Motor de búsqueda semántico de contenido académico en repositorios digitales bajo el protocolo OAI-PMH

José Federico Medrano[1], José Luis Alonso Berrocal[2], Carlos G. Figuerola[2]

jfmedrano@fi.unju.edu.ar, berrocal@usal.es, figue@usal.es

[1]VRAIn / Visualización y Recuperación Avanzada de Información / Facultad de Ingeniería

Universidad Nacional de Jujuy - Ítalo Palanca 10, +54 (388) 4221587

[2] REINA / Recuperación de Información Avanzada / Facultad de Traducción y Documentación

Universidad de Salamanca – España - C/ Francisco de Vitoria, 6-16

RESUMEN

A menudo las búsquedas de material académico que uno realiza no son del todo acertadas o no son tan exactas como uno quiere. Los resultados ofrecidos por los motores de búsqueda dependen en gran medida de los mecanismos internos utilizados y de los algoritmos de ordenación, tal es el caso de *Google Scholar* que emplea el *Page Rank* (Page, Brin, Motwani, & Winograd, 1999) para ordenar los resultados. Por otro lado, los repositorios digitales institucionales carecen de un buscador de material relacionado, puesto que los resultados de las búsquedas se basan en la existencia de algunos de los términos buscados en los campos de metadatos de los registros almacenados. Por esta razón, este proyecto propone la construcción de un metarepositorio que recolecte todos los registros de los repositorios digitales argentinos pertenecientes a instituciones educativas que implementan el protocolo *Open Archives Initiative Protocol for Metada Harvesting* y que permita realizar búsquedas semánticas de contenido relacionado a partir de una búsqueda inicial.

Palabras clave: *Buscador semántico; Repositorios digitales; OAI-PMH; PLN*

CONTEXTO

La línea de investigación aquí presentada se encuentra enmarcada dentro del Proyecto Consolidado D/B029 denominado "*Aplicación de técnicas de Inteligencia Artificial para evaluar la producción científico-académica de investigadores de Universidades públicas del Noroeste Argentino*", aprobado y financiado por la Secretaría de Ciencia y Técnica y Estudios Regionales de la Universidad Nacional de Jujuy. Este proyecto es llevado a cabo en conjunto por dos grupos de investigación. En primer lugar liderado por el grupo de investigación VRAIn (Visualización y Recuperación Avanzada de Información) de la Facultad de Ingeniería de la Universidad Nacional de Jujuy, y en segundo lugar como colaboradores, el grupo REINA (Recuperación de Información Avanzada) de la Facultad de Traducción y Documentación de la Universidad de Salamanca

1. INTRODUCCIÓN

Son numerosas las instituciones y entidades que necesitan no solo preservar el material y las publicaciones que producen, sino también, estas

tienen como tarea publicar, divulgar y poner a disposición del público los resultados de la investigación y cualquier otro material científico-académico. Para este propósito existen los repositorios de libre acceso, que a través de iniciativas como la *Open Archives Initiative* (OAI) y de la aparición de instrumentos como el protocolo *Open Archives Initiative Protocol for Metada Harvesting* (OAI-PMH), facilitan esta tarea en gran medida.

El protocolo OAI-PMH propone unos lineamientos generales tanto para listar y recuperar (cosechar) metadatos de un repositorio (OAI Service Providers), como también para exponer recursos (OAI Data Providers) para que puedan ser cosechados por aplicaciones externas. Estos lineamientos proponen la organización de los recursos en conjuntos (sets), el uso del estándar XML para la representación y transporte de recursos (vía HTTP), y un conjunto de seis verbos necesarios para interactuar, como por ejemplo identificar el repositorio, listar conjuntos, listar formatos de metadatos soportados u obtener registros (Medrano, 2017)

La mayor parte del tiempo, los investigadores deben filtrar varios documentos académicos para encontrar aquellos relevantes para su investigación. Este filtrado muchas veces es engorroso y requiere emplear una considerable cantidad de tiempo. En la búsqueda de este tipo de material resulta útil contar con un listado de objetos relacionados no sólo con la temática buscada, sino también material que pueda estar relacionado semánticamente con el objeto de la búsqueda. Los repositorios institucionales (RI) poseen un motor de búsqueda interno, el cual ofrece un conjunto de resultados a partir de una búsqueda por palabra o frase, autor o fecha; buscando dentro de todas las colecciones o una colección en particular. Para los RI que utilizan como software de base a *DSpace*, el motor de búsqueda, dependiendo de la versión, es *Lucene* o SOLR o combinaciones de ambos. Estos motores permiten un gran número de opciones al momento de realizar la búsqueda (Prasad & Patel, 2005), pero son búsquedas sintácticas, es decir, se busca entre los registros almacenados la coincidencia de los términos ingresados realizando distintas combinaciones con los campos de metadatos de los mismos. Los resultados presentados, a partir de una búsqueda inicial, no ofrecen ninguna otra relación más allá de la existencia/coincidencia de algún término en común.

