Adecuación de un Sistema de Recuperación de Información para su utilización en un Contexto Jurídico

Osvaldo Sposito[1], Hugo Ryckeboer[1], Julio Bossero[1], Edgardo Moreno[1], Viviana Ledesma[1], Gastón Procopio[1], Lorena Matteo[1], Cecilia Gargano[1], Victoria Saizar[1], Patricio Macias[1], Juan Ojeda[1], Fabio Quintana[1], Laura Conti[2], Sergio García[3] y Gustavo Pérez Villar[4]

[1] Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigación Tecnológicas. Florencio Varela 1903. San Justo. La Matanza.

{sposito, hugor, jbossero, ej_moreno, vledesma, gprocopio, lmatteo, cgargano, vsaizar, pmacias, fquintana, jmojeda}@unlam.edu.ar

[2] Universidad Nacional de La Matanza. Departamento Derecho y Ciencia Política. lconti@unlam.edu.ar

[3] Palacio de Tribunales. Departamento Judicial de Morón. Alte. Brown. Piso 4. Morón. sergiogabriel.garcia@pjba.gov.ar

[4] Subsecretaría de Tecnología Informática del Poder Judicial de la Provincia de Buenos Aires. Palacio de Justicia, avenida 13 entre 47 y 48, primer piso (La Plata). Argentina. gperez@scba.gov.ar

RESUMEN

En las últimas décadas, las instituciones públicas, particularmente el Poder Judicial (PJ), con el desarrollo de las TICs, han generado un importante aumento en: la generación de documentos digitales, en los repositorios de los mismos y en los Sistemas de Recuperación de Información (SRI). Este trabajo se orienta a estudiar y proponer soluciones para la recuperación de documentos judiciales, se hace una propuesta para la construcción de la matriz de términos en un proceso de indización.

Palabras clave: SRI, Modelo Vectorial, Indización, Lematización.

CONTEXTO

La línea de investigación aquí presentada es parte del proyecto de investigación "*Implementación de un Sistema Web de Recuperación de la Información Orientado a Documentación Jurídica con el Proceso de Indexación Semántica Latente Paralelizado",* perteneciente al programa de Investigaciones PROINCE (Programa de Incentivos para Docentes Investigadores) de la Secretaría de Políticas Universitarias del Ministerio de Educación de la Nación. Los integrantes del equipo son docentes e investigadores dependientes de las siguiente Unidades Académicas de la Universidad Nacional de La Matanza (UNLaM): el Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT), y el Departamento Derecho y Ciencia Política, además, colaboran personal técnico de la Subsecretaría de Tecnología Informática del Poder Judicial de la Provincia de Buenos Aires.

1. INTRODUCCIÓN

En este trabajo se describe un proceso de indización, que consiste en extraer una serie de términos, representativos de los temas tratados en un documento, para utilizarlos después como puntos de acceso para la recuperación de esos documentos de un corpus jurídico. El propósito es brindar jurisprudencia similar a los profesionales del derecho luego de realizar una consulta. Entendiendo el concepto de jurisprudencia, como el conjunto de las sentencias de distintos fallos dictados por los tribunales de justicia u organismos judiciales de un Estado. En el campo del derecho, la jurisprudencia juega un papel importante como fuente del derecho; por ser la comprensión e interpretación de las normas jurídicas basada en las sentencias pasadas emitidas por órganos oficiales,

estas sustentan la aplicación de la ley en un caso concreto. En el PJ se producen una enorme cantidad de documentos jurídicos (dictámenes, expedientes, etc.) cada año, lo cual produce que esta fuente de derecho sea cada vez mayor, lo que impulsa a los profesionales del derecho a dedicar más tiempo a la búsqueda de una decisión relevante.

Basándonos en [1] coincidimos, en que los SRI están en continua mejoría, esto se debe a: la incorporación de utilidades dependientes de la expansión de su uso, el avance de las aplicaciones tecnológicas y el claro deslinde de sus funciones.

En [2], se referencia a Calvin N. Mooers como quien introdujo por primera vez en 1950 el término Recuperación de Información (en inglés Information Retrieval) en la literatura de documentación, la definió como «*la búsqueda de información en un stock de documentos, efectuada a partir de la especificación de un tema*». Sólo un año más tarde, el mismo autor ampliaba esta definición al manifestar que la recuperación de información abarca los aspectos intelectuales de la descripción de información y su especificación para la búsqueda, y también cualquier sistema, técnica o máquina que se utilice para llevar a cabo la operación [3].

