INFORMATIZACIÓN PARA EL DESARROLLO SOSTENIBLE

TÍTULO: DESARROLLO DE UN BUSCADOR ESPECIALIZADO CON SOFTWARE LIBRE.

TITLE: DEVELOPMENT OF SPECIALIZED OPEN SOURCE SEARCH ENGINE

Armando de Jesús Plasencia Salgueiro[1], Sandra Lisset González Pedrón[2], María Josefa Peralta González[3]

1- Departamento de Control del Instituto de Cibernética, Matemática y Física, Cuba. Jefe de Departamento, Dr.C., Investigador Auxiliar, E-mail: armando@icimaf.cu
2- Universidad de la Isla de la Juventud "Jesús Montané Oropesa", Cuba. Profesora Auxiliar. MSc. E-mail: spedron@uij.edu.cu
3- Universidad Central de las Villas. Cuba. Profesora Titular. Dra. E-mail: peralta.glez@uclv.edu.cu.

Resumen: Encontrar literatura relevante es crucial para muchas actividades de investigación, en las energías renovables y la biomédica y en la práctica de la medicina basada en evidencia. Los Motores de búsqueda como PubMed proporcionan un medio para buscar y recuperar literatura publicada, dada una consulta. Sin embargo, estos están limitados en la forma en que los usuarios pueden controlar el procesamiento de consultas y los artículos, por el motor de búsqueda. Para posibilitar este control tanto a investigadores biomédicos como informáticos que trabajan en la recuperación de información biomédica o las energías renovables, se desarrolló una herramienta para la búsqueda de literatura biomédica y de otro tipo. La configuración propuesta se guía por la configuración del NIST de las tareas de evaluación TREC relevantes en la medicina de precisión. Se propuso un motor de búsqueda construido con Lucene SolR que debe estar disponible de forma pública para los investigadores que quieran explorar diferentes estrategias de búsqueda sobre literatura científica y en particular biomédica. Describimos

varias estrategias de formulación de consultas y proponemos sus evaluaciones con juicios humanos conocidos para una gran cantidad de temas, por ejemplo, energía renovable y medicina de precisión utilizando algoritmos especializados como Lingo y STC.

Palabras Clave: Evaluación, Recuperación de Información, Energías renovables, medicina de precisión, Buscador especializado.

Abstract: Finding relevant literature is crucial for many biomedical research activities and in the practice of evidence-based medicine. Search engines such as PubMed provide a means to search and retrieve published literature, given a query. However, they are limited in how users can control the processing of queries and articles—or as we call them *documents*—by the search engine. To give this control to both biomedical researchers and computer scientists working in biomedical information retrieval, was developed a tool for searching over biomedical and other literature. Our setup is guided by the NIST setup of the relevant TREC evaluation tasks in precision medicine. Was proposed a search engine building with Lucene SolR which should be publicly available for the researchers who want to explore different search strategies over published biomedical literature. We outline several query formulation strategies and present their evaluations with known human judgements for a large pool of topics, for example renewable energy and precision medicine using specialized algorithms like Lingo and STC.

Keywords: Evaluation, Information retrieval, renewables energies, Precision medicine, Specialized Search Engine.

1. Introducción

La comunidad de las ciencias, y en particular las ciencias de la vida, está experimentando un crecimiento sin precedentes en la disponibilidad de recursos textuales. Los repositorios de citas, como PubMed [1], han registrado recientemente más de un millón de artículos nuevos en año. Es decir, PubMed, pudo tener un aumento de 2.4 millones de registros en un lapso de dos años, desde principios de 2017 hasta principios de 2019. Por un lado, significa que hay cada vez más información al alcance de la mano. Sin embargo, por otro lado, encontrar la literatura y la evidencia científica relevantes en tal volumen de datos es un desafío.

El área de Ciencias de la Computación que se enfoca en mejorar los mecanismos de búsqueda de información relevante sobre una colección de recursos, se conoce como Recuperación de Información (RI). El acceso a la información es fundamental en las ciencias de la vida; por ejemplo, los científicos deben mantenerse al día con los últimos hallazgos de la ciencia y contrastar su trabajo con los hallazgos reportados, y los profesionales médicos deben recomendar tratamientos clínicos adecuados a sus pacientes. Como resultado, la RI biomédica se ha convertido en un campo de investigación relevante.

