AVANCES EN EL DESARROLLO DE MÉTODOS DE DESAMBIGUACIÓN Y RECOMENDACIÓN DE AUTORES CIENTÍFICOS PARA UN METABUSCADOR DE LAS CIENCIAS DE LA COMPUTACIÓN

H. Kuna, A. Cantero, A. Canteros, M. Rey, E. Zamudio, A. Rambo, E. Martini, G. Pautsch, C. Biale, S. Krujoski, F. Rauber

Departamento de Informática, Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones.

hdkuna@gmail.com

RESUMEN

En el ámbito de la recuperación y procesamiento de datos relacionados a publicaciones científico-tecnológicas existen diversos problemas a resolver. En este trabajo son de especial interés aquellos relacionados con el tratamiento de datos de los autores de tales publicaciones.
La identificación unívoca de un autor, la constitución de un perfil del mismo compilando su historial de publicaciones, un conjunto de métricas que permitan estimar su impacto en la comunidad científica y la selección de un subconjunto de autores a partir de una consulta del usuario de un metabuscador, son algunos de los problemas que se tratan en la presente línea de investigación.
En este trabajo se presentan los avances en dos soluciones planteadas: las correspondientes a un método de desambiguación de autores y, por otra parte, un método de recomendación de autores basado en clases previamente definidas.
Mediante los resultados que se reseñan en este documento se pretende contribuir a la mejora del desempeño de procesos de explotación de información asociados a la recuperación de producciones científico-tecnológicas.

Palabras clave: producción científico-tecnológica, autores científicos, recuperación de información, desambiguación, 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 INTRODUCCIÓN

La recuperación de información relativa a producciones científicas a través de internet es uno de los desafíos actuales de la actividad científica. Los grandes volúmenes de publicaciones, la diversidad de fuentes de consulta y de herramientas de ayuda al investigador junto a la expansión de los datos que se generan a partir de la interacción entre científicos, son las bases de este problema.
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 [1], [2]. Dentro de este último grupo, se agrupan herramientas que proveen algunos de los siguientes servicios [1], [3]: integración de resultados; generación automática de perfiles de autores, publicaciones o centros de investigación; visualización de redes de colaboración entre autores; clasificación automática de contenidos; cálculo de métricas relativas al impacto o reconocimiento de autores y/o publicaciones; comparación de perfiles de entidades, entre otros.
En la presente 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 [4]–[6]. 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 [7]–[9]. En el desarrollo de esta BD se ha identificado uno de los ejes del presente trabajo, el problema de la desambiguación de entidades [9] y haciendo uso del contenido de la misma se ha generado el método de recomendación de autores que completa los avances a reseñar.

Se denomina entidades a todos los actores involucrados en los procesos del metabuscador, de los cuales se han definido perfiles para el almacenamiento de sus datos y su disponibilidad para implementar procesos que hicieran uso de los mismos para contribuir a mejorar los resultados que el metabuscador presenta a sus usuarios [7]. La gestión de este tipo de contenido requiere implementar procesos que ayuden a identificar unívocamente aquellas entidades asociadas a las producciones científicas. En paralelo, se consideró adecuado asistir al investigador generando recomendaciones de entidades que pudieran resultar relevantes para su investigación. En función de este objetivo se generó el método de recomendación de autores que considera diferentes aspectos de los mismos para establecer una clasificación [8] y posteriormente elabora un listado para presentar al usuario aquellos que pueden ser de utilidad dada su consulta.

1.1 DESAMBIGUACIÓN DE AUTORES

La identificación de un autor de una publicación científica es primordial para conocer la obra del mismo y poder obtener referencias a trabajos previos que permitan conocer su historial de trabajo. En la actualidad esta actividad no puede ser desarrollada manualmente por los grandes volúmenes de datos disponibles, es así como surgen diferentes procesos y métodos que tienen por objeto facilitar esta tarea. Tales recursos son los que se agrupan dentro de lo que se denomina desambiguación de entidades. En una instancia inicial, dentro del presente grupo de investigación se ha determinado comenzar por la desambiguación de los autores registrados en la BD del metabuscador. Casos como: publicaciones de un mismo autor con distintos nombres, diferentes autores con un mismo nombre, ausencia de datos en los registros de la publicación, entre otros, son los que generan inconsistencias en la identificación unívoca de autores.

