



Mejoras Algorítmicas para Problemas de Búsquedas en Datos Masivos

Gabriel Tolosa^{1,2}, Tomás Delvechio¹, Pablo Lavallén¹, Andrés Giordano¹, Agustín González¹, Claudia Reinaudi², Santiago Ricci¹, Tomás Juran^{1,2}, Esteban A. Ríssola^{1,3}
{tolosoft, tdevechio, plavallen, agiordano, agonzalez, creinaudi, sricci, tjuran}@unlu.edu.ar
esteban.andres.rissola@usi.ch



¹Departamento de Ciencias Básicas, Universidad Nacional de Luján

²CIDETIC, Universidad Nacional de Luján

³Faculty of Informatics, Università della Svizzera italiana

Introducción

El procesamiento de **datos masivos** se enfrenta a diario a nuevos desafíos debido a su crecimiento a tasas exponenciales, la variedad y diversidad de fuentes disponibles.

Los **algoritmos** que resuelven **problemas de búsquedas** requieren de mejoras tanto conceptuales como ingenieriles que les permitan escalar con el tamaño del problema. La **eficiencia** es un requerimiento fundamental para procesar datos masivos, debido al tamaño, la complejidad y la dinámica de las fuentes actuales de información digital.

Este proyecto presenta el abordaje de problemas relacionados con dos escenarios actuales. Por un lado, el procesamiento de **colecciones masivas de documentos**, para la construcción de motores de búsqueda de escala web. Por otro lado, el procesamiento de **grafos** en cuanto a las métricas de distancias, para aplicar, por ejemplo, a búsquedas de caminos más cortos entre usuarios de redes sociales.

Formación de Recursos Humanos

En el marco de estas líneas de investigación se están dirigiendo **tres tesis de Licenciatura en Sistemas de Información** (UNLu). Además, asociados al proyecto de investigación hay una **Beca Estímulo a las Vocaciones Científicas** (CIN) y **dos pasantías internas en la UNLu**.

Search

Lineas de I+D

Algoritmos de Poda para Top-k

Técnicas eficientes de recorrido de listas invertidas para la resolución de consultas sobre índices masivos.

MaxScore DAAT WAND Upper Bounds

Búsquedas sobre Flujos de Documentos

Procesamiento de documentos que ocurren en tiempo real, junto con estrategias eficientes de *sharding*, cache y selección de recursos.

Sharding Redes Sociales Selective Search Cache

Compresión de Índices

Métodos híbridos de representación en espacio comprimido de los índices invertidos.

Particionado Multicompresión PForDelta Elias Fano

Estimación/Corrección de Distancias en Grafos

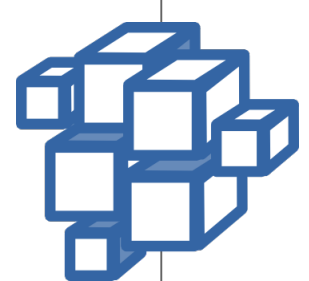
Cálculo estimado (y corrección) de la distancia del camino más corto en grafos masivos para acelerar la resolución de consultas.

Landmarks Grafos Masivos Estimación Función Ajuste

Estrategias de Partición y Procesamiento de Grafos

Distribución de un grafo masivo en diferentes nodos de procesamiento para el cálculo de métricas características.

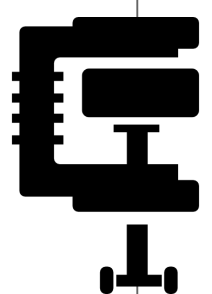
Particionamiento Cálculo Distribuido Grafos Dinámicos Escala del Grafo



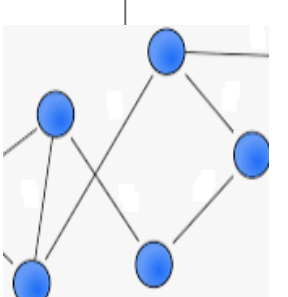
Diseñamos versiones optimizadas de los algoritmos de procesamiento de consultas que recorren las estructuras de datos eficientemente y reducen los tiempos de resolución de consultas.



Desarrollamos técnicas que permiten asignar flujos de documentos a un índice particionado en shards, y la selección de recursos para satisfacer consultas combinando estrategias de caching y de fusión de resultados.



Diseñamos estrategias alternativas (híbridas) de representación de los índices invertidos, considerando el tradeoff entre espacio y tiempo de descompresión.



Proponemos técnicas de estimación y corrección de distancias en grafos masivos para utilizar en búsquedas (por ejemplo, contextuales), junto con estrategias de procesamiento distribuidas y eficientes.

Referencias

Ahmed, Duffield, Willke, Rossi. **On sampling from massive graph streams**. SIGIR, 2018

Giordano, Tolosa. **Improved landmark-based shortest path length estimation in large graphs with distance correction**. IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Tech., 2020

González, Tolosa. **Multicompresión de grandes listas de enteros para sistemas de búsquedas**. JAIIO, 2020

Mallia, Ottaviano, Porciani, Tonello, Venturini. **Faster blockmax wand with variable-sized blocks**. ACM, 2017

D. Lemire and L. Boytsov. **Decoding billions of integers per second through vectorization**. Softw. Pract. Exper., 2015

Mallia, Porciani. **Faster blockmax wand with longer skipping**. Advances in Information Retrieval, 2019.

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