# Recuperación de la Información

Ryckeboer Hugo, Spositto Osvaldo, Bossero, Julio César, Barone Miriam
***Departamento de Ingeniería e Investigación Tecnológica***

***Universidad Nacional de La Matanza***

***hryckeboer@unlam.edu.ar spositto@unlam.edu.ar mbarone@unlam.edu.ar jbossero@unlam.edu.ar***

## Resumen

Las Técnicas de Recuperación de Información que responden a inquietudes puntuales, se han hecho populares gracias a los buscadores ofrecidos gratuitamente a quienes recurren al Internet. Y no se hace referencia únicamente a grandes repositorios, sino también a medianos y pequeños, con miles de documentos, donde también es un inconveniente la localización del documento o los documentos que respondan a la inquietud del Usuario.

El grupo posee sus propios motores orientados a corpus estáticos. De las diversas concepciones existentes ha prestado especial atención a la indexación semántica latente conocidas por sus siglas LSI.

Una vez construidos los motores las líneas de investigación se han orientado a enfoques que permitan acelerar los mismos tanto en la búsqueda como en los preprocesos.

Uno de ellos es el uso del procesamiento paralelo, tanto en clúster de máquinas como en el uso de placas de video.

La técnica LSI es particularmente dependiente en su preproceso de un eficiente cálculo de autovalores y autovectores de matrices de gran tamaño, lo que hace incluir en nuestra temática el cálculo numérico.

Se investiga si es factible acelerar la selección de los documentos que responden a un requerimiento por medio de un particionado del corpus basándose en criterios de similitud propia de minería de datos y técnicas de selección de la parte usando redes neuronales.

En este sentido se exponen distintas líneas de trabajo a seguir, teniendo como objetivo diseñar, implementar y probar modificaciones en los procesos de filtrado y ordenamiento de documentos, en un Sistema de Recuperación de Información (SRI), aplicando algoritmos de clustering tradicionales.

**Palabras claves**: Examinar, Indexar, Buscar, SRI, LSI, Minería de Datos, Agrupamiento, Algoritmos de Clasificación (Browse, Indexing, Search)

## Contexto

Esta propuesta de trabajo se lleva cabo dentro de dos proyectos de investigación "*Optimización de la Recuperación de Documentos, usando como técnica base el LSI (Lematización Semántica Latente)",* y el proyecto "*Uso de Minería de Datos para acelerar la recuperación de documento***"**

Los cuales son desarrollados por el grupo de investigación del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza, en el marco de investigación del programa PROINCE.

## 1.Introducción

Los contenidos de los documentos digitales hoy en día, son una materia prima muy valiosa, tanto para empresas u organizaciones como para simples usuarios. Es por esto que en la Sociedad de la Información se destinan gran cantidad de recursos en almacenar grandes volúmenes de documentos, organizarlos para luego recuperar los adecuados, debido entre otras cosas, a la explosión en el número de fuentes de información disponibles en Internet que sobrepasa a toda información manual. También hay conjuntos de documentos

cerrados, por ejemplo, los legislativos, las obras de filósofos famoso, sobre los cuales se desean efectuar consultas puntuales.

Este es uno de los motivos por lo que, desde hace años, se dispone de los denominados Sistemas de Recuperación de Información *(SRI o IRS en inglés Information Retrieval Systems),* que permiten almacenar, buscar y mantener documentos, extendiendo esto a textos, imágenes, vídeos, audios y otros objetos multimedia, los cuales, utilizan técnicas de búsqueda relativas a su contenido, que son específicas para cada tipo de información.

Para evitar una dispersión en un grupo humano pequeño, los proyectos se han centrado sobre información textual en español

Las manifiestas similitudes existentes entre la recuperación de información y otras áreas vinculadas al procesamiento de la información, se repiten en el campo de los sistemas encargados de llevar a cabo esta tarea. Para Salton en [Sal86] "*...la recuperación de información se entiende mejor cuando uno recuerda que la información procesada son documentos…*", con el fin de diferenciar a los sistemas encargados de su gestión de otro tipo de sistemas, como los gestores de bases de datos relacionales. Salton piensa que "…cualquier SRI puede ser descrito como un conjunto de ítems de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo (SIMILAR) que determine qué ítem satisfacen las necesidades de información expresadas por el usuario en la petición"

Para poder hacer aportes originales en esta temática fue fundamental tener motores completos en estado operativo.