En la literatura del área de ciencias de la computación se encuentran diferentes métodos que constituyen alternativas de solución para este problema. Entre las técnicas de base utilizadas se pueden identificar métodos de agrupamiento y de clasificación. En todos los casos se plantea la necesidad de contar con un volumen adecuado de datos para asegurar la calidad de los resultados [10]–[12].

1.2 RECOMENDACIÓN DE AUTORES MEDIANTE PERFILES

Las tareas de búsqueda de expertos en un área de conocimiento son conocidas en la comunidad científica. La construcción de perfiles de expertos o *expert profiling* busca tomar y organizar datos a fin de evidenciar la experiencia de los autores en una temática en particular. Estas evidencias podrán variar en términos de las fuentes de datos empleadas para la construcción del perfil, sin embargo, las publicaciones científico-tecnológicas de un autor se consideran el mejor reflejo de su desempeño.

Para contar con una BD de autores se comenzó por el diseño de los perfiles a través de los que serían registrados los mismos y sus publicaciones. Posteriormente se desarrollaron de métodos de extracción, transformación y carga (ETL, por su sigla en inglés) para obtener datos desde diferentes fuentes y unificarlos en la BD generada. Sobre este conjunto de datos se generó una serie de clases en las cuales se etiquetó a los autores en función de sus antecedentes, para luego determinar su recomendación al usuario [8]. En este trabajo se presentan los avances relacionados a las acciones finales del método correspondientes con la selección de los autores a recomendar.

2 LÍNEAS DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN

La presente línea de investigación propone como objetivo general, el desarrollo de procesos de explotación de información para su implementación en un sistema de recuperación de información de producciones científicas del área de las Ciencias de la Computación.

Actualmente se está trabajando sobre dos tareas derivadas de tal objetivo. Los procesos de desambiguación de entidades, en particular de autores, son necesarios para obtener datos de calidad sobre los cuales luego se pueda aplicar algún tipo de técnica de reconocimiento de patrones. Por otra parte, la recomendación automática de autores constituye un primer acercamiento a los objetivos planteados, permitiendo presentar al usuario del metabuscador datos relativos a su consulta que van más allá de las publicaciones que vayan a ser recuperadas.

El aspecto de desambiguación, ha sido abordado a partir del relevamiento de técnicas de procesamiento de lenguaje natural, aprendizaje automático, y análisis de redes sociales, a partir de los datos extraídos de las producciones científicas. Los resultados parciales se presentan en las próximas secciones. Mientras que para el caso del método de recomendación se presentan los avances en el desarrollo de los componentes finales del mismo.

3 RESULTADOS Y OBJETIVOS

En este apartado se presentan los avances logrados en ambas líneas de trabajo.

3.1 MÉTODO DE DESAMBIGUACIÓN DE AUTORES

Se hizo un relevamiento de métodos de desambiguación actuales, considerando aspectos tales como: los datos que utilizan, las métricas para validar su funcionamiento y su impacto en los sistemas de información [9].

Uno de los métodos es utilizado por un repositorio académico actual, AMiner [1]. El mismo consiste en un *framework* basado en Campos Ocultos Aleatorios de Markov (o *Hidden Markov Random Fields*). Consiste en un modelo probabilístico que explora tanto las relaciones entre documentos como los atributos de cada documento.

Se formaliza el problema de desambiguación de la siguiente manera. Se tiene un conjunto de publicaciones, que representan las variables observables del Modelo de Markov. Dichos artículos están escritos por autores con un mismo nombre. Se aplica un método de *clustering* sobre el conjunto de *papers* en el que cada cluster corresponde a un autor diferente. Dado que se desconoce el número de autores, es por ello que el conjunto de autores representa el campo oculto aleatorio en el modelo de Markov. El mismo *framework* contiene un método de estimación del número de autores para el conjunto de publicaciones.

