# ESTRATEGIAS DE DESAMBIGUACION DE PERFILES Y SIMILITUD TEMÁTICA PARA UN METABUSCADOR DE LAS CIENCIAS DE LA COMPUTACIÓN

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

Instituto de Investigación Desarrollo e Innovación en Informática (IIDII)
Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones.

hdkuna@gmail.com

## RESUMEN

Un metabuscador académico para el área de las Ciencias de la Computación requiere gestionar los resultados de sus búsquedas y brindar sus servicios adecuadamente. En particular, los servicios de recomendación y gestión de resultados requieren el abordaje de problemas de desambiguación de las entidades que recupera, así como recomendación de autores. En este trabajo se presentan las líneas de investigación relacionadas con la evaluación de estrategias para la desambiguación de autores, junto con una línea relacionada con la recomendación de autores en base a los contenidos temáticos de sus perfiles.

Los resultados obtenidos en la evaluación de una estrategia de desambiguación demuestran que se pueden obtener un desempeño equivalente a la referencia. Asimismo, se describe un conjunto de datos en desarrollo para la evaluación de la recomendación de perfiles de autores en base a contenidos temáticos para el área de las Ciencias de la Computación.

**Palabras clave:** desambiguación, perfil, similitud, academico, tópico

## CONTEXTO

Esta línea de investigación se desarrolla en el ámbito Programa de Investigación en Computación (PICom), perteneciente al Instituto de Investigación Desarrollo e Innovación en Informática de la Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones (IIDII, FCEQyN, UNaM).

El PICom desarrolla líneas de investigación relacionadas con la explotación de información y la robótica. Previamente se ha desarrollado un prototipo de metabuscador académico para las Ciencias de la Computación en el que se ha contribuido principalmente en áreas como la generación de estrategias de detección de datos anómalos (outliers e inliers), expansión de consultas, algoritmos de rankings, generación automática de perfiles de entidades, desambiguación de entidades, recomendación de expertos, y selección de grupos de expertos.

## 1. INTRODUCCIÓN

El metabuscador académico para las Ciencias de la Computación que desarrolla y

mantiene este grupo de investigación, actualmente en estado de prototipo, incorpora las contribuciones que se van desarrollando de acuerdo a las líneas de investigación propuestas en el marco del proyecto de investigación y sus actividades asociadas. En el último año, se han abordado problemas relacionados con el mejoramiento de procesos de recomendación, a partir de los resultados arrojados en los procesos de búsqueda, y como consecuencia se ha iniciado el tratamiento del problema de la desambiguación de entidades [1], incluyendo instituciones, autores, lugares de publicación.

La desambiguación de entidades es una tarea específica que involucra la generación de perfiles de dichas entidades, y luego un conjunto de procesos que contribuyan a determinar en qué medida dos o más perfiles con un mismo identificador (nombre) son similares entre sí [1].

Adicionalmente, la similitud de perfiles de de entidades puede tener otras aplicaciones, como la recomendación de contenidos adecuados para los perfiles, por ejemplo para determinar preferencias, o como la recomendación de perfiles, a partir de contenidos específicos. En este último caso, un metabuscador para las Ciencias de la Computación puede ofrecer un servicio de recomendación de perfiles en base a los contenidos temáticos de un documento.

En consonancia con la línea general de trabajo, se decide profundizar en el tratamiento de desambiguación de entidades, para intentar determinar las características adecuadas de una estrategia que permita adaptarse a los servicios brindados por un metabuscador académico para las Ciencias de la Computación.

Adicionalmente, y como consecuencia de los avances en el área de recomendación para el servicio del metabuscador, se identifica la necesidad de ampliar las estrategias de recomendación. En particular, se decide abordar la problemática de la generación de perfiles de autores de las producciones científico-tecnológicas y la recomendación de éstos en base a la similitud de los temas asociados a éste.

