Minería de texto y Deep Learning aplicados a determinar la pertenencia de las consultas realizadas a un metabuscador a cada área temática dentro de las ciencias de la computación

Kuna H.D[1]., Rambo A.R.[1], Canteros A.[1], Rey, M.[1], Zamudio, E.[1], Martini, E.[1], Pautsch, G. [1], Biale, C. [1], Krujoski, S. [1], Rauber, F. [1]

[1]Depto. de Informática, Facultad de Ciencias Exactas Químicas y Naturales (FCEQyN), Universidad Nacional de Misiones (UNaM) Posadas, Misiones 3300/Argentina.

hdkuna@gmail.com

RESUMEN

Al trabajar en un metabuscador que permita identificar cuáles son las mejores recomendaciones en un área temática específicas según las consultas ingresadas por el usuario existen diversos problemas que abordar. En el presente trabajo se realiza el análisis de la pertenencia de una consulta escrita en lenguaje coloquial a una o varias áreas de temáticas dentro de las ciencias de la computación. Para lo cual se podrán tomar varios criterios y se realiza un relevamiento del estado de arte específico con la finalidad de determinar las mejores estrategias tanto para el tratamiento del texto como para la clasificación de los artículos. En este trabajo se presenta el relevamiento realizado para abordar y resolver esta problemática.

Palabras clave: *minería de texto, metabuscadores, recuperación de información, pertenencia a áreas temáticas, recomendación.*

CONTEXTO

Esta línea de investigación articula el Programa de Investigación en Computación (PICom) de la Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones (FCEQyN/UNaM) con el grupo de investigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España, y con el Departamento de Matemáticas de la Universidad de Sonora, México.

1 INTRODUCCION

Dentro de las actividades científicas se encuentra la búsqueda de antecedentes y datos científico-tecnológicos que permitan determinar el estado del arte en el área temática bajo análisis. En la actualidad no solo el volumen de información generada es un problema si no también contar con herramientas que permitan filtrar y cruzar si no todas al menos la gran parte de fuentes confiables teniendo en cuenta estándares predefinidos y consensuados por el cuerpo científico. Entre las diferentes herramientas que ayudan al investigador a recuperar datos de calidad y relevantes se pueden mencionar: los buscadores científicos, los repositorios digitales, redes sociales del ambiente y soluciones de procesamiento de grandes bases de datos (Tang, 2016), (Ortega y Aguillo, 2014), (Lee et. al., 2006).

En el presente proyecto de investigación acreditado “Diseño y construcción de procesos de explotación de información para el área de las ciencias de la computación”, aprobada por RES RECTORAL 601/18 y RES CD 274/18 se han abordado diferentes problemáticas como el análisis sobre los nombres de los autores permitiendo la desambiguación de los mismos (Kuna et. al. 2019). La búsqueda y definición de expertos tema de una tesis de grado (Cantero, 2019), la evaluación y recomendación de autores definiendo diferentes métricas tema de una tesis de grado (Canteros et. al., 2018b), (Canteros A., 2020). Para la prueba e integración de estas propuestas elaboradas y siguiendo la línea de

investigación se han abordado algunos de los problemas mencionados y se ha desarrollado un Sistema de Recuperación de Información (SRI), concretamente un metabuscador, que opera sobre documentos científicos del área de Ciencias de la Computación (Kuna et. al., 2013), (Kuna et. al., 2014). En el desarrollo del SRI se han planteado diferentes avances en torno a la conformación de una base de datos (BD) interna para el metabuscador que permitiera la generación de soluciones como las mencionadas en el apartado anterior (Kuna et. al, 2017), (Canteros et. al., 2018b).

2 LÍNEAS DE INVESTIGACION, DESARROLLO E INNOVACIÓN

Si bien se ha trabajado y logrado avances en varias cuestiones sobre el metabuscador, la presente línea de investigación propone abordar otra de las problemáticas detectadas al momento de generar las búsquedas hacia la estructura del SRI, el desarrollo de una técnica para determinar la pertinencia de la búsqueda a determinadas áreas específicas, lo que sería otorgar un nivel de simplicidad y especificidad para el metabuscador.

Sobre esta línea de trabajo se detectan a su vez que dado el conjunto de datos con los que contamos en la estructura propuesta, lo que nos sería de utilidad plantear para la presente implementación sería por un lado definir el método que permita determinar a qué área temática dentro de las ciencias de la computación, según la taxonomía de la *Association for Computing Machinery* - ACM (ACM 2008), (ACM, 2012), pertenece cada consulta realizada, replicar similar criterio para los autores y para los artículos, tomando como referencia previa los artículos analizados y guardados.

