DISEÑO Y CONSTRUCCIÓN DE PROCESOS DE EXPLOTACIÓN DE INFORMACIÓN PARA EL ÁREA DE CIENCIAS DE LA COMPUTACIÓN

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

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

hdkuna@gmail.com

RESUMEN

El desarrollo de procesos de explotación de información aplicado a la recuperación de producciones científico-tecnológicas, exige resolver problemáticas específicas relacionadas con las entidades del dominio, tales como publicaciones, instituciones, y autores.
Continuando con la línea general de este grupo de investigación, este trabajo aborda dos problemáticas actuales para la adecuada gestión de un servicio de recuperación de información en general, y del área de las Ciencias de la Computación en particular. Estas problemáticas incluyen la desambiguación y la recomendación de entidades asociadas a las producciones cientifico-tecnológicas.
El abordaje propuesto de estas problemáticas se enfoca principalmente en la aplicación de técnicas como el Procesamiento de Lenguaje Natural, Aprendizaje de Máquina y Análisis de Redes Sociales.
Estas propuestas preveen la evaluación experimental en el contexto de un servicio de metabuscador de publicaciones científicas del área de las Ciencias de la Computación, desarrollado y mantenido por este grupo de investigación.
Mediante este trabajo se pretende contribuir en la mejora del desempeño de procesos actuales de explotación de información asociados a la recuperación de producciones científico-tecnológicas.

Palabras clave: producciones científico-tecnológicas, recuperación de información, desambiguación, recomendación

CONTEXTO

Está 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-Universidad de Sonora-México.

1 INTRODUCCIÓN

La recuperación de producciones científico-tecnológica de relevancia a través de la web es uno de los desafíos actuales que involucra la actividad científica. El volumen de publicaciones disponibles, la variedad de fuentes y los datos que se generan a partir de la interacción entre publicaciones, autores y demás entidades involucradas, constituyen los ejes de este problema. En la actualidad, existen distintas alternativas que permiten a los usuarios acceder a tales contenidos, parte de ellas son las herramientas de búsqueda web y los repositorios digitales.

Este tipo de herramientas, se encuentran en auge en la actualidad. En particular, el área de las Ciencias de la Computación agrupa organizaciones que implementan servicios de búsqueda de algún tipo[1]–[3] como son los repositorios de instituciones como ACM, IEEE, Elsevier, Springer, ArXiv, DBLP y DOAJ, entre otros; los motores de búsqueda de

Google Scholar y Microsoft Academics [4]; y los repositorios digitales de documentos científicos que son gestionados por instituciones educativas o científicas, y que contienen datos primarios de su producción científico-tecnológica[5]–[7].
En nuestro país, a partir de la creación del Sistema Nacional de Repositorios Digitales [8], se cuenta con una herramienta de búsqueda, que incluye más de 100.000 documentos de diferentes tipos y de todo el país.

Este contexto evidencia la necesidad de enfoques que asistan al proceso de recuperación de información a través de la implementación de funcionalidades destinadas a satisfacer las necesidades de los usuarios. Así, los procesos de explotación de información pueden servir de soporte para acciones como: integración de resultados, especificación del orden de relevancia de los resultados, clasificación automática de los materiales, visualización de redes de colaboración entre entidades (autores, equipos de investigación, instituciones) y la identificación unívoca de contenidos, entre otros.

Este problema ha sido abordado previamente por este grupo de investigación desde el área de Recuperación de Información (RI), particularmente dirigido a la búsqueda de producciones científicas en el área de Ciencias de la Computación.
Como consecuencia de trabajos previos, este grupo ha desarrollado un metabuscador de producciones científicas para el área de las Ciencias de la Computación. Este metabuscador ha experimentado varias transformaciones desde el momento de su concepción[9]–[12].
Entre los elementos o módulos que conforman el mencionado metabuscador, se incluyen: un algoritmo de ranking basado en las propiedades de las publicaciones recuperadas [9], un método de expansión de consultas basado en ontologías de un área disciplinar [10] y un conjunto de operaciones destinadas a la gestión de los datos de las entidades con las que opera el metabuscador [11], [12].

Estos avances han contribuido significativamente a mejorar la recuperación de publicaciones científicas en el área de las Ciencias de la Computación. Sin embargo, estos mismos avances han planteado nuevos desafíos en relación a la gestión interna del metabuscador, como así también a los servicios ofrecidos por éste.