El objetivo a corto plazo en esta línea de trabajo es reproducir el algoritmo de AMiner para comprender mejor su funcionamiento y contrastarlo con un método de desambiguación a definir. Posteriormente se realizará un análisis de ambos métodos y en base a los resultados se determinará cuáles de estos métodos se adapta mejor para su implementación en el metabuscador.

3.2 MÉTODO DE RECOMENDACIÓN DE AUTORES

Tanto el esquema empleado para el almacenamiento de los datos de las entidades con las que opera el metabuscador, como una reseña parcial del proceso de recomendación de autores han sido presentados previamente [7], [8]. En lo que respecta al método de recomendación, en la publicación mencionada se detallaron cuestiones relativas a las fuentes de datos a utilizar, los aspectos que serían incluidos en la evaluación de los autores y las métricas asociadas a los mismos, dando lugar a la definición de las clases en función de las que serían catalogados y al algoritmo que efectúa la clasificación en sí misma.

A continuación, se presentan los componentes restantes del método, en particular aquellos que serán utilizados a partir del ingreso de la consulta del usuario en el metabuscador. Se propuso para el proceso de recomendación el uso de la técnica de filtrado basado en contenido [13], la cual plantea la generación

de recomendaciones a partir del estudio del contenido de los ítems a recomendar y las preferencias del usuario.
El proceso se compone de tres etapas sucesivas que se desarrollan a continuación:
[a] Para la etapa inicial, denominada de "**similaridad**" entre consulta y autores, se utilizaron los datos pertenecientes a las publicaciones de cada autor de la BD del SRI. Sobre los campos de título, resumen y palabras clave de cada publicación almacenada, se buscaron coincidencias con los términos de la consulta del usuario.
Resultando en un conjunto reducido de autores "*candidatos*" que se corresponden con los requerimientos del usuario. En este caso se simplificó el proceso seleccionando a todos los autores en cuyos perfiles se encontraron coincidencias con los términos de la consulta.
[b] Para la segunda etapa, que se denominó de "**relevancia**", el objetivo es determinar la relevancia de los autores candidatos para la recomendación. Para ello se establecieron los criterios, la técnica y las definiciones consideradas para la evaluación de los autores, los elementos utilizados y adaptados para el escenario propuesto, y la propuesta de un algoritmo para su aplicación.
Entre las técnicas aplicables para esta fase [14], [15], se determinó emplear un modelo de espacios vectoriales el cual permite desarrollar el siguiente postulado: "*los autores más relevantes para un término dado, pueden ser evaluados según la calidad de sus trabajos, para lo que existen métricas ampliamente aceptadas en la comunidad científica*". En función de ese principio, el proceso de filtrado toma en consideración dos factores: las métricas que determinan el impacto de un autor, siendo las mismas empleadas en el método de clasificación [8]; y métricas que estiman el peso de los términos de la consulta en los datos presentes en las publicaciones del autor. Para esto último se empleó el índice TF-IDF (*Term frequency – Inverse document frequency*) que es utilizado para este tipo de tareas en la literatura del área [16].
[c] Finalmente, en la fase de "**reparto**", se listan los ítems que resultan similares y relevantes para los usuarios. El desafío en esta etapa radica en la tarea de presentar el mayor número de ítems relevantes con el menor número de falsos positivos y por ello se busca la visibilidad de ítems con la mayor relevancia posible. Se propuso aplicar un método que permita emplear las clases de autores identificadas previamente [8].
Complementariamente, se utilizaron métodos de reparto como los empleados en elecciones gubernamentales, particularmente el de Sante-Laguë [17]. Considerando a las listas o partidos políticos equivalentes a cada una de las clases que pueden pertenecer los autores; y a la cantidad de votos de una lista o partido equivalente a la cantidad numérica de autores agrupados dentro de cada clase. Es decir, que la cantidad de autores de una clase definirá la cantidad de "*lugares*" que tendrá cada categoría en la presentación al usuario.
Estas operaciones se ejecutan sobre el conjunto de autores procesado en las etapas previas, resultando en un conjunto de autores a presentar al usuario como recomendación ante su consulta.
Actualmente el método completo se encuentra implementado y su fase de validación está siendo finalizada para ser presentado integralmente en una futura publicación.

