Hacia la Optimización de un Sistema de Recuperación de Información

Osvaldo Sposito[1], Viviana Ledesma[1], Gastón Procopio[1], Hugo Ryckeboer[1]
[1]Departamento de Ingeniería e Investigaciones Tecnológicas,
Universidad Nacional de La Matanza
{sposito, vledesma, gprocopio, hugor}@unlam.edu.ar

RESUMEN

Con el desarrollo de los repositorios digitales cada vez ha cobrado mayor interés el estudio de los Sistemas de Recuperación de Información. El volumen de la información contenida en dichos repositorios crece de forma exponencial con lo cual búsqueda de los documentos que respondan a la necesidad de los usuarios se torna una tarea difícil. En este contexto este grupo de investigación ha estado trabajando por más de ocho años en la construcción de sus propios motores de búsqueda y recuperación orientados a corpus estáticos. Una vez construidos dichos motores las líneas de investigación se han orientado a distintos enfoques que pretenden acelerar los mismos tanto en la búsqueda como en los preprocesos. En particular, en esta etapa se tiene por objetivo la investigación, desarrollo e implementación de algoritmos paralelos, principalmente para resolver el proceso de la de Descomposición en Valores Singulares en arquitecturas basadas en Unidades de Procesamiento Gráfico y su comparación con clústeres de multicores, así como el empleo de soluciones híbridas que combinen ambos enfoques.

Palabras clave: Sistemas de Recuperación de Información, Descomposición en Valores Singulares, Bidiagonalización, Indexación Semántica Latente.

CONTEXTO

La línea de investigación que se presenta se encuentra inmersa en el proyecto de investigación C225 "*Resolución Eficiente de la Descomposición en Valores Singulares en una Arquitectura Híbrida y su Posterior Inserción en un Sistema de Recuperación de Información*" llevado a cabo en el marco del programa PROINCE de la Universidad Nacional de La Matanza (UNLaM). El mismo se desarrolla en el Polo Tecnológico dependiente del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLaM.

Este trabajo continúa con la línea de investigación de los proyectos PROINCE C151, C177 y C205 cuya temática se orientó: primero al estudio del tema y posteriormente a la realización de un prototipo de un sistema de recuperación de la información, la optimización de la recuperación de documentos usando como técnica base el LSI (Lematización Semántica Latente), y el uso de Minería de Datos para acelerar la recuperación de documentos.

1. INTRODUCCIÓN

A partir de la década del 90 los avances tecnológicos de la informática dieron lugar a un incremento exponencial en la generación y almacenamiento de información que continúa hasta la actualidad. La enorme cantidad de información almacenada hace que su búsqueda y recuperación sea cada vez más dificultosa, lo cual impulsó el estudio de la Recuperación de la Información (RI) como disciplina. Existen dos tendencias principales en el desarrollo de Sistemas de RI (SRI) según el contexto y el ámbito de la fuente documental [1]: La RI vertical que se enfoca en la indexación de fuentes documentales específicas, por ejemplo, una biblioteca de Ciencias Jurídicas. La RI horizontal que se enfoca en fuentes documentales generales, por ejemplo, la Web. Esta línea de investigación se relaciona con el primer grupo dado que se trata la implementación y optimización de un SRI de fuentes documentales específicas y la evaluación de su rendimiento.

En la literatura existen diversas propuestas sobre la organización interna de un SRI [2, 3]. La Figura 1 muestra una representación simplificada de un SRI, tal como se visualiza, los procesos más importantes que intervienen

en dicho sistema son los siguientes:

Indexación – los documentos que alimentan el sistema se representan como objetos indexados.

Búsqueda – se analiza la consulta del usuario y se compara con los objetos indexados, de tal modo se pueden obtener los objetos recuperados que se le presentarán al usuario.

Ranking – se determina la relevancia de cada documento recuperado para dar solución a la consulta que haya ingresado el usuario, finalmente los documentos se ordenan en base a los valores obtenidos en este proceso.

Este trabajo está enmarcado dentro del proceso de indexación.

Figura 1. Sistema de RI.

Se han ideado diferentes modelos basados en distintos paradigmas para representar tanto documentos como consultas en SRI y comparar la similitud de esas representaciones [4]. Entre estos se encuentran el modelo booleano, el modelo vectorial y el modelo probabilístico, denominados clásicos. Esta línea de investigación se enmarca en una variante del método de RI vectorial, la Indexación Semántica Latente (ISL) [5].