## 1.1 DESAMBIGUACION, SIMILITUD TEMÁTICA Y SU IMPACTO EN EL ÁMBITO DEL SISTEMA CIENTÍFICO TECNOLÓGICO NACIONAL

La importancia de la aplicación de estrategias de desambiguación y de similitud temática puede ser apreciada en contextos como el del Sistema Científico Tecnológico Nacional Argentino (SCTNA). En particular, el área de las Ciencias de la Computación incluye un conglomerado de organizaciones que tienen incidencia en la producción científico-tecnológica, tales como eventos de carácter científico, en la que se reciben trabajos que serán presentados durante la duración de los mismos. La evaluación que determina la aceptación o no del trabajo para ser presentado en un evento en particular se lleva a cabo por un conjunto de expertos que pertenecen a la nómina de varias organizaciones dentro del SCTNA.

La generación de perfiles de entidades, en particular de autores pertenecientes a un

sistema científico tecnológico, resulta determinante para permitir identificar en qué temas, áreas o tópicos, un autor define su expertise, y en qué medida lo logra. Asimismo, para propósitos como la recomendación, resulta necesario proporcionar herramientas que permitan desambiguar los perfiles de los expertos, así como el de proveer mecanismos para recomendar perfiles en base a la similitud temática.

## 2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

En este trabajo, presentamos dos derivaciones de la línea de investigación principal referida al metabuscador para el ámbito académico de las Ciencias de la Computación. Estas derivaciones corresponden en primer lugar, a la evaluación de estrategias de desambiguación de autores de publicaciones científicas; y en segundo lugar, al desarrollo de estrategias para determinar similitudes temáticas entre perfiles de autores y documentos asociados a la producción científico tecnológica.

### 2.1 DESAMBIGUACIÓN DE AUTORES

La primer línea de trabajo define como objetivo, la evaluación de un enfoque de desambiguación de entidades, principalmente autores de producciones científico-tecnológicas. En particular, se evalúa la aplicación de estrategias de desambiguación a nombres de autores en el contexto de los resultados ofrecidos por el metabuscador de las Ciencias de la Computación.

Teniendo en cuenta el esquema del metabuscador, las entidades que lo componen y los datos con los que opera, se realizó un relevamiento y análisis de distintos enfoques de desambiguación de nombres de autores [2], [3].

Como resultado del relevamiento y análisis de las alternativas de desambiguación, se optó por la evaluación de AMiner en los resultados del metabuscador. AMiner presenta un framework de desambiguación de nombres de autores que se enfoca en resolver problemas tales como: determinar el número preciso de personas que comparten el mismo nombre; integrar datos continuamente, mejorando el proceso de desambiguación y la inclusión del usuario en el proceso [4].

### 2.2 SIMILITUD TEMÁTICA

La segunda línea de trabajo define como objetivo, la generación de procedimientos que permitan determinar el grado de similitud temática entre perfiles de autores y documentos asociados a producciones científico-tecnológicas. En particular, se pretende aplicar las estrategias de similitud temática [5] para asistir en la búsqueda de expertos adecuados para la evaluación de propuestas de trabajo y de artículos enviados a su publicación en eventos y revistas.

Esta línea de investigación abarca la identificación de estrategias del Procesamiento de Lenguaje Natural y del Aprendizaje Automático, destinadas al desarrollo de procesos que permitan elaborar perfiles de expertos a partir de artículos

académicos y otros contenidos textuales. Asimismo, se pretende desarrollar alternativas que permitan la evaluación de la similitud entre los perfiles de expertos y trabajos académicos del área de las Ciencias de la Computación.

## 3. RESULTADOS Y OBJETIVOS

La evaluación de la estrategia de desambiguación involucró la definición de un conjunto de datos generados a partir de consultas realizadas al metabuscador académico de Ciencias de la Computación. El conjunto de datos utilizados para la evaluación se constituyó a partir de 30 nombres de autores, de los cuales 18 (60%) elementos de la lista fueron tomados para entrenamiento y 12 (40%) para validación. Cada elemento de la lista contiene uno o más autores con el mismo nombre. Además, el conjunto de datos contiene los documentos de cada uno de esos autores, sumando en total 4.317 artículos.

Se utilizaron las métricas Precision, Recall y F1-score para evaluar los resultados del proceso de desambiguación. El experimento arrojó los siguientes resultados promedio. Precision: 0,7359; Recall: 0,60407; F1-score: 0,6635.

A modo de referencia, la ejecución del experimento sobre un conjunto de datos generados por AMiner [1] arrojó los siguientes resultados. Precision: 0,7685; Recall: 0,61661; y F1-score: 0,68423.