Por una parte, se evidencia la necesidad de gestionar adecuadamente las entidades que almacena el metabuscador para asistir a los procesos de búsqueda, recuperación, y presentación de resultados. Esta gestión de las entidades, requiere implementar procesos de desambiguación que ayuden a determinar unívocamente aquellas entidades asociadas a las producciones científicas. Estas entidades incluyen autores, instituciones, y títulos, entre otros.
Por otra parte, es necesario asistir al usuario del metabuscador a partir de recomendaciones de producciones científicas que pudieran resultar relevantes para su investigación. Esta tarea requiere implementar procesos de recomendación que ayuden a identificar entidades que podrían ser relevantes para el investigador. En particular, la recomendación de autores requiere identificar, entre otros aspectos, las áreas de experiencia de dichos autores. Así es que resulta necesario también investigar sobre los temas asociados a los autores de las producciones científicas.

1.1 RECOMENDACIÓN MEDIANTE PERFILES DE EXPERTOS

En particular, cuando las entidades recomendadas son autores, el sistema de recuperación de información puede intentar determinar las áreas de conocimiento asociadas a dichos autores. Esta tarea se conoce como Perfilado de Expertos o Expert Profiling. La construcción de Perfiles de Expertos se encuentra estrechamente relacionada con las fuentes de datos disponibles para determinar la evidencia de experiencia de las personas. En el ámbito académico, pueden utilizarse varias fuentes de evidencia de experiencia, como los CV, páginas personales, y otros. Sin embargo, una de las principales fuentes de evidencia de

experiencia son las mismas producciones científico-tecnológicas de los autores.
El principal problema de utilizar producciones científico-tecnológicas como fuente primaria de experiencia en la construcción de perfiles de expertos, es que estas producciones se componen principalmente por texto, el cual debe procesarse de algún modo, para extraer algún tipo de información.
La generación automática de perfiles de expertos en el ámbito académico utilizando producciones científico-tecnológicas como principal fuente de experiencia, implica la necesidad de desarrollar y evaluar técnicas avanzadas de computación, principalmente pertenecientes al área del Procesamiento de Lenguaje Natural y del Aprendizaje de Máquina.
El desarrollo de estrategias para la generación automática de perfiles de expertos se presenta como una alternativa para ofrecer nuevas funcionalidades a los sistemas de recuperación de información en el ámbito académico. Asimismo, la generación de los perfiles de expertos en ámbitos académicos permite extender sus aplicaciones a problemas frecuentes del contexto, como ser la recomendación de expertos para evaluación de proyectos de investigación, becarios, asignación de subsidios, o conformación de paneles de expertos para el tratamiento de temas específicos o consultoría. Asimismo, y considerando los ámbitos de los primeros desafíos en la generación de perfiles de expertos relacionados con el ámbito empresarial, las contribuciones alcanzadas en un contexto incentivan su adaptación y aplicación en otros contextos como la selección de expertos para el tratamiento de temas específicos, selección de personal en el ámbito laboral, y selección de comités, entre otros.

1.2 DESAMBIGUACIÓN DE AUTORES

Cuando se publica una producción científica, se desea conocer al autor de la misma, y en algunos casos, si éste ha publicado otras obras [13]. Con la masividad de información en internet, dicha búsqueda se vuelve más compleja. Una tarea que antes era llevada a cabo manualmente analizando datos descriptivos del autor, y a veces sus textos, es casi imposible de realizar hoy en día.
La actividad para solucionar esta problemática es lo que se conoce como desambiguación de autores.
Algunos ejemplos en que se requiere desambiguar los autores, incluyen: publicaciones de autores con distinto nombre, o diferentes autores con un mismo nombre. Además, puede ocurrir que la información sobre el autor sea insuficiente o presente inconsistencias en su identificación. Por ejemplo: algunas editoriales no publican el primer nombre de los autores, o su información geográfica, títulos que poseen o áreas en las que son expertos, entre otros.
En un SRI, la identificación del autor es necesaria en la aplicación de las métricas de calidad, por ejemplo, en un algoritmo de ranking.
Los desafíos que presenta la desambiguación de autores han llevado al desarrollo de varios métodos. En [14], se presenta una taxonomía jerárquica caracterizando los distintos métodos, y haciendo referencia a aquellos que son más representativos. De acuerdo con esta taxonomía, los métodos se clasifican: según su aproximación, en agrupamiento de autorer [15]–[32] y asignación de autores. Esta última a su vez se divide en clasificación [33], [34] y clustering [35]–[37]. Alternativamente, estos métodos se pueden clasificar según la evidencia explorada en aquellos que utilizan datos presentes en las citas, los que utilizan información en la web [23], [27], [32], [34], y los que extraen datos implícitos a partir de lo que esté disponible [22].

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.

Este objetivo deriva en dos aspectos principales. El primer aspecto consiste en investigar sobre los procesos de recomendación automática de datos asociados a producciones científicas, como autores o publicaciones. El segundo aspecto consiste en investigar sobre los procesos de desambiguación de entidades correspondientes a producciones científicas, principalmente autores.