Según la bibliografía consultada [4-6], una SRI es un programa que interactúa entre un corpus y sus usuarios. Su efectividad depende del adecuado control del lenguaje de representación de los elementos de información y las búsquedas de sus usuarios. Para cumplir con sus objetivos, según Gabriel H. Tolosa y otros [5], un SRI debe realizar las siguientes tareas básicas:

- Representación lógica de los documentos y, opcionalmente, alma_ cenamiento del original.
- Representación de la necesidad de información del usuario en forma de consulta.
- Evaluación de los documentos respecto de una consulta para establecer la relevancia de cada uno.
- Ranking de los documentos considerados relevantes para formar el "conjunto solución o respuesta. Presentación de la respuesta al usuario.
- Retroalimentación de las consultas para aumentar la calidad de la respuesta.

Jaime Robredo en [5], asevera qué en cualquier área del conocimiento, los términos con significado se pueden utilizar como descriptores para representar el contenido de documentos escritos, en los procesos de indización y organización de la información, así como para formular preguntas en el proceso de recuperación de información. Tolosa en [5] afirma que el proceso se puede dividir en las siguientes etapas:

- Análisis lexicográfico: Se extraen las palabras y se normalizan.
- Reducción (Tokenización) de palabras vacías o de alta frecuencia.
- Lematización: Se reducen palabras morfológicamente parecidas a una forma base o raíz, con la finalidad de aumentar la eficiencia de un SRI.
- Asignación de pesos o ponderación de los términos que componen los índices de cada documento.

Los SRI implementan una gama diversa de estructuras de datos, algoritmos y técnicas de recuperación de información, por ello, se precisa de un modelo conceptual donde se determinen: el tipo de almacenamiento, operaciones sobre los términos, modelos de búsqueda con base patrones exactos o los modelos inexactos los cuales contendrán las técnicas probabilísticas, los modelos lógicos y los espacios vectoriales [8]. En el trabajo de Martínez Méndez, se puede encontrar un estudio detallado de los distintos modelos de RI existentes. Uno de los modelos más utilizados [4-5], es el Modelo de Espacio

Vectorial. En este modelo, el texto es representado por un vector de términos, los términos comúnmente son palabras; cualquier texto puede ser representado por un vector en un espacio dimensional Salton en el año 1975 [9]. En un espacio de documento que consiste en documentos D_i, cada uno identificado por uno o más términos de índice T_j; los términos pueden ser ponderados de acuerdo a su importancia, o no ponderados con pesos restringidos a 0 y 1. En el modelo los documentos se representan a partir de vectores, de la siguiente manera:

$$D_i = (T_1, T_2, ..., T_j) \quad (1)$$

En la Figura 1 se muestra un espacio de índice tridimensional, donde cada elemento se identifica con hasta tres términos distintos.

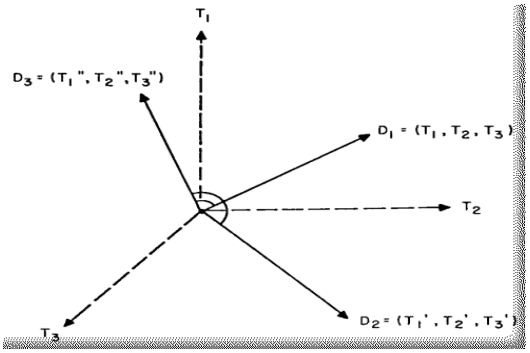

Figura. 1. Representación vectorial espacial de los documentos [9].

Una consulta se puede ver como un documento por lo tanto se puede ver como un vector.

Matemáticamente, una de formas de establecer la cercanía de dos vectores es calcular el coseno del ángulo que forman los dos vectores entre sí. Esta fórmula tiene la ventaja de su bajo esfuerzo computacional y es independiente de los módulos de los vectores. De manera similar, se puede calcular el coseno del ángulo entre cada vector de documento y el vector de consulta para encontrar su cercanía. Para encontrar un documento relevante para el término de la consulta, se calcula la puntuación de similitud entre cada vector del documento y el vector del término de la consulta aplicando la similitud del coseno. Finalmente, aquellos documentos con puntajes de similitud altos se considerarán documentos relevantes para la consulta. [9].