3 RESULTADOS Y OBJETIVOS

En el presente proyecto se prevé tomar el almacenamiento, del metabuscador desarrollado, las consultas y los perfiles de los autores en la base de datos según se observa en la fig. 1. Los datos de consultas almacenados en formato JavaScript Object Notation, (notación de objeto de JavaScript) JSON y ejecutadas con el metabuscador, los datos de perfiles de autores almacenados en la base de datos, los artículos devueltos y los resultados seleccionados por los usuarios.

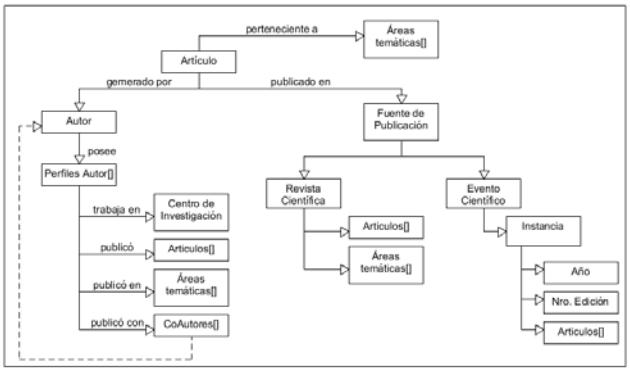

Figura 1: Esquema de artículos, autores y fuentes de publicación en el metabuscador

Uno de los temas principales en la minería de texto es la representación de texto, que es fundamental e indispensable para el procesamiento de información inteligente basado en texto. En general, la representación de texto incluye dos tareas: indexación y ponderación, por lo tanto, el primer abordaje es definir el criterio para la representación de las frases a ser analizadas. Los espacios vectoriales semánticos los cuales se basan en la idea de que el significado de una palabra puede ser aprendido de un entorno lingüístico, el de las áreas de la ciencias de la computación en nuestro caso, y poseen dos enfoques, la semántica distribucional y la semántica composicional (Torres López & Arco García, 2016).

Se ha estudiado comparativamente TF-IDF *(term frequency inverse document frequency)* y LSI *(latent semantic indexing)* como representación de varias palabras para texto. Donde se ha demostrado que, en la categorización de texto, LSI tiene un mejor rendimiento que otros métodos. Donde se observa que LSI tiene una calidad semántica y estadística favorable y es diferente con la afirmación de que LSI no puede producir poder discriminatorio para la indexación (Zhang et. al, 2011). Una aplicación de este mismo esquema de representación se observa en (Xamena et.al., 2019) donde se utiliza bolsa de palabras básica y una representación TF-IDF para los documentos históricos y para la tarea de modelado de temas, se empleó el método *LDA (Linear discriminant analysis)*.

El modelo Fast Text propuesto por (Bojanowski et. al., 2017) donde según este enfoque, basado en el modelo de skipgram, cada palabra se representa como una bolsa de caracteres n-gramas. Una representación vectorial está asociada a cada carácter n-gramo; palabras representadas como la suma de estas representaciones. El método denominado Fast Text demuestra ser rápido y permite calcular representaciones de palabras para palabras que no aparecieron en los datos de entrenamiento.

Otros trabajos (Mikolov et. al., 2013a) presentan varias extensiones que mejoran la calidad de los vectores y la velocidad de entrenamiento para el modelo continuo de Skip-gram (Mikolov et. al, 2013b). Por submuestreo de las palabras frecuentes obtienen una aceleración significativa y también se logra que aprenda representaciones de palabras más regulares. También describen una alternativa simple al softmax jerárquico llamado muestreo negativo donde logran buenas representaciones para los vectores de aprendizaje para millones de frases.

De esta manera se obtendrán representaciones de las secuencias de texto, para alimentar esto a arquitecturas Deep Learning de Clasificación de Textos que actualmente se encuentran en revisión. Como es la propuesta del estudio que toma como base la Wikipedia[1], teniendo en cuenta que este es un gran volumen de datos y proponen la creación del corpus que define una categoría mediante técnicas de Procesado de Lenguaje Natural (PLN) que analizan sintácticamente los textos a clasificar. El resultado final del sistema propuesto presenta un alto porcentaje de acierto, incluso cuando se compara con los resultados obtenidos con técnicas alternativas de Aprendizaje Automático (Quinteiro-González et. al, 2011).

(Zhang et. al, 2015), como el trabajo donde presentan una nueva arquitectura (*Very Deep Convolutional Neuronal Networks* - VDCNN) para el procesamiento de texto que opera directamente a nivel de personaje y usos solo pequeñas convoluciones y operaciones de agrupamiento. Demostrando que el rendimiento de este modelo aumenta con la profundidad en clasificaciones de textos públicos (Conneau et.al., 2016).