De acuerdo a los resultados obtenidos en la evaluación de la estrategia de desambiguación, se puede observar que la aplicación de dicha estrategia a distintos conjuntos de datos genera resultados similares. Por lo tanto, el método de desambiguación podría ser adecuado para su aplicación en el metabuscador.

En forma paralela a la evaluación de la estrategia de desambiguación de entidades, se comenzó a trabajar con la elaboración de un conjunto de datos que permita la evaluación exploratoria de estrategias para la generación de perfiles de expertos a partir del contenido temático de sus producciones científico-tecnológicas. Asimismo, se comenzó con el relevamiento de estrategias para determinar similitudes en base a dichas representaciones de los perfiles.
Se logró elaborar un conjunto de datos con 27.812 registros de producciones científico-tecnológicas en idioma español publicadas en el repositorio de la Universidad Nacional de La Plata (SEDICI). Estas producciones se corresponden principalmente con publicaciones en eventos, tesis, trabajos finales de grado y posgrado y artículos en revistas del ámbito académico de las Ciencias de la Computación de la República Argentina.

La evaluación exploratoria del enfoque para la generación de perfiles se realizó sobre un conjunto de 22.380 autores a partir de los registros contenidos en el conjunto de datos. Esta evaluación incluyó la generación de perfiles de expertos utilizando la técnica de word embeddings [6] a partir de atributos de

[1] https://github.com/neozhangthe1/disambiguation

los registros contenidos en el conjunto de datos, incluyendo: título y resumen [7]. Asimismo, se generaron conjuntos de tópicos mediante Latent Dirichlet Allocation (LDA) [8].

Sobre los resultados de la evaluación experimental, se han evaluado métricas de similitud (ej: Jensen-Shannon), los cuales aún deben ser contrastados con la opinión de expertos, debido a la falta de un conjunto de datos que permita contrastar los resultados en forma automática.

## 4. FORMACIÓN DE RECURSOS HUMANOS

Las líneas de investigación presentadas cuentan con doce integrantes relacionados con las carreras de Ciencias de la Computación de la UNaM. El grupo de investigación desarrolla dos tesis de grado articulando sus trabajos con becas de Estímulo a las Vocaciones Científicas del Consejo InterUniversitario Nacional (CIN) y becas UNaM; dos tesis de maestría en curso y una finalizada, de las cuales dos de ellas enmarcadas en becas del Programa Estratégico de Formación de Recursos Humanos en Investigación y Desarrollo (PERHID) del CIN. Asimismo, las líneas de investigación y sus integrantes se vinculan con grupos de la Universidad de Castilla-La Mancha, España y la Universidad de Sonora, México.

## 5. BIBLIOGRAFÍA


[1] I. Bhattacharya and L. Getoor, "Collective Entity Resolution in Relational Data," *ACM Trans Knowl Discov Data*, vol. 1, no. 1, Mar. 2007.

[2] A. Canteros, E. Zamudio, and H. Kuna, "Desambiguación de autores para un sistema de recuperación de expertos en un contexto académico," Simposio Argentino de Inteligencia Artificial, Sep. 2018, pp. 54–57.

[3] N. R. Smalheiser and V. I. Torvik, "Author name disambiguation," *Annu. Rev. Inf. Sci. Technol.*, vol. 43, no. 1, pp. 1–43, Jan. 2009.

[4] Y. Zhang, F. Zhang, P. Yao, and J. Tang, "Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop.," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2018, pp. 1002–1011.

[5] O. Hanif, Z. Donghua, W. Xuefeng, and M. S. Nawaz, "Refining the Measurement of Topic Similarities Through Bibliographic Coupling and LDA," *IEEE Access*, vol. 7, pp. 179997–180011, 2019.

[6] M. Liu, B. Lang, Z. Gu, and A. Zeeshan, "Measuring similarity of academic articles with semantic profile and joint word embedding," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 619–632, 2017.

[7] B. Zhang and M. Al Hasan, "Name Disambiguation in Anonymized Graphs using Network Embedding," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Nov. 2017, pp. 1239–1248.

[8] M. Amami, G. Pasi, F. Stella, and R. Faiz, "An lda-based approach to scientific paper recommendation," in *International conference on applications of natural language to information systems*, 2016, pp. 200–210.