Como se comentó, dentro de la indización se encuentra la lematización, que es una técnica empleada en la recuperación de datos en los SRI, que sirve para reducir variantes morfológicas de la forma de una palabra a raíces comunes o lexemas; con el fin de mejorar la habilidad de los motores de búsqueda y, a consecuencia, los resultados de las consultas. Básicamente, este consiste en remover el plural, el tiempo, o los atributos finales de la palabra [5,6,10]. En el trabajo de González [6], afirma que "*cuando se realiza la extracción de palabras de un texto se obtiene una gran cantidad de entradas con formas verbales conjugadas y variantes de concordancia. Logrando la reducción morfológica de todas estas variantes se busca que el usuario recupere tanto los textos que contienen sus términos de búsqueda, como aquellos que contienen las formas derivadas de esos términos…*". Cabe aclarar que, en este proyecto, nosotros también simplificamos las apariciones de sustantivos y adjetivos. Los algoritmos de lematización más conocidos son: Lovins[1] (1968), Porter[2] (1980) y Paice[3] (1990). Originalmente todos fueron hechos para el inglés, y se diferencian en la eficiencia

[1] http://snowball.tartarus.org/algorithms/lovins/stemmer.html

[2] https://tartarus.org/martin/PorterStemmer/

[3] https://www.scientificpsychic.com/paice/paice.html

del código y la elección de sufijos que identifican y eliminan. Una modificación del algoritmo trabajo de Porter, es el algoritmo de Snowball[4]. Este puede mapear palabras que no están en inglés. Estos algoritmos permiten realizar "*derivaciones*", esto es remover los sufijos comunes morfológicos e inflexionales de palabras literalmente diferentes, pero con una "*raíz*" común, que pueden ser consideradas como un sólo término. Este algoritmo requiere de un conjunto de pasos para llegar a la raíz.

2. LÍNEAS de INVESTIGACIÓN y DESARROLLO

El presente trabajo tiene como eje central el desarrollo de un SRI. Entre las líneas de investigación a considerar en este proyecto se pueden mencionar:

- El problema de la recuperación de información, el modelo vectorial y la forma de almacenar los términos de una colección (corpus) de pruebas.
- La paralelización del proceso de Indexación Semántica Latente (ISL). Se estudian las librerías: Compute Unified Device Architecture (CUDA) y CUDA Basic Linear Algebra Subprograms (CuBlas), aplicadas a una arquitectura híbrida.
- La aplicando el patrón de arquitectura Modelo-Vista-Controlador (MVC), para desarrollos WEB. Aplicando el lenguaje de programación C#.
- Estudio de la librería REGEX., para resolver las Expresiones Regulares (ER).
- Estudio y evaluación de distintos algoritmos de ranking para Documentos. Las pruebas serán realizadas tomando como base un corpus jurídico real.

3. RESULTADOS OBTENIDOS/ESPERADOS

Durante el año 2021 se ha trabajado, principalmente, en dos temas, por un lado, en el estudio, análisis y modificación de algoritmos y técnicas que permitan la lematización de términos, y por otro, en el proceso que permita incorporar, de un corpus jurídico, las fechas y las referencias de la norma jurídica actual, mediante el Reconocimiento de Entidades Nombradas (tales como Acordadas, Artículos, Leyes, entre otros), que componen los distintos textos judiciales, utilizando Expresiones Regulares (ER).

Se presentaron en distintos congresos las siguientes publicaciones:

1. "***Propuesta para la construcción de un corpus jurídico utilizando Expresiones Regulares***". Presentado en el XXVII Congreso Argentino de Ciencias de la Computación (CACIC). Salta. Argentina [8].

Una ER es una notación algebraica para caracterizar un conjunto de cadenas [11]. Son particularmente útiles para la búsqueda en textos, cuando se tiene un patrón y un corpus de textos donde buscar. En este trabajo se demostró que es posible incorporar en el proceso de Análisis lexicográfico Expresiones Regulares para incorporar fechas y Entidades Nombradas a una matriz de términos. Dentro de las tareas a desarrollar, durante este año, se puede mencionar:

- Incorporar la codificación propuesta al SRI implementado por el proyecto PROINCE mencionado en la introducción.
- Analizar otros algoritmos y técnicas de derivación.
- Estudiar otras librerías existentes de ER.
- Realizar una clasificación de todas las EN dentro de la norma jurídica Argentina.