El grupo ha privilegiado la metodología LSI porque además de su conocida habilidad para resolver ambigüedad, equivocidad y sinonimia, provee vectores descriptivos de los documentos y de consultas de menor dimensión lo que beneficia a la minería de datos.

Los motores construidos son lo suficientemente abiertos y flexibles para ser utilizado en docencia y en puestos de investigación.

Cada día se utilizan técnicas más avanzadas de análisis del contenido de los documentos con vistas a mejorar los tiempos de acceso a los documentos y la efectividad del resultado.

Por lo tanto, el problema al que se enfrenta la RI se puede definir como: "Dado un conjunto de documentos, ordenar los documentos de mayor a menor según la relevancia para una determinada necesidad ya expresada como consulta, las limitaciones perceptivas del usuario aconsejan entregar los elementos que encabezan la lista".

Aunque una buena parte del pre proceso de organización y proceso de consultas recurren a técnicas básicas de la computación se pueden señalar algunas áreas que dan lugar a mejoras y optimizaciones las que influirán en la calidad de las prestaciones:

Análisis Textual: es una práctica ya establecida que en lugar de recurrir a la presencia o no de palabras identificadas en la consulta dentro de los documentos, es conveniente reducirlas a lexemas ignorando deflexiones propias del desarrollo del texto, pero salvando un concepto común que la palabra encierra en sus distintas formas morfológicas.

Esta actividad en sus detalles es dependiente del idioma y del campo de aplicación.

Proceso del Corpus: Los coeficientes de los vectores que describen la temática de los documentos requieren un ajuste a la luz de la totalidad de los documentos disponibles. Esto en el caso del LSI requiere hallar autovalores y autovectores de elevada dimensión, problema no trivial por los errores de redondeo del cálculo con reales. Continuamente aparecen propuestas que intentan acelerar o mejorar tales cálculos.

Resolución de las Consultas: Una forma de saber el valor de un documento frente a una consulta es enfrentar los vectores

representativos de ambos. El tiempo que insume la construcción de la lista ordenada crece con el tamaño del corpus.

Dos líneas de trabajo surgen frente a esta situación:

- Distribuir los documentos en varios procesadores
- Dividir el corpus en conjuntos más pequeños conteniendo documentos "similares" y comenzar la recuperación por la parte más promisoria.

Tanto en el dividir como elegir esta parte lo estamos encarando con técnicas de minería de datos.

Los cálculos de autovalores involucran a todos los coeficientes de la matriz lo que crea un dilema y necesidad de hallar un compromiso entre distribuir el cálculo aumentando las comunicaciones entre procesadores, o concentrarlos para no tener tales recargos de tiempos.

## 2. Líneas de Investigación y desarrollo

Mejorar el trabajo con RI implica, por lo que se ha dicho, considerar diferentes líneas de investigación:

a) Acelerar la velocidad de cómputo, recurriendo a un procesamiento en paralelo
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.
c) Mejorar la lematización disponible del idioma español, dado que la misma no da resultados satisfactorios.

Respectivamente, se señalan las observaciones que serán las inquietudes centrales en estas líneas de investigación:

Los sistemas que operan en gran escala deben recurrir necesariamente al uso en paralelo de varios procesadores. Se estudia la forma de paralelizar algunos algoritmos para acelerar adecuadamente los cómputos.

Dada la posibilidad de extender la selección de documentos a corpus muy voluminosos, existen diversas ideas de subdividir el corpus en grupos aplicando técnicas de agrupamiento. Para que la subdivisión sea efectiva a los fines propuestos se debe agrupar documentos de iguales características y separar los que manifiestamente difieren. En este proyecto se pretende dominar e incorporar estas tecnologías a nuestro prototipo, con la intención de evaluar si la mejora en velocidad compensa una eventual pérdida de exhaustividad a las mismas. Las tecnologías que resuelve esto se conoce como "clustering" [Her04], de modo tal de disminuir el espacio de búsqueda cuando se procesa una consulta. El proyecto pretende subdividir un corpus en varios grupos o clúster, para realizar la búsqueda de documentos pertinentes a una consulta dada. Por lo que la hipótesis principal es que la utilización de técnicas de clustering y de aprendizaje supervisado, en un SRI, acelerará la obtención de documentos pertinentes. Los beneficios alcanzarían a los usuarios directos y a la reducción de recursos computacionales. La subdivisión encierra un riesgo de respuestas no exhaustivas que afecten a la calidad del servicio. Esto se encuentra en etapa de evaluación.

