Grandes Datos y Algoritmos Eficientes para Búsquedas de Escala Web

Gabriel Tolosa[1,2], Santiago Banchero[1], Esteban Ríssola[1], Tomás Delvechio[1], Santiago Ricci[1] y Esteban Feuerstein[2]
{tolosoft, sbanchero, earissola, tdelvechio, sricci}@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 evolución del contenido en la web a crecido en número y complejidad exponencial en los últimos años. El enfoque general para acceder a esta información es el uso de motores de búsqueda. Estos motores intentan satisfacer la consulta de los usuarios realizando procesos de recuperación sobre la porción de la web que han recorrido, recopilado y procesado [1].

La arquitectura interna de una máquina de búsqueda de gran escala posee una complejidad desafiante [3], con múltiples oportunidades de optimización. Como operan sobre un sistema dinámico y en constante evolución, las soluciones existentes pueden ya no ser eficientes a futuro y nuevas necesidades aparecen constantemente.

La aparición de nuevas fuentes de información agrega mayores desafíos a los motores de búsqueda (ej. grandes volúmenes de datos de naturaleza diversa donde se requieren respuestas en tiempo real [10]). El área de Grandes Datos se enfoca en dar solución a estos problemas aportando modelos escalables y flexibles [11].

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) y tesis de doctorado del primer autor.

Tesis en curso de la maestría en "Exploración de Datos y Descubrimiento de Conocimiento", DC, FCEyN, Universidad de Buenos Aires.

Dirección de cuatro trabajos finales correspondientes a la Lic. en Sistemas de Información de la Universidad Nacional de Luján. Dos pasantes alumnos realizan tareas con el grupo dentro del proyecto principal. Se espera dirigir al menos dos estudiantes más por año e incorporar al menos otro pasante.

Proponemos optimizar los algoritmos de búsqueda incorporando estrategias provenientes del análisis de grandes datos.

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

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

Procuramos diseñar y evaluar estrategias de indexación distribuida para estructuras de datos avanzadas usando *frameworks* del área de *grandes datos*.

Líneas de Investigación

1) Estructuras de Datos

Distribuidas: Índices invertidos particionados por documentos o términos y enfoques híbridos (índice 2D y 3D [6, 5]), donde se organizan los nodos en un array bidimensional (C col x R filas) aplicando particionado por documentos en cada columna y por términos a nivel de filas (y se replican en el 3D).

Escalables: Mecanismos que soporten búsquedas en tiempo real sobre índices invertidos en memoria [2], considerando la dinámica del vocabulario para evaluar aquella información que resultará relevante (ej. invalidación de entrada a un índice invertido).

2) Algoritmos Eficientes para Búsquedas

Utilizando técnicas de *caching*. En motores de búsqueda, habitualmente se implementan caches para resultados de búsqueda, listas de posting, intersecciones y documentos. Incorporando complementariamente, información de redes sociales [12] para optimizar las técnicas propuestas.

3) Grandes datos en Motores de Búsqueda

Utilizar la información presente en *logs* de consultas de motores de búsqueda para encontrar patrones de comportamiento empleando técnicas de descubrimiento del conocimiento [7, 8], aplicando dicha información para optimizar procesos internos de un buscador.

4) Indexación Distribuida

Incorporar estrategias del ámbito de Grandes Datos a los procesos de construcción y de índices, con interés en la evaluación de técnicas de indexación distribuida sobre estructuras de índices avanzadas [4, 9].

Referencias

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval: The Concepts and Technology behind Search (2 Ed). Addison-Wesley Professional. 2011.
[2] C. Chen, F. Li, B. C. Ooi, and S. Wu, "TI: An Efficient Indexing Mechanism for Real-time Search on Tweets," in Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, 2011.
[3] B. Croft, D. Metzler, and T. Strohman, Search Engines: Information Retrieval in Practice, 1st ed. Addison-Wesley Publishing Company, 2009.
[4] S. Ding and T. Suel, "Faster Top-k Document Retrieval Using Block-max Indexes," in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011.
[5] E. Feuerstein, V. G. Costa, M. Marín, G. Tolosa, and R. A. Baeza-Yates, "3D Inverted Index with Cache Sharing for Web Search Engines," in 18th International Conference, Euro-Par 2012, August 27-31, 2012., 2012.
[6] E. Feuerstein, M. Marín, M. J. Mizrahi, V. G. Costa, and R. A. Baeza-Yates, "Two-Dimensional Distributed Inverted Files," in 16th International Symposium of String Processing and Information Retrieval, {SPIRE'09}, August 25-27, 2009.
[7] Y. Hu, Y. Qian, H. Li, D. Jiang, J. Pei, and Q. Zheng, "Mining query subtopics from search log data," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012.
[8] P. Kaushik, S. Gaur, and M. Singh, "Use of query logs for providing cache support to the search engine," in International Conference on Computing for Sustainable Global Development (INDIACom), 2014.
[9] R. Konow, G. Navarro, C. L. A. Clarke, and A. López-Ortíz, "Faster and Smaller Inverted Indices with Treaps," in Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013.
[10] E. Rissola and G. Tolosa, "Inverted Index Entry Invalidation Strategy for Real Time Search," in Proceedings of the XXI CACIC, 2015.
[11] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis," Nat. Rev. Genet., vol. 11, no. 9, 2010.
[12] S. Ricci and G. Tolosa, "Efecto de los Trending Topics en el Volumen de Consultas a Motores de Búsqueda," in XVII CACIC., 2013.