El aspecto de recomendación, será abordado utilizando datos internos del metabuscador, con el objetivo de analizar los perfiles de los autores ahí almacenados.
Se prevé la evaluación de un conjunto de algoritmos de recomendación para ser aplicado sobre un conjunto de autores que podrían ser de interés para el usuario del metabuscador, a partir de la consulta que haya ingresado.

El aspecto de desambiguación, será abordado a partir de técnicas integradas de procesamiento de lenguaje natural, aprendizaje automático, y análisis de redes sociales, a partir de los datos extraídos de las producciones científicas.

Ambos aspectos de la línea de investigación general, prevee evaluaciones de tipo experimental en el contexto del metabuscador ya desarrollado.

3 RESULTADOS Y OBJETIVOS

El desarrollo de los procesos de explotación de información, que conforman el objetivo principal de esta línea de trabajo, incluye los siguientes objetivos particulares:

- Desarrollar estrategias que permitan la generación automática de perfiles de expertos en el ámbito de sistemas de recuperación de información académicos mediante la aplicación y desarrollo de técnicas avanzadas de Aprendizaje de Máquina y Procesamiento de Lenguaje Natural.
- Desarrollar métodos de recomendación de contenido para las entidades con las que opera el metabuscador, siguiendo la línea del sistema de recomendación de autores, brindando resultados en forma paralela a la ejecución de las consultas del usuario.
- Desarrollar métodos para la gestión de los datos con los que opera el metabuscador, tal como el de desambiguación de entidades, detección de outliers, y la especificación de una taxonomía propia para las entidades almacenadas, entre otros.
- Analizar la factibilidad de implementación de métodos basados en sistemas inteligentes para la determinación de la relevancia de los resultados del metabuscador, considerando los datos de los artículos, autores, lugares de publicación y otras entidades involucradas.
- Finalizar la implementación de un método que genere perfiles de los usuarios del metabuscador y los considere para la mejora de la experiencia del usuario con la herramienta.
- Evaluar el desempeño del meta-buscador en relación con soluciones de recuperación de información que operen sobre contextos similares.

4 FORMACIÓN DE RECURSOS HUMANOS

Este proyecto es parte de las líneas de investigación en recuperación y explotación de información del Programa de Investigación en Computación (FCEQyN/UNaM), 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, y un trabajo de investigación posdosctoral. Asimismo, la línea y el equipo de investigación se vinculan con el Grupo de Investigación SMILe de la Universidad de Castilla-La Mancha, España.

5 BIBLIOGRAFÍA

[1] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, “Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses,” *FASEB J.*, vol. 22, no. 2, pp. 338–342, Feb. 2008.

[2] A. N. Guz and J. J. Rushchitsky, “Scopus: A system for the evaluation of scientific journals,” *Int. Appl. Mech.*, vol. 45, no. 4, pp. 351–362, Apr. 2009.

[3] M. Ley, “The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives,” in *String Processing and Information Retrieval*, A. H. F. Laender and A. L. Oliveira, Eds. Springer Berlin Heidelberg, 2002, pp. 1–10.

[4] 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.

[5] C. De Volder, “Los repositorios de acceso abierto en Argentina: situación actual,” *Inf. Cult. Soc.*, no. 19, pp. 79–98, 2008.

[6] 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.

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

[8] *Sistema Nacional de Repositorios Digitales, & Ministerio de Ciencia*. .

[9] H. Kuna, E. Martini, and M. Rey, “Evolution of a Ranking Algorithm for Scientific Documents in the Computer Science Area,” in *XX Argentine Congress of Computer Science Selected Papers*, La Plata, Buenos Aires, Argentina: EDULP, 2015, pp. 145–155.

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

[11] M. Rey *et al.*, “Propuesta de Esquemas de Perfiles para la Recuperación de Datos Científicos para un Sistema de Recuperación de Información del Área de Ciencias de la Computación,” presented at the XXII Congreso Argentino de Ciencias de la Computación, San Luis, Argentina, 2016.

[12] 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.

[1]..... Malietzis, and G. Pappas, “Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses,” *FASEB J.*, vol. 22, no. 2, pp. 338–342, Feb. 2008.

[2].....A system for the evaluation of scientific journals,” *Int. Appl. Mech.*, vol. 45, no. 4, pp. 351–362, Apr. 2009.

[3]..... Bibliography: Evolution, Research Issues, Perspectives,” in *String Processing and Information Retrieval*, A. H. F. Laender and A. L. Oliveira, Eds. Springer Berlin Heidelberg, 2002, pp. 1–10.