La ISL es un método para la búsqueda de información en documentos a través de la indexación de términos [5]. Con dicho método se pretende resolver perturbaciones en la RI causados por problemas de sinonimia y polisemia o equivocidad del habla corriente. Para ejemplificar, si se desea buscar la palabra "estación", la cual tiene múltiples significados (polisemia) una búsqueda literal de la palabra produciría muchos resultados posibles (estación de tren, estación del año, etc.). Si lo que se desea buscar es "estación del año", podrían interesar resultados de palabras distintas, pero con un significado igual o similar, por ejemplo "temporada", "época" y así por el estilo (sinonimia). La ISL permite la búsqueda por conceptos o definiciones en contraposición a la búsqueda literal.

La aplicación de la ISL implica la utilización de algoritmos matemáticos especializados, que como resultado simulan el análisis que realizaría una persona. Una técnica ampliamente utilizada a tal fin es la Descomposición en Valores Singulares (DVS), luego la recuperación se realiza utilizando como punto de partida los valores singulares y vectores obtenidos a partir de la aplicación de dicha técnica [6].

La DVS [5,7] consiste en descomponer una matriz en varias matrices que exhiben las propiedades más importantes de la matriz original. Así, una matriz A de tamaño t × d descompuesta con DVS (ver Figura 2) produce tres matrices de la forma:

$$A = T_0 S_0 D_0$$

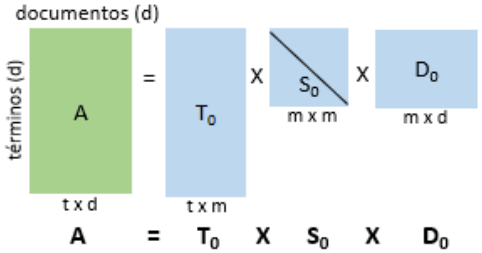

Figura 2. Reducción de dimensiones en DVS. Fuente: [5]

T_0 y D_0 tienen columnas ortonormales (ortogonales y de tamaño uno) y son las matrices izquierda y derecha, respectivamente, de vectores singulares y S_0 es una matriz diagonal compuesta de los valores singulares de A.

Disponer de modelos de orden reducido tiene como ventaja el simplificar la comprensión del sistema, reducir el coste computacional en los problemas de simulación, lo cual a su vez implica menor esfuerzo computacional en el diseño de controladores numéricamente más eficientes y se obtienen leyes de control más simples [8]. De ahí la necesidad de buscar modelos

matemáticos simplificados que aproximen al máximo el comportamiento del sistema original. El modelo resultante, que poseerá un número menor de estados que el sistema original, se denomina modelo reducido o modelo de orden reducido y al procedimiento utilizado para conseguirlo se lo conoce como reducción de modelo.

Existen dos tipos principales de algoritmos que se aplican al cálculo computacional de la DVS de una matriz real, el método unilateral de Jacobi y aquellos algoritmos que se basan en la bidiagonalización [6]. El número de operaciones para los distintos algoritmos se encuentra en el orden de On^3, las diversas propuestas y mejoras que han surgido buscan disminuir operaciones costosas en tiempo.

Mediante este proyecto se pretende optimizar la resolución de la DVS, en especial mediante implementar algoritmos para resolver la primera fase de este proceso, a través de la bidiagonalización. Los métodos más tradicionales de bidiagonalización utilizan las transformaciones de Householder por la izquierda y por la derecha de la matriz [9, 10]. Como desventaja, cuando las matrices son de grandes dimensiones requieren tiempos de computación elevados y además repercuten negativamente en los costos de comunicación de una implementación paralela del algoritmo en sistemas de memoria distribuida [11,12]. Así es que se han realizado diversos trabajos, entre estos se encuentran la propuesta de Ralha [13], mejorada más adelante por Barlow [14], orientada a conseguir un método más sencillo aplicando las transformaciones de Householder solamente por el lado derecho de la matriz. Posteriormente, Da Silva Sanches de Campos [12] presenta una mejora al método de Barlow con el objetivo de reducir el número de comunicaciones necesarias para una implementación paralela destinada a sistemas de memoria distribuida.

Para este proyecto se pondrá especial interés en los algoritmos alternativos de bidiagonalización propuestos por Ralha y Barlow [13,14], dado que están pensados para soportar el paralelismo. La decisión se sustenta en que el resultado final de esta etapa de la investigación se orienta hacia algoritmos que puedan ser implementados en plataformas paralelas y, en particular aprovechando la capacidad de las unidades de procesamiento gráfico (GPU, por sus siglas en inglés).

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

A lo largo de varios proyectos de investigación, se ha profundizado la investigación de la RI. La meta inicial del equipo fue construir íntegramente un prototipo de un sistema de organización de documentos para su posterior recuperación mediante un buscador. Habiendo concluido esa primera etapa se buscó optimizar el trabajo de RI lo cual implica dar atención a distintas líneas de investigación:

a) Mejorar la lematización disponible, en particular para este caso del idioma español;
b) Subdividir el corpus en forma inteligente de modo tal que sin gran pérdida de exhaustividad se pueda resolver la consulta examinando una o más partes de la subdivisión, excluyendo a muchas de ellas; y
c) Acelerar la velocidad de cómputo.