[1] Wikipedia. La Enciclopedia Libre. Ultima visita: 03/03/2020. Disponible en: https://www.wikipedia.org/

4 FORMACION DE RECURSOS HUMANOS

Este proyecto es parte de las líneas de investigación del "Programa de Investigación en Computación" de la FCEQyN de la UNaM, con cuatro integrantes relacionados con las carreras de Ciencias de la Computación de la UNaM. De los cuales dos están realizando su tesis de pos-grado, dos se encuentran realizando tesis de grado.

5 BIBLIOGRAFIA

1. Association for Computing Machinery: "CS2008 Curriculum Update: The Computing Curricula Computer Science Volume is complete and approved", 2008.
2. Association for Computing Machinery: "Computer Science Curricula 2013 Strawman Draft (February 2012)", 2012.
3. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.
4. Cantero, L. (2019) "Desarrollo de un Proceso de Recomendación de Autores para un Sistema de Recuperación de Información de Ciencias de la Computación". Tesis de Grado, Licenciatura en Sistemas de Información. F.C.E.Q.yN. – Universidad Nacional de Misiones. Argentina.
5. Canteros A., Rey M., Kuna H., (2018a) "Clasificación de autores para un proceso de recomendación integrado a un metabuscador científico," in XXIV Congreso Argentino de Ciencias de la Computación - Libro de Actas, Tandil, Argentina.
6. Canteros A., Zamudio E., Kuna H. D., (2018b) "Desambiguación de autores para un sistema de recuperación de expertos en un contexto académico," in XIX Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 47, CABA, Argentina.
7. Canteros, A. (2020) "Desambiguación de Autores para un Sistema de Recuperación de Información de Ciencias de la

Computación". Tesis de Grado, Licenciatura en Sistemas de Información. F.C.E.Q.yN. – Universidad Nacional de Misiones. Argentina.
8. Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for text classification. arXiv preprint arXiv:1606.01781.
9. Kuna H., Rey M., Martini E., Solonezen L., Podkowa L., (2013) "Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación," Rev. Latinoam. Ing. Softw., vol. 2, no. 2, pp. 107–114.
10. Kuna H., Rey M., Podkowa L., Martini E., Solonezen L., (2014) "Expansión de Consultas Basada en Ontologías para un Sistema de Recuperación de Información," XVI Workshop de Investigadores en Computación. RedUNCI. Argentina.
11. Kuna H. et al., (2017) "An Entity Profile Schema for Data Integration in an Academic Metasearch Engine," in Proceedings of the 2017 International Conference on Artificial Intelligence, Las Vegas, USA, pp. 281–285.
12. Kuna H. Rambo A., Zamudio E., Cantero A., Canteros A., Biale C., Martini R., Rauber F., Pautsch G. Krujoski S., Rey M. (2019) "Avances en el Desarrollo de Métodos de Desambiguación y Recomendación de Autores Científicos para un Metabuscador de las Ciencias de la Computación", XXI Workshop de Investigadores en Ciencias de la Computación, San Juan, RedUNCI - USJ. Argentina.
13. Li, H., Councill, I., L, e W.-C, Giles , C. L., (2006) "CiteSeerx: An Architecture and Web Service Design for an Academic Document Search Engine," in Proceedings of the 15th International Conference on World Wide Web, New York, NY, USA, pp. 883–884.
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
15. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. ICLR Workshop.
16. Ortega J. L., Aguillo I. F., (2014) "Microsoft academic search and Google scholar citations: Comparative analysis of author profiles," J. Assoc. Inf. Sci. Technol., vol. 65, no. 6, pp. 1149–1156.
17. Quinteiro-González, J. M., Martel-Jordán, E., Hernández-Morera, P., Ligero-Fleitas, J. A., & López-Rodriguez, A. (2011). Clasificación de textos en lenguaje natural usando la Wikipedia. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, (8), 39-52.
18. Tang J., (2016) "AMiner: Mining Deep Knowledge from Big Scholar Data," in Proceedings of the 25th International Conference Companion on World Wide Web, Geneva, Switzerland, pp. 373–373.
19. Torres López, Carmen, & Arco García, Leticia. (2016). Representación textual en espacios vectoriales semánticos. Revista Cubana de Ciencias Informáticas, 10(2), 148-180. Recuperado en 30 de marzo de 2020, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992016000200011&lng=es&tlng=es.
20. Xamena, E., Marmanillo, W. G., & Mechaca, A. L. (2019). Rebuilding the Story of a Hero: Information Extraction in Ancient Argentinian Texts. In V Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2019)-JAIIO 48 (Salta).
21. Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765.
22. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).