Explorar a fondo las distintas alternativas que se presentan en las diversas etapas de esta solución constituyen el centro de uno de los proyectos en curso.

Las mismas se pueden resumir en:

- ✓ Elegir técnica de clustering o incluso diseñar una nueva.
- ✓ Diseñar el algoritmo que oriente una búsqueda en particular hacia uno o varios de los subconjuntos. Las redes neuronales son una de las técnicas que se vislumbran como promisorias para esta tarea.
- ✓ Proponer técnicas de evaluación en cuanto a cobertura lograda versus tiempo de respuesta.

Pensando en el preproceso está la lematizacion, esta se puede desdoblar o sea repartiendo a distinto procesador, distinto documento o inclusive repartiendo párrafos. La suma de presencia de lexemas en distintos documentos también es paralelizable. La transposición de la tabla documento-termino también es realizable en paralelo diseñando cuidadosamente el algoritmo.

Objetivos Secundarios

De lo enunciado ut-supra se desprenden los siguientes objetivos secundarios:

a) Obtención de un Corpus en español, más voluminoso que el utilizado hasta ahora, para la aplicación de los algoritmos.

b) Evaluar y proponer algoritmos de agrupamiento para el conjunto de documentos disponibles.

c) Ajustar los modelos predictivos de tipo supervisados, para resolver el problema de clasificación.

d) Evaluar el alcance de los resultados a través de las métricas propuestas.

e) Aplicar los modelos desarrollados en nuevos Corpus

## 3. Resultados Obtenidos/Esperados

Las técnicas de minerías de datos en uso en esta línea de investigación fueron profundizadas en proyectos previos lo que dio lugar a las siguientes publicaciones:

1- "Predicción del riesgo de abandono universitario utilizando métodos supervisados" En colaboración con Edwards, Diego y Pérez, Silvia (UNLaM). Trabajo presentado en el Workshop de la V Jornadas Nacionales y I Latinoamericanas de Ingreso y Permanencia en Carreras Científico – Tecnológicas. Facultad Regional Bahía Blanca. Universidad Tecnológica Nacional. Bahía Blanca. Mayo de 2016. IPECyT 2016

2- "Modelos de minería de datos para el diagnóstico de enfermedad de Parkinson mediante el análisis de voz". En colaboración con el Ing. Osvaldo Spositto, Ing. Gabriel Blanco, Mg. Mónica Giuliano y el Ing. Luis Fernández (UNLaM). Trabajo presentado en el Workshop del V Congreso Nacional de Ingeniería en Informática/Sistemas de Información. Publicación en línea - ISSN. CONAIISI 2017. Santa Fe. Argentina.

3- "Comparación de Algoritmos de Aprendizaje Supervisado para la obtención de perfiles de alumnos desertores". En colaboración con el Ing. Osvaldo Spositto (UNLaM). Trabajo presentado en el Workshop del IV Congreso Nacional de Ingeniería en Informática/Sistemas de Información. Publicación en línea - ISSN 2347-0372. CONAIISI 2016. Salta. Argentina

4- "Una paralización del método de Householder" En colaboración con el Ing. Osvaldo Spositto, Ing. Hugo Ryckeboer. (UNLaM). Trabajo presentado en el XXII Congreso Argentino de Ciencias de la Computación- CACIC 2016- Universidad Nacional de San Luis San Luis.

En el marco de la línea de investigación para acelerar la velocidad de cómputo, recurriendo a un procesamiento en paralelo a lo que respecta a la lematización del idioma español disponible que no daba resultados satisfactorios, se puede concluir que con el uso de los hilos se realiza un procesamiento más rápido que con la forma secuencial.

Finalmente, respecto a los sistemas que operan en gran escala, estos deben recurrir necesariamente al uso en paralelo de varios procesadores. La utilización de los GPU, es el acelerador dominante para realizar procesamiento de cálculos en paralelo, principalmente en matrices, ello tuvo un resultado positivo: los tiempos bajan drásticamente comparando un proceso secuencial en una Pc típica de escritorio contra el procesamiento sobre cualquiera de las GPU.

En cuanto a la comparación entre las diferentes GPU, se observa que los mejores tiempos se obtuvieron para la GPU R9 390X. A medida que haya más documentos, la

diferencia de tiempos entre cada una de ellas se va notando considerablemente.