[4]..... academic search and Google scholar citations: Comparative analysis of author profiles,” *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 6, pp. 1149–1156, 2014.

[5].....C abierto en Argentina: situación actual,” *Inf. Cult. Soc.*, no. 19, pp. 79–98, 2008.

[6]..... 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.

[7]..... Knowledge from Big Scholar Data,” in *Proceedings of the 25th International*

Conference Companion on World Wide Web, Republic and Canton of Geneva, Switzerland, 2016, pp. 373–373.

[8] *Digitales, & Ministerio de Ciencia*. .

[9] "Evolution of a Ranking Algorithm for Scientific Documents in the Computer Science Area," in *XX Argentine Congress of Computer Science Selected Papers*, La Plata, Buenos Aires, Argentina: EDULP, 2015, pp. 145–155.

[10] and L. Solonezen, "Expansión de Consultas Basada en Ontologías para un Sistema de Recuperación de Información," presented at the XVI Workshop de Investigadores en Ciencias de la Computación, 2014.

[11] Perfiles para la Recuperación de Datos Científicos para un Sistema de Recuperación de Información del Área de Ciencias de la Computación," presented at the XXII Congreso Argentino de Ciencias de la Computación, San Luis, Argentina, 2016.

[12] 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.

[13] *de información en la web*. Editorial de la Universidad Nacional de La Plata (EDULP), 2011.

[14] H. F. Laender, "A Brief Survey of Automatic Methods for Author Name Disambiguation," *SIGMOD Rec*, vol. 41, no. 2, pp. 15–26, Aug. 2012.

[15] fast method based on multiple clustering for name disambiguation in bibliographic citations," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 3, pp. 634–644, Mar. 2015.

[16] and J. Pei, "An Effective Approach to Entity Resolution Problem Using Quasi-clique and Its Application to Digital Libraries," in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2006, pp. 51–52.

[17] "Author Name Disambiguation Using a Graph Model with Node Splitting and Merging Based on Bibliographic Information," *Scientometrics*, vol. 100, no. 1, pp. 15–50, Jul. 2014.

[18] Entity Resolution in Relational Data," *ACM Trans Knowl Discov Data*, vol. 1, no. 1, Mar. 2007.

[19] "Disambiguating Authors in Academic Publications Using Random Forests," in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2009, pp. 39–48.

[20] Liu, "Dynamic Author Name Disambiguation for Growing Digital Libraries," *Inf Retr*, vol. 18, no. 5, pp. 379–412, Oct. 2015.

[21] "Efficient Name Disambiguation for Large-scale Databases," in *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Heidelberg, 2006, pp. 536–544.

[22] and C. L. Giles, "Efficient Topic-based Unsupervised Name Disambiguation," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2007, pp. 342–351.

[23] "Improving Author Coreference by Resource-bounded Information Gathering from the Web," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA, 2007, pp. 429–434.

[24] 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, Apr. 2017.

[25] Giles, "Large scale author name disambiguation in digital libraries," in *2014 IEEE International Conference on Big Data (Big Data)*, 2014, pp. 41–42.

[26] disambiguation in author citations using a K-

way spectral clustering method," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*, 2005, pp. 334–343.
[27]..... author disambiguation," *Inf. Process. Manag.*, vol. 45, no. 1, pp. 84–97, Jan. 2009.
[28]..... Lv, "On Graph-Based Name Disambiguation," *J Data Inf. Qual.*, vol. 2, no. 2, p. 10:1–10:23, Feb. 2011.
[29]..... Giles, "Online Person Name Disambiguation with Constraints," in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2015, pp. 37–46.
[30]..... Disambiguation using Multi-level Graph Partition," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, 2007, pp. 575–580.
[31]..... with the same name," *Scientometrics*, vol. 72, no. 2, pp. 281–290, Jun. 2007.
[32]..... A. H. F. Laender, M. A. Gonçalves, and A. A. Ferreira, "Using Web Information for Author Name Disambiguation," in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2009, pp. 49–58.
[33]..... Gonçalves, and A. H. F. Laender, "Self-training author name disambiguation for information scarce scenarios," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 6, pp. 1257–1278, Jun. 2014.
[34]..... Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations," in *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.*, 2004, pp. 296–305.
[35]..... "A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations," in *Proceedings of the 2005 ACM Symposium on Applied Computing*, New York, NY, USA, 2005, pp. 1065–1069.
[36]..... Dirichlet Model for Unsupervised Entity Resolution," in *Proceedings of the 2006 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, 2006, pp. 47–58.
[37]..... Zhang, "A Unified Probabilistic Framework for Name Disambiguation in Digital Library," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 975–987, Jun. 2012.