En particular, esta etapa de la investigación se relaciona a la línea de investigación mencionada en el ítem c. El objetivo principal de este trabajo apunta a diseñar, implementar y evaluar algoritmos secuenciales y paralelos para resolver eficientemente cada uno de los algoritmos utilizados en la DVS. En función a tal objetivo se han trazado los siguientes objetivos específicos:

a) Estudiar el problema matemático de la DVS y las variantes algorítmicas existentes para una mejor implantación en GPU.
b) Estudiar a bajo nivel las arquitecturas de los equipos CPU y GPU, sobre los cuales se realizarán los desarrollos, así como de las herramientas de software necesarias para su máximo aprovechamiento.
c) Estudiar las principales librerías disponibles que resuelven problemas relacionados con el álgebra lineal, especialmente aquellas que lo realicen en

arquitectura de cálculo paralelo, por ejemplo, CUDA, CUBLAS, OPENCL, etc.

d) Desdoblamiento de los algoritmos para resolverlos sobre una configuración híbrida.
e) Realizar un estudio comparativo en cuanto al rendimiento al bidiagonalizar matrices de variados tamaños cuando se utilizan distintas implementaciones variando la arquitectura. Determinando que algoritmo e implementación resulta más eficiente.
f) Calcular la DVS utilizando el algoritmo identificado en el punto anterior, y finalmente implementarlo en el SRI desarrollado por el equipo.

3. RESULTADOS OBTENIDOS/ESPERADOS

A continuación, se enumeran los resultados ya alcanzados:

i. Se construyeron dos prototipos de SRI, uno basado en lo que ahora se pueden llamar métodos clásicos y otro según el método LSI. Estos se diseñaron de forma modular de tal manera que permitan mejoras locales con su consiguiente experimentación. Los mismos se almacenaron con código abierto a fin de facilitar a nuevos grupos interesados en esta tecnología para su iniciación en el tema y la base sobre la cual efectuar experimentos propios.
ii. Se extendió la selección de documentos a corpus voluminosos a través de la utilización de DVS y clustering. La DVS proporciona una salida donde las diferencias de las distancias entre los documentos son muy cercanas, mientras que clustering utiliza esas distancias para poder agrupar los documentos, parte de este trabajo ha sido publicado en [15].
iii. Se estudió como línea alternativa fraccionar el corpus. Esto requiere dos algoritmos preparatorios, uno que particione el corpus utilizando una noción de vecindad o similitud y el entrenamiento de un algoritmo de clasificación que direccione la consulta hacia la parte más promisoria. Para ambos servicios se aplicaron técnicas de minería de datos y de la selección de la parte usando redes neuronales. Los resultados se han reflejado parcialmente en [16].
iv. Adicionalmente se desarrolló un generador de corpus, para ser utilizado en esta línea de investigación y pretendiendo, además, con esto, colaborar con la propagación de la Lingüística de Corpus como metodología para investigaciones en RI, el trabajo realizado se describe en [17].

En las próximas etapas de investigación, se profundizará el análisis de la posibilidad de optimizar la DVS, dando atención en particular a la primera fase, la bidiagonalización. Para ello se realizará un estudio comparativo en cuanto al rendimiento al bidiagonalizar matrices de variados tamaños cuando se utilizan distintos algoritmos y distintas implementaciones variando la arquitectura. De esta manera se intentará determinar qué algoritmo e implementación resulta más eficiente. La investigación se centrará en el desarrollo e implementación de algoritmos paralelos, principalmente el de DVS, en arquitecturas basadas en GPU y su comparación con clústeres de multicores, así como el empleo combinado de GPU y multicores. El alto grado de paralelismo de las GPU que sin lugar a duda disminuye el tiempo de cálculo, sufre una mengua en la misma a causa de la lentitud de sus comunicaciones transversales. De allí surge el interés del equipo en explorar si soluciones híbridas pudieran aportar una aceleración en los cómputos, delegando en cada parte, CPU o GPU, aquellas tareas en las cuales mejor se desempeñan.

Por otro lado, se llevará a cabo la puesta a prueba de dicha optimización en el SRI desarrollado por el equipo con el fin de comprobar el nivel de impacto alcanzado en la productividad del proceso. No es objetivo de este proyecto obtener una resolución en forma abstracta y genérica como para enriquecer las bibliotecas del cálculo matricial, sino resolverlo para ciertas arquitecturas concretas disponibles en la sede del proyecto. Esto marcaría el camino para que otros, con más recursos económicos y tamaño del equipo

humano, puedan extenderlo y parametrizarlo para que funcione en otras configuraciones.