Los documentos de un Corpus son transformados en vectores descriptivos. Una consulta de usuario es también convertida en otro vector descriptivo. Para obtener un documento que satisfaga la necesidad de información del usuario, el vector de la consulta se debe enfrentar con todo el corpus, en búsqueda de similitudes. Este proceso genera un índice de relevancia, que será la salida que recibe el usuario, en forma de lista ponderada. Dividir el Corpus, de modo tal de poder desechar uno o más grupos, debería acelerar la búsqueda.

## 4. Formación de Recursos Humanos

*Resultados en cuanto a la producción de conocimiento:*

Disponer de un buen lematizador del español es una contribución al estado del conocimiento en Recuperación de Información en lengua española.

*Resultados en cuanto a la difusión de resultados:*

Del mismo modo que en el proyecto precedente se puso un Motor de Búsqueda a disposición de toda la comunidad académica, se hará lo mismo con el lematizador español.

Los resultados en materia de lematización y del uso de "clusters" de computadoras, serán expuestos en Congresos/Revistas.

Los lematizadores no exponen normalmente la metodología con la cual los diseñaron, con lo cual, la exposición de estos detalles, puede ser valiosa para profesionales de otras lenguas.

El armado de "clusters" obliga a resolver problemas prácticos de conexión y administración, lo que puede acortar el camino a otros investigadores que se inicien en el tema.

Los profesionales de Informática disponen de baja formación lingüística, de modo tal que su participación en este proyecto les abrió nuevos campos de actividades: la lingüística computacional, el manejo de semántica, traducción automática. Todas estas actividades requieren un adecuado manejo morfológico del lenguaje.

Respecto a la minería de Datos se utilizaron parte de los conocimientos en 2 trabajos de tesis desarrolladas en sendas maestrías en informática:

Tesis aprobada: El Soporte Informático y su Aporte para la Inclusión Universitaria

Tesis en curso: Estudio Comparativo de Técnicas de Minería de Datos para la predicción de deserción universitaria

*Resultados en cuanto a transferencia hacia las actividades de docencia y extensión:*

Los sistemas de IR son eficaces en la medida que diseñen buenas estructuras de datos. Estas son estudiadas en materias intermedias de la Ingeniería de Sistemas, poder ilustrar su uso práctico, beneficiará a los estudiantes.

Del mismo modo, el uso de clusters ilustra los tópicos más avanzados de la Arquitectura de Computadoras, que también integran el Plan de Estudios.

## 5. Bibliografía

[Her04] "Introducción a la minería de datos". José Hernández Orallo y otros. Editorial: Pearson. Edición: I. Año 2004

[Ven03] "*Análisis Semántico Latente: una panorámica de su desarrollo*". René Venegas

Rev. signos [online]. 2003, vol.36, n.53, pp.121-138. ISSN 0718-0934. Pontificia Universidad Católica de Valparaíso. Chile

Disponible en: http://dx.doi.org/10.4067/S0718-09342003005300008.

[Sal86]"Introduction to Modern Information Retrieval". Gerard Salton, Michael J. Michael J. McGill. Ed. McGraw-Hill, Inc. New York, NY, USA. ISBN: 0070544840. 1986

[Sal68]"Automatic Information Organization and Retrieval". Salton, G. McGraw−Hill, N.Y. 1968.

[1] PCI-Express http://pcisig.com/specifications/pciexpress/resources

https://nvlabs.github.io/moderngpu/performance.html

[2] *"Clustering de documentos con restricciones de tamaño"*. Diego Fernando Vallejo Huanga. Trabajo Fin de Máster. Universitario en Gestión de la Información. Escola T. S. d'Enginyeria Informàtica Universitat Politècnica de València. 2015. Disponible: https://riunet.upv.es/bitstream/handle/10251/69089/VALLEJO-ClusteringdeDocumentosconRestriccionesdeTamaño.pdf?sequence=23

[3] "Clusterdoc, un sistema de recuperación y recomendación de documentos basado en algoritmos de agrupamiento". Marylin Giugni O.;Luis León G.;Joaquín Fernández.

Telematique, volumen 9 - número 2 - año 2010.

Disponible en:http://publicaciones.urbe.edu/index.php/telematique/article/view/913/ pdf

[4] Lindholm, Erik and Nickolls, John and Oberman, Stuart and Montrym, John,NVIDIA Tesla: A unified graphics *and computing architecture*,IEEE micro, 2008.