Ofrecer contenido semánticamente relacionado aportaría significativamente mejores resultados que búsquedas sintácticas, como fue comprobado en (Medrano, 2018). Por ello la necesidad de encarar este proyecto y extender el trabajo inicial mencionado aportando una herramienta capaz de ofrecer resultados "*más*" relevantes. La búsqueda semántica es una técnica de búsqueda de datos en la que una consulta de búsqueda tiene como objetivo no solo encontrar palabras clave, sino también determinar la intención y el significado contextual de las palabras que una persona está utilizando para la búsqueda.

La idea central es recolectar periódicamente los repositorios institucionales argentinos, para esto se consultará el ROAR (*Registry of Open Access Repositories*) que es una de las bases de datos de repositorios de libre acceso más grandes que existen, posee alrededor de 4734 repositorios registrados, con 57 repositorios pertenecientes a instituciones argentinas. El interés fue puesto en los repositorios que soportan el protocolo OAI-PMH, puesto que es un protocolo estándar para la recolección de metadatos y es el objeto de estudio de este trabajo.

La herramienta que se pretende desarrollar, una vez recuperados y almacenados los registros, se presentará como un sistema de recomendación

basado en contenido (Ricci, Rokach, & Shapira, 2011; Bai, y otros, 2019), en este tipo de sistemas el contenido desempeña un papel principal en el proceso de recomendación, en el que las calificaciones de los usuarios y las descripciones de los atributos de los elementos se aprovechan para hacer predicciones. La idea básica es que los intereses del usuario se puedan modelar sobre la base de las propiedades (o atributos) de los elementos que han calificado o accedido en el pasado. Los "elementos" suelen ser textuales, por ejemplo, correos electrónicos (Paik, y otros, 2001) o páginas web (Seroussi, 2010). La "interacción" generalmente se establece mediante acciones, como descargar, comprar, crear o etiquetar un artículo. Los elementos están representados por un modelo de contenido que contiene las características de los elementos. Las características suelen estar basadas en palabras, es decir, palabras sueltas, frases o n-*grams* (Beel, Gipp, Langer, & Breitinger, 2016).

Algunas de las aproximaciones relacionadas con sistemas de recomendación de contenido académico o relacionado a la educación están bien clasificadas y resumidas en (Beel, Gipp, Langer, & Breitinger, 2016), los autores mencionan algunos enfoques como sistemas de recomendación de libros, sistemas de recomendación educativa, servicios de alerta académica, búsqueda de expertos, resumen automático de artículos académicos, recomendadores de conjuntos de datos académicos y detección de plagio. En el mismo sentido, otras de las aproximaciones es el enfoque de (Son & Kim, 2018) el cual propone un sistema de recomendaciones para artículos académicos que combina el análisis de citas y el análisis de redes. Por otro lado, (Hwang, Wei, & Liao, 2010) plantea un sistema de recomendación basado en un esquema híbrido que emplea redes de coautoría y técnicas basadas en contenido.

El enfoque que aquí se propone consta de 2 etapas bien diferenciadas:

1. Proceso de recolección de información: el cual, como se mencionó previamente, consistirá en recolectar por completo, y de forma periódica, los RI Argentinos listados en ROAR y que soportan el protocolo OAI-PMH, es decir, aquellos que respetan el formato de metadatos Dublin Core[1] de manera obligatoria.
2. Diseño del modelo de predicción: para este proceso es necesario aplicar técnicas de Procesamiento del Lenguaje Natural (PLN) que permitan representar de forma adecuada la información y calcular/computar el grado de cercanía semántica entre una búsqueda inicial y el resto de registros almacenados.

Para el diseño del modelo de predicción, es necesario representar de algún modo la información, en este caso el conjunto de registros, para esto los campos utilizados en este estudio serán los títulos (campo *title*) y las descripciones (campo *description*, empleado para almacenar el resumen del elemento digital) por ser los más representativos de cada registro.