2. "***Implementación de un lematizador***

[4] https://snowballstem.org/demo.html

para la lengua española". Trabajo presentado en el Workshop del IX Congreso Nacional de Ingeniería en Informática/Sistemas de Información. CONAIISI 2021. Mendoza. Argentina.

En este trabajo se muestra una modificación realizada al algoritmo de Snowball. Mejorando en un 26% la lematización de términos. Se prevé para este año:

- Modificar el orden de los pasos, propuesto en el algoritmo de Snowball, para mejorar los tiempos de procesamiento.
- Estudiar nuevos métodos de derivación.
- Profundizar en el estudio de la morfología léxica, ciencia que estudia la estructura de las palabras y las pautas que permiten formarlas o derivarlas a partir de otras.

4. FORMACIÓN DE RECURSOS HUMANOS

La presente línea de investigación la lleva adelante un equipo de 15 integrantes provenientes de dos departamentos de la UNLaM, el DIIT y el Departamento de Derecho y Ciencia Política.

- 1 alumno de grado. En el año 2021 se gradúo en la carrera de Ingeniería de Informática.
- 2 asesores especialistas externos. (uno perteneciente al Poder Judicial de la Provincia de Buenos Aires y un Secretario de Juzgado).

5. BIBLIOGRAFÍA

[1]Galindo Ayuda, F. (2020). Avances en sistemas jurídicos de recuperación de documentos. Scire: Representación Y organización Del Conocimiento, 26(1), 63–74. https://doi.org/10.54886/scire.v26i1.4698. Fecha de consulta: 07/02/22

[2] S. Oliván, J.A., & Arquero Avilés, Rosario. (2006). Una aproximación al concepto de recuperación de información en el marco de la ciencia de la documentación. Investigación bibliotecológica, 20(41), 13-43. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2006000200002&lng=es&tlng=es. F. de consulta: 07/02/22

[3] C.N. Mooers, "The theory of digital handling of non–numerical information and its implications to machine economics", en Technical Bulletin No. 48. Cambridge, MA: Zator Co., 1950 (Ponencia presentada en Association for Computing Machinery, Rutgers Univ., New Brunswick, NJ, 1950, March 29).

[4] 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. Rev. Latinoamericana de Ingeniería de Software, (2014). 2(2): 107-114. http://revistas.unla.edu.ar/software/article/view/81. Fecha de consulta: 07/02/22

[5] Tolosa G. & Bordignon, F. Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos. UNdeL, Argentina, (2008). En línea: http://eprints.rclis.org/12243/1/Introduccion-RI-v9f.pdf. Fecha de consulta: 07/02/22

[6] González, C. M. La recuperación de información en el siglo XX. Revisión y aplicación de aspectos de la lingüística cuantitativa y la modeliza-ción matemática de la información UNLP. (2008) Disponible en: https://memoria.fahce.unlp.edu.ar/tesis/te.350/te.350.pdf. Fecha de consulta: 07/02/22

[7] Robredo, J. (2019). Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico. Ciência Da Informação, 47(1). Recuperado de http://revista.ibict.br/ciinf/article/view/4431. Fecha de consulta: 07/02/22.

[8] Martínez Méndez, F. (2004). Recuperación de información: modelos, sistemas y evaluación. Disponible en: http://eprints.rclis.org/16262/1/libro-ri.PDF. Fecha de consulta: 07/02/22.

[9] Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Information Retrieval. Communications of the ACM, 18(11), 613–620. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.446.5101&rep=rep1&type=pdf. F. consulta: 07/02/22.

[10] Zazo Rodríguez A. y otros. (2002). Recuperación de información utilizando el modelo vectorial. U. de Salamanca. Disponible en: http://eprints.rclis.org/13963/1/zazo2002recuperacion.pdf. Fecha de consulta: 07/02/22.

[11] Robaldo, L. y otros. Compiling regular expressions to extract legal modifications. 250. 133-141. 10.3233/978-1-61499-167-0-133. (2012).