Tanto las comunidades de RI como las de las ciencias de la vida han reconocido ampliamente la necesidad de mejores sistemas de búsqueda [2].

Medir la eficacia de los sistemas de búsqueda requiere evaluaciones exhaustivas. Un órgano conocido para establecer métricas y marcos de evaluación comunes en la comunidad de RI es la Conferencia de Recuperación de Texto (TREC) organizada anualmente por el Instituto Nacional de Ciencia y Tecnología (NIST). TREC ha proporcionado una serie de medios de evaluación técnica y biomédica. Estos medios incluyen TREC Precision Medicine (2017-2020) [3-5] (PM), así como TREC-COVID [6].

Si bien TREC ha proporcionado recursos cruciales, la investigación biomédica en RI todavía se ve obstaculizada por tres factores importantes: (1) la reproducibilidad relativamente baja de los métodos de investigación y sus resultados reportados; (2) el componente de ingeniería de software de la búsqueda biomédica; y (3) la falta de líneas de base universalmente accesibles y reproducibles.

Sin duda, las tareas compartidas de TREC han llevado a una proliferación en la actividad de investigación de RI por la introducción de bases comunes para la evaluación. No obstante, la eficacia de una búsqueda realizada por el sistema incorpora muchas decisiones de diseño complejas, que incluyen: formulación de consultas, preprocesamiento de documentos, implementación de motores de búsqueda y métodos de clasificación. Esto conlleva a tres obstáculos en la evaluación. Primero, si se evalúa un sistema de búsqueda complejo, no es posible aislar la contribución de ninguna técnica o método específico sin un experimento específico. En segundo lugar, la ejecución de experimentos de RI requiere la implementación de un motor de búsqueda, que a su vez requiere indexar documentos, diseñar una estructura de índice, analizar los archivos de

origen, y transmitir los resultados de la búsqueda a una herramienta de evaluación. En otras palabras, un investigador del procesamiento del lenguaje, por ejemplo, con una técnica prospectivamente para la formulación de consultas, como la detección de entidades nombradas con expansión de sinónimos, debe configurar una infraestructura completa de nivel industrial solo para ejecutar un experimento conceptual de prueba. Finalmente, el tercer problema es que la investigación de RI implica tomar muchas decisiones en ocasiones contradictorias, como qué motor elegir o cómo preprocesar documentos. Las líneas de base universalmente reproducibles son difícilmente alcanzables y, por lo tanto, los nuevos métodos no se basan en la investigación existente. Esto significa que es difícil saber si un nuevo método mejora la efectividad de la recuperación en general, o solo para componentes específicos de un experimento en particular.

En virtud de ello se desarrolló una herramienta con el objetivo de minimizar esos tres problemas. Esta herramienta proporciona una interfaz gráfica de usuario (GUI) basada en la web fácil de usar, que permite a los usuarios definir y ejecutar experimentos de IR, incluidos los biomédicos y explorar los resultados de su evaluación.

Esta herramienta permite la evaluación de técnicas de formulación de consultas dentro de un marco predefinido, pero parametrizable. Cualquiera puede reproducir fácilmente los resultados de la evaluación de cada experimento proporcionando entradas idénticas al sistema.

2. Metodología

La falta de resultados comparables en RI es un problema no resuelto en la comunidad de RI. Por otro lado, hay varios motores de búsqueda de código abierto disponibles para la comunidad de RI, como Apache Solr, Lucene, Elastic search, Terrier [7] y Galago [8], que pueden ser utilizados por desarrolladores e investigadores profesionales con preparación en la recuperación de información. Sin embargo, conlleva a un proceso de aprendizaje con una pendiente elevada al necesitar configurarlos incluso para una tarea de recuperación básica [9].

La metodología para el desarrollo del motor de búsqueda especializado empleada fue la CRISP-DM.

3. Resultados

3.1 Construcción del sistema de RI.

En el Departamento de Control del Instituto de Cibernética, Matemática y Física se desarrolló un buscador especializado.

Como resultado del trabajo realizado se detallan los pasos llevados a cabo siguiendo las recomendaciones de la metodología para desarrollo de buscadores [10].

Apache Nutch (en su versión 1.9) es la herramienta con la que se ha implementado el rastreador y va a nutrir al buscador.

