | Redes Sociales | Métricas | Landmarks

Diseñamos versiones optimizadas de los algoritmos de procesamiento de *queries* dotando a los algoritmos para *top-k* de información accesoria que les permita recorrer las estructuras de datos eficientemente.

Desarrollamos técnicas para la asignación de flujos de documentos a particiones de un índice que soporte búsquedas, maximizando la performance y considerando parámetros de la arquitectura.

Definimos y evaluamos modelos de cómputo distribuido sobre hardware commodity para problemas de cálculo de métricas en grafos masivos evolutivos.

Diseñamos estrategias de estimación de distancias entre nodos de un grafo masivo para problemas de búsqueda, junto con métodos de corrección de la estimación.

Referencias

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

Ferone, Festa, Napoletano & Pastore. **Shortest paths on dynamic graphs: A survey**. Pesquisa Operacional 2017.

Mallia, Ottaviano, Porciani, Tonellotto & Venturini. **Faster blockmax wand with variable-sized blocks**. SIGIR 2017.

Ríssola & Tolosa. **Improving Real Time Search Performance using Inverted Index Entries Invalidation Strategies**. JCS&T 2016.

Wang, Wu, Luo, Zhang & Dong. **Short-term internet search using makes people rely on search engines when facing unknown issues**. PloS 2017.

Zhang & Pang. **Distance and friendship: A distance-based model for link prediction in social networks**. Spinger 2015.