4. FORMACIÓN DE RECURSOS HUMANOS

En el proyecto participan seis investigadores, uno de ellos en formación y dos son alumnos de grado. La línea de investigación presentada aquí es parte directa de la tesis "*Estudio comparativo de DBSCAN, KMEANS con redes neuronales en un Sistema de Recuperación de Información*", correspondiente a la Maestría en Informática que está desarrollando el Ing. Casuscelli Marcos en UNLaM.

Durante el último año la Ing. Viviana Ledesma presentó su tesis de maestría y su posterior defensa, la cual desarrolló en la UNLaM, y por su parte, el alumno Gastón Procopio, finalizó su carrera de Ingeniería en Informática con la aprobación del proyecto final de carrera.

Parte de los resultados de esta investigación son divulgados en la cátedra de Diseño de Sistemas que se dicta para la carrera de Ingeniería en Informática de la UNLaM. Se espera además que esta investigación contribuya a la formación de recursos humanos en RI y que el sistema desarrollado pueda servir de base para una transferencia de tecnología a las PYMEs de la región.

5. BIBLIOGRAFÍA

[1] Cleverdon, C.W. "Progress in documentation. Evaluation of information retrieval systems", Journal of Documentation, 26, 55-67, 1970.

[2] Kowalski, G. "Information Retrieval Systems: Theory and Implementation", 1st ed. Norwell, MA, USA: Kluwer Academic Publishers, 1997.

[3] Kuna, H., Rey, M., Martini, E., Solonezen, L. & Podkowa, L. "Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación", Revista Latinoamericana de Ingeniería de Software, 2014. 2(2): 107-114.

[4] Tolosa G. & Bordignon, F. "Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos". Universidad Nacional de Luján, Argentina, 2008. Recuperado el 01/08/2019 de: http://eprints.rclis.org/12243/1/Introduccion-RI-v9f.pdf

[5] Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. "Indexing by latent semantic analysis". Journal of the American Society for Information Science (SIAM),1990. 41(6):391–407.

[6] Lahabar, S. & Narayanan, P. "Singular Value Decomposition on GPU using CUDA". IEEE International Symposium on Parallel & Distributed Processing, 2009. 1-10.

[7] Berry, M., Dumais, S. & O'Brien, G. "Using Linear Algebra For Intelligent Information Retrieval". Society for Industrial and Applied Mathematics, Review 37(4): 573-595. Philadelphia, USA, 1995.

[8] L. Fortuna, G. Nunnari & A. Gallo. "Model order reduction techniques with applications in electrical engineering". Springer-Verlag, 1992.

[9] J. Demmel, M. Gu, S. Eisenstat, et al. "Computing the Singular Value Descomposition with Hight Relative Accuracy". Linear Algebra and its Application, 299, 21-80, 1999.

[10] T. Chan. "An Improved Algorithm for Computing the Singular Value Decomposition". ACM Transactions on Mathematical Software, 8(1): 72-83, 1982.

[11] Sangwine, S. & Le Bihan, N. "Quaternion Singular Value Decomposition based on Bidiagonalization to a Real Matrix using Quaternion Householder Transformations" Applied Mathematics and Computation, ELSEVIER, 182(1): 727-738, 2006.

[12] Da Silva Sanches de Campos, C. "Algoritmos de Altas Prestaciones para el Cálculo de la Descomposición en Valores Singulares y su Aplicación a la Reducción de Modelos de Sistemas Lineales de Control". Tesis Doctoral. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, España, 2014.

[13] Ralha, R. "One-sided reduction to bidiagonal form". Linear Algebra and Its Applications, ELSEVIER, 358(1-3): 219-238, 2003.

[14] Barlow, J., Bosner, N., Drmač, Z. "A new stable bidiagonal reduction algorithm". Linear Algebra and Its Applications, ELSEVIER, 397: 35-84, 2005.

[15] Spositto, O., Procopio, G., Quintana, F.& Ryckeboer H. "Una paralelización del método de Householder", CACIC 2016, pp. 1291-1300. Universidad Nacional de San Luis San Luis, 2016.

[16] Spositto, O., Casuscelli, M., Bossero, J., Matteo, L., Ryckeboer, H. "Aceleración en la Recuperación de Información utilizando Algoritmos de Minería de Datos de R". CACIC 2018, pp.491-500.Universidad Nacional del Centro, Tandil, 2018.

[17] Spositto, O., Procopio, G. Bossero, J. "Método para la Construcción de un Corpus Periodístico mediante Expresiones Regulares". CONAIISI 2018, pp. 491-500. Universidad CAECE, Mar del Plata, 2018.