La cobertura se refiere al tamaño del conjunto de páginas recuperadas dentro de un cierto período de tiempo. Un rastreador exitoso trata de maximizar su cobertura con el objetivo de proporcionar una colección grande localizable a los usuarios. De forma similar, el refrescamiento de la colección es importante con el objetivo de minimizar la diferencia entre las copias captadas de la Web y sus originales y así mantener la información del servidor actualizada.

Apache Solr (en su versión 4.3.0) es la herramienta en el proyecto que contiene el servidor de búsqueda y va a interactuar con la interfaz que se le muestre al usuario. Esta herramienta depende en su funcionamiento de un contenedor o servelets, para la solución se utilizó Tomcat (en su versión 7.0.21).

Para poder garantizar el correcto funcionamiento de Solr es necesario dominar los conceptos referidos a la configuración del fichero schema.xml, que es donde se definen las características con que se va a realizar el indexado.

También en este fichero se establece la Tokenización, proceso mediante el cual Solr analiza el contenido de los documentos, descomponiendo el texto en elementos ("tokens"). Con ello, puede entregar como resultado de una búsqueda resultados que contienen sinónimos de los términos buscados, elementos que son fonéticamente equivalentes, o elementos que tienen la misma raíz (stem), elimina las palabras poco significativas en las búsquedas (stop words), utilizando para ello los analizadores (Analyzers, Tokenyzers, Filters).

Otras configuraciones adicionales que se agregaron fue el código para integrar Nutch y Solr, la comunicación con la herramienta Luke que permite el análisis de los índices de Lucene y Tika para el análisis de documentos estructurados.

Luego de realizado el proceso de puesta a punto de Solr se puede acceder a la interfaz de administración vía web (se muestra en el anexo). Esta interfaz permite gestionar las colecciones de documentos y realizar análisis y consultas sobre las mismas, entre otras muchas funcionalidades.

Carrot 2 en su versión 3.9.3 es la herramienta que contiene los algoritmos de agrupamiento para documentos. Esta herramienta permite el trabajo mediante dos alternativas diferentes, desplegado como servicio web con el fichero carrot-webapp.war o instalado como aplicación de escritorio utilizando el paquete carrot2-workbench. Para el desarrollo del presente trabajo se ejecutaron ambas alternativas y se integraron satisfactoriamente con Solr.

Al finalizar la configuración y puesta en marcha de los módulos del sistema de RI se obtuvo una herramienta capaz de rastrear, indexar y mostrar resultados de las búsquedas realizadas.

3.2. Desarrollo de la etapa de Preparación de los Datos

Para el desarrollo de la etapa de *Preparación de los Datos:*

Se inicializó el crawler implementado con Nutch 1.4 indexando con Solr 4.0. Como resultado se obtuvo un set de datos de 22163 documentos de las urls detalladas en el documento con anterioridad, el resultado se muestra en la figura 1.

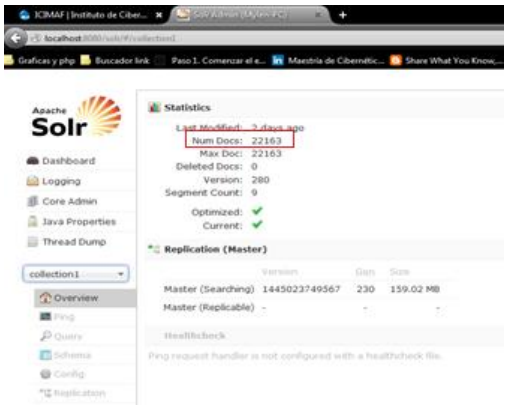

Figura 1. Cantidad de documentos recuperados [10].

Se realizó un estudio del índice creado utilizando la herramienta LUKE en su versión 4.6.1 cómo se puede observar en la Figura 2. Del análisis realizado se obtuvo la siguiente información.

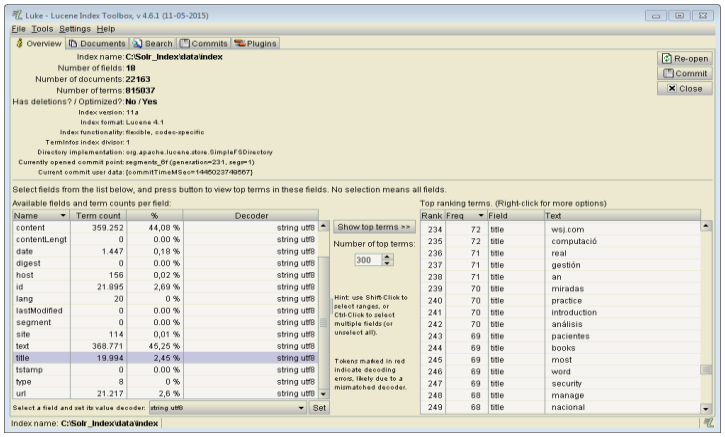

Figura 2 Análisis del índice con la herramienta Luke.