Para este trabajo se decidió utilizar la librería *Gensim* (Rehurek & Sojka, 2010) del lenguaje *Python*[2]. La elección del lenguaje y librería responden al hecho de que en el ámbito de ciencia de datos, *Python* es el lenguaje más versátil y quien mejor desempeño ofrece, además existe un gran número de librerías que reducen enormemente las tareas de desarrollo al incorporar implementaciones de los algoritmos y técnicas más utilizados. En este sentido las técnicas de representación de documentos seleccionadas para el desarrollo de los

[1] http://dublincore.org/

[2] https://www.python.org/

experimentos fueron: *Term Frequency Inverse Document Frequency* (TF-IDF) (Aizawa, 2003), *Latent Semantic Indexing/Latent Semantic Analysis* (LSI/LSA) (Landauer & Dumais, 1997; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) y *Word Mover's Distance* (WMD) (Kusner, Sun, Kolkin, & Weinberger, 2015; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Para cada una de las técnicas mencionadas se diseñará un modelo de predicción el cual, tomando el título y la descripción de un resultado elegido, ofrecerá un conjunto de publicaciones relacionadas semánticamente con el mismo.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de trabajo principal de este proyecto de investigación es el estudio, evaluación e implementación de técnicas de representación de documentos (PLN) que trabajarán sobre los registros recolectados de RIs para ofrecer un motor de búsqueda semántico de material académico. Este acercamiento pretende mejorar los procesos de búsqueda y filtrado que son llevados a cabo al momento de realizar una búsqueda bibliográfica.

Abarca las siguientes áreas y disciplinas: Recuperación de Información, Bases de Datos, Desarrollo Web, Inteligencia Artificial, PLN, entre otras.

3. RESULTADOS OBTENIDOS/ESPERADOS

En una primera instancia se espera emplear las técnicas de representación de documentos elegidas y evaluar los resultados entregados tanto de forma manual como comparándolos con motores de búsqueda actuales. Para luego seleccionar la técnica o combinación de estas que mejor resultados entregue.

Una vez que los resultados demuestren ser fiables y se ajusten a ciertos parámetros de calidad, se espera poder publicar la herramienta para que sea accesible y utilizada de forma libre por la comunidad académica.

4. FORMACIÓN DE RECURSOS HUMANOS

Los autores de este trabajo han desarrollado investigaciones sobre la temática presentada. El director del mismo, el Dr. J. Federico Medrano, se encuentra realizando investigaciones postdoctorales relacionadas con bases de datos bibliográficas de libre acceso, entre las que se encuentran los Repositorios Institucionales como una fuente adicional de material científico-académico. Actualmente es becario postdoctoral de la Fundación Carolina[3] y realiza su investigación en conjunto con la Universidad de Salamanca, donde presentan su apoyo y colaboración el Dr. José Luis Alonso Berrocal y el Dr. Carlos G. Figuerola.

Por otro lado, en este proyecto formarán parte alumnos del último año de la carrera Ingeniería Informática de la Universidad Nacional de Jujuy en donde imparte docencia el director. De los cuales se espera puedan desarrollar su trabajo fin de carrera con temáticas afines.

5. BIBLIOGRAFÍA

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management, 39*(1), 45-65.

[3] https://www.fundacioncarolina.es/

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific Paper Recommendation: A Survey. *IEEE Access, 7*, 9324--9339.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries, 17*(4), 305-338.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science, 41*(6), 391.

Hwang, S., Wei, C., & Liao, Y. (2010). Coauthorship networks and academic literature recommendation. *Electronic Commerce Research and Applications, 9*(4), 323-334.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. *International Conference on Machine Learning*, (pp. 957-966).

Landauer, T., & Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*(2), 211.

Medrano, J. F. (2017). Calidad en repositorios digitales en argentina, estudio comparativo y cualitativo. *VII Conferencia Internacional BIREDIAL-ISTEC'17 y XII SIBD.* La Plata.

Medrano, J. F. (2018). Filtrado basado en contenido para artıculos académicos en repositorios institucionales. *Congreso Argentino de Ciencias de la Computación (CACIC).* Tandil.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Neural Information Processing Systems*, 3111-3119.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* Stanford InfoLab.

Paik, W., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., & Amice, C. (2001). Applying Natural Language Processing (NLP) based metadata extraction to automatically acquire user preferences. *Proceedings of the 1st international conference on Knowledge capture.*, (pp. 116-122).

Prasad, A., & Patel, D. (2005). Lucene search engine: An overview. *DRTC-HP International.*

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* (pp. 45-50). Valletta, Malta: ELRA.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. *Recommender systems handbook*, 1-35.

Seroussi, Y. (2010). Utilising user texts to improve recommendations. *International Conference on User Modeling, Adaptation, and Personalization* (pp. 403-406). Springer.

Son, J., & Kim, S. (2018). Academic paper recommender system using multilevel

simultaneous citation networks.
Decision Support Systems, 105, 24-33.