4 FORMACIÓN DE RECURSOS HUMANOS

Este proyecto cuenta con once integrantes relacionados con las carreras de Ciencias de la Computación de la UNaM. En resumen, el grupo de investigación desarrolla tres tesis de grado articulando sus trabajos con becas de Estímulo a las Vocaciones Científicas del Consejo InterUniversitario Nacional (CIN), tres tesis de maestría, dos de ellas enmarcadas en becas del Programa Estratégico de Formación de Recursos Humanos en Investigación y Desarrollo (PERHID) del CIN, y un trabajo de investigación posdoctoral con beca del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Asimismo, la línea y el equipo de investigación se vinculan con investigadores de la Universidad de Castilla-La Mancha, España y de la Universidad de Sonora, México.

5 BIBLIOGRAFÍA

[1] J. Tang, "AMiner: Mining Deep Knowledge from Big Scholar Data," in *Proceedings of the 25th International Conference Companion on World Wide Web*, Geneva, Switzerland, 2016, pp. 373–373.

[2] J. L. Ortega and I. F. Aguillo, "Microsoft academic search and Google scholar citations: Comparative analysis of author profiles," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 6, pp. 1149–1156, 2014.

[3] H. Li, I. Councill, W.-C. Lee, and C. L. Giles, "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, 2006, pp. 883–884.

[4] H. Kuna, M. Rey, E. Martini, L. Solonezen, and L. Podkowa, "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, 2013.

[5] H. D. Kuna, E. Martini, and M. Rey, "Evolución de un algoritmo de ranking para documentos científicos del área de las ciencias de la computación," XX Congreso Argentino de Ciencias de la Computación (Bs. As., 2014), 2014.

[6] H. Kuna, M. Rey, L. Podkowa, E. Martini, and L. Solonezen, "Expansión de Consultas Basada en Ontologías para un Sistema de Recuperación de Información," XVI Workshop de Investigadores en Ciencias de la Computación, 2014.

[7] H. Kuna *et al.*, "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, 2017, pp. 281–285.

[8] A. Cantero, M. Rey, and H. Kuna, "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, 2018.

[9] A. Canteros, E. Zamudio, and H. D. Kuna, "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, 2018)*, CABA, Argentina, 2018.

[10] Y. Liu, W. Li, Z. Huang, and Q. Fang, "A fast method based on multiple clustering for name disambiguation in bibliographic citations," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 3, pp. 634–644, 2015.

[11] Y. Qian, Q. Zheng, T. Sakai, J. Ye, and J. Liu, "Dynamic author name disambiguation for growing digital libraries," *Inf. Retr. J.*, vol. 18, no. 5, pp. 379–412, Oct. 2015.

[12] A. F. Santana, M. A. Gonçalves, A. H. F. Laender, and A. A. Ferreira, "Incremental author name disambiguation by exploiting domain-specific heuristics," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 4, pp. 931–945, 2017.

[13] R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation," in *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*, 2000, pp. 47–56.

[14] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[15] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based Recommender Systems: State of the Art and Trends," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 73–105.

[16] F. Cacheda, "Introduction to the Classic Models of Information Retrieval," *Rev. Gen. Inf. Doc.*, vol. 18, p. 365, 2008.

[17] V. R. González and A. L. Carmona, "Sistemas electorales basados en la representación proporcional," *eXtoikos*, no. 6, pp. 29–39, 2012.