3.3. Desarrollo de la etapa Modelado

Para la mejor comprensión de los algoritmos utilizados se detallan elementos de su funcionamiento considerados de importancia en el estudio realizado.

STC (Suffix Tree Clustering)

El algoritmo Clúster de árbol de sufijos funciona en dos fases principales: la fase de descubrimiento del clúster de base. En la primera fase es construido un árbol de sufijos generalizado de todas las frases del texto, utilizando las palabras como elementos básicos. Después se procesan todas las sentencias, los nodos del árbol contienen información acerca de los documentos en los que aparecen las frases particulares.
El uso de esta información permite que los documentos que comparten la misma frase se agrupan en racimos de base de los cuales sólo se conservan aquellos cuya puntuación supera una predefinida como Puntuación Base Mínima del Clúster (del inglés Minimal Base Cluster Score).

El algoritmo Lingo por otra parte, para convertirse en un algoritmo de agrupamiento con todas las funciones, debe en el proceso de búsqueda de etiquetas ir precedido de algún pre procesamiento de la colección de entrada. Esta etapa debe abarcar el filtrado de texto, reconocimiento del lenguaje de documento, despalillado y de identificación de palabras.

También se recomienda que se eliminen los grupos con idéntico contenido. Por lo que el trabajo del algoritmo dependerá en gran manera de los resultados que ofrezca el buscador Solr implementado.

3.4. Desarrollo de la etapa Evaluación del proceso

Para evaluar los resultados de los algoritmos y métodos de análisis, en la etapa de modelado se desarrolló un caso de estudio. El objetivo es poder, de la información ya rastreada y centralizada por el buscador, sintetizar mediante la creación de clúster, los documentos similares y lograr que los investigadores ganen en tiempo a la hora de definir el estado de arte de una determinada temática.

Para el caso de estudio se escogió la temática control de aerogeneradores. Se definieron como palabras clave: *aerogenerador*, *energía eólica*, *renovable*.

Para el término *aerogeneradores*, se obtuvieron un total de 58 coincidencias y 21 clúster creados con Lingo:

La información que se pudo identificar relevante fue (Figura 3):

- ✓ La etiqueta *turbina de viento* con 8 documentos con información general de aerogeneradores, tipos de energías eólicas, material que se utiliza para fabricarla.

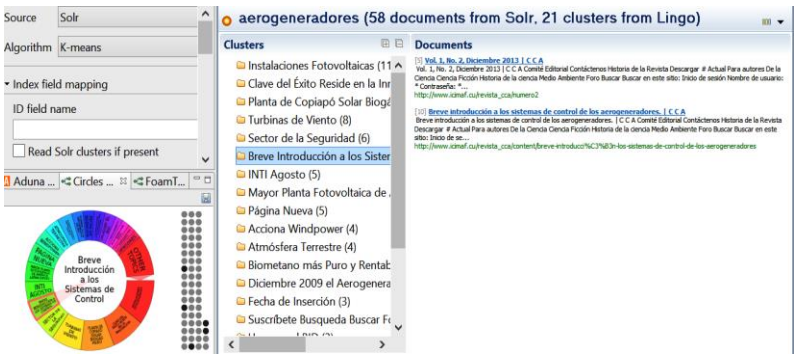

Figura 3. Resultados de la consulta aerogeneradores utilizando el agrupamiento mediante algoritmo Lingo.

Haciendo uso del algoritmo Kmeans se crearon 23 clusters. El uso de este algoritmo para el conjunto de datos resultó ser de gran utilidad para la identificación de los términos más manejados en la temática dentro de la colección documental creada. Tal como se muestra en la figura 4, cada grupo contiene documentación referente a los términos identificados.

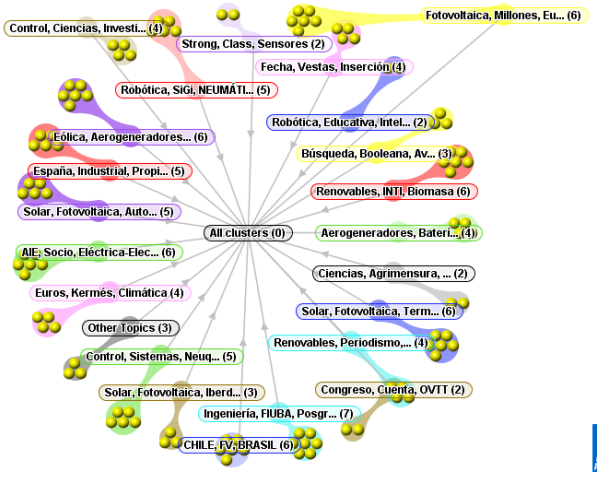

Figura 4. Resultados de la consulta energía eólica renovable utilizando el agrupamiento mediante algoritmo Kmeans y la visualización de esquema relacional aduna.

Se utilizó en las búsquedas el algoritmo STC, pero las etiquetas ofrecidas no fueron de impacto en el estudio.

Otra de las ventajas de trabajar con Carrot 2 es que permite exportar los resultados de las búsquedas a formato XML. Esto brindó la posibilidad de realizar análisis utilizando técnicas de mapas auto organizados mediante la herramienta VosViewer para analizar la información referente a la ocurrencia de palabras lo cual constituye un indicador de gestión muy útil.

4. Conclusiones

Se proporciona una plataforma de software que permite a investigadores de diversos perfiles comparar los sistemas de recuperación, en campos específicos como el de la energía eólica la medicina de precisión por poner dos ejemplos. Con el fin de realizar una evaluación completa de diferentes metodologías de búsqueda, la herramienta permite al usuario especificar formalmente los métodos que utilizan, volver a ejecutar experimentos y analizar los resultados en función del marco de evaluación oficial de TREC.

Su interfaz gráfica de usuario permite un análisis cualitativo y cuantitativo detallado del rendimiento de recuperación utilizando algoritmos como Lingo y STC.

Al utilizar la plataforma, los investigadores de RI podrán informar de sus métodos de manera coherente y evaluar sus resultados frente a una referencia común. En el futuro, nuestro objetivo es utilizar esta plataforma para evaluar sistemáticamente una combinación más diversa de métodos búsqueda.

5. Referencias bibliográficas

1. NCBI: Pubmed-NCBI. https ://www.ncbi.nlm.nih.gov/pubme d/. Accessed 23 Oct 2021.

2. Hersh W, Detmer WM, Frisse ME. Information-retrieval systems, chap. 15, p. 539–72.

3. Roberts K, Demner-Fushman D, Voorhees E, Hersh WR, Bedrick S, Lazar A, Pant S. Overview of the TREC 2017 Precision Medicine track. In: TREC, Gaithersburg, MD. 2017.

4. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ. Overview of the TREC 2018 Precision Medicine track. In: TREC, Gaithersburg, MD. 2018.

5. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, Pant S, Meric-Bernstam F. Overview of the TREC 2019 Precision Medicine track. In: TREC, Gaithersburg, MD. 2019.

6. Roberts K, Alam T, Bedrick S, Demner-Fushman D, Lo K, Soboroff I, Voorhees E, Wang LL, Hersh W. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. J Am Med Inform Assoc. 2020;. https ://doi.org/10.1093/jamia /ocaa0 91.

7. Ounis I, Amati G, Plachouras V, He B, Macdonald C, Johnson D. Terrier information retrieval platform. In: ECIR, vol 3408; 2005. p. 517–9.

8. Cartright M-A, Huston S, Feild H. Galago: a modular distributed processing and retrieval system. In: OSIR@SIGIR; 2012. p. 25–31.

9. Rybinski M., et al. A2A: a platform for research in biomedical literature search. *From* Joint NETTAB/BBCC 2019 Meeting - Network Tools and Applications in Biology (NETTAB) & Bioinformatics and Computational Biology Conference Salerno, Italy. 11-13 November 2019.

10. Montero M. Sistema de recuperación y análisis de información no estructurada para apoyar el proceso de investigación en el ICIMAF. Tesis presentada en opción al título de Máster en Cibernética Aplicada. La Habana. 2015.