Desambiguación de autores para un sistema de recuperación de expertos en un contexto académico

Canteros, A. V.; Zamudio, E.; Kuna, H. D.

Programa de Investigación en Computación
Facultad de Ciencias Exactas, Químicas y Naturales
Universidad Nacional de Misiones

{anacanteros, eduardozamudio}@fceqyn.unam.edu.ar,
hdkuna@gmail.com

Abstract. La identificación de un autor y sus obras es un componente importante en los sistemas que administran información de publicaciones científicas. Dicha tarea es determinante para identificar expertos y conocer sus obras. El presente trabajo es una investigación para el relevamiento y desarrollo de estrategias de desambiguación de autores para un sistema de recuperación de información del área de las Ciencias de la Computación.

Keywords: desambiguación de autores, recuperación de expertos, producción científica.

1 Introducción

La recuperación de expertos es el área de recuperación de información que se encarga de determinar las áreas en las que una persona es experta, y a su vez determinar las personas expertas en un área determinada.

En la recuperación de expertos en contextos científicos, se desea conocer a las personas que son autores de las publicaciones y si ha publicado otras obras.

En términos computacionales, la identificación unívoca de personas autoras de dichas publicaciones resulta dificultosa, ya que implica analizar un conjunto de datos estructurados o no estructurados en una base de datos y aplicar métodos para encontrar similitudes entre sus registros que establezcan una referencia a una entidad del mundo real [1, 2]. Esto se conoce como resolución de entidades [3].

La tarea que se ocupa de la identificación unívoca de los autores se conoce como desambiguación de autores. Previamente, se han desarrollado diversos métodos para tratar esta dificultad. Estos métodos tratan en forma resumida los siguientes desafíos [4, 5]:

- Un autor puede publicar bajo distintos nombres: ya sea por errores ortográficos, uso de seudónimos u otros cambios.
- Varios autores usan nombres similares. En tales casos es necesario el análisis de sus metadatos para conocer a qué individuo corresponde.

- Problemas de metadatos: en algunos casos la información adicional sobre un autor, además del nombre, es insuficiente o presenta inconsistencias a la hora de identificar a la persona.

Estos desafíos de la desambiguación de autores han llevado al desarrollo de varios métodos. Ferreira et al. [6] presenta una taxonomía jerárquica caracterizando los distintos métodos, y haciendo referencia a aquellos que son más representativos.

Dicha taxonomía clasifica a los métodos de desambiguación según el enfoque que utilizan para su resolución en dos categorías: agrupamiento de autores y asignación. Los métodos de agrupamiento explotan las similitudes entre las referencias a un autor y las agrupan mediante alguna técnica de clustering. En cambio, los métodos de asignación toman las referencias de un autor y se las asignan a dicho autor.

A su vez los métodos de desambiguación pueden ser clasificados según el tipo de información que utilizan en: métodos que utilizan solamente información en las citas, los que utilizan información de la web; y los que utilizan evidencia implícita que puede ser obtenida a partir de la información disponible.

A partir del crecimiento continuo de las bases de datos académicas con información de autores, se ha detectado la importancia de desarrollar procesos de desambiguación de autores que consideren el carácter incremental de las bases de datos [7]. Esto implica que la desambiguación debe tomar provecho de los datos existentes en las bases de datos de los sistemas de recuperación de información, y a su vez, considerar información actualizada de los posibles autores desde otros contextos, particularmente de los recursos en línea como redes sociales y publicaciones en línea.

Mediante el desarrollo de este trabajo, se buscará la aplicación de distintas técnicas de búsqueda, recuperación y explotación de información que aporte a mejorar los procesos de desambiguación de autores de publicaciones científicas en particular, y de entidades en general.

1.1 Un metabuscador para el área de las ciencias de la computación

En los últimos años se han producido avances en el desarrollo de un Sistema de Recuperación de Información (SRI) de artículos científicos de las Ciencias de la Computación [8]. Dicho sistema consiste en un metabuscador, en el cual se han implementado módulos que permiten procesar las búsquedas de un usuario, acceder a distintas librerías digitales, recuperar datos de los documentos científicos, sus autores y los lugares donde fueron publicados, entre otros. Asimismo, se desarrollaron módulos que permiten procesar dichos datos para su presentación al usuario [9].

Uno de los módulos que destacan dentro del metabuscador es un algoritmo de ranking [10]. El mismo se encarga de procesar distintas métricas de calidad de unidades de producción científica (UPC) para ordenar los resultados que se le presentarán al usuario. Las métricas mencionadas evalúan cada UPC desde tres propiedades o dimensiones diferentes: la calidad de la fuente de publicación, la calidad del documento en sí y la calidad de sus autores.

En el metabuscador, la ambigüedad del nombre de un autor puede llevar a una mala precisión del algoritmo de ranking. El algoritmo de ranking arroja resultados de

acuerdo a las propiedades del autor. Si el autor no ha sido identificado unívocamente, el orden de relevancia de los resultados no será el correspondiente.

2 Propuesta y tareas en progreso

Este trabajo propone analizar los distintos métodos de desambiguación actuales, identificando sus características en relación a los datos que utilizan, los requisitos para su implementación, las herramientas para su evaluación y su impacto en los sistemas de recuperación de información.

Posteriormente, se propone diseñar e implementar y evaluar una estrategia de desambiguación de autores para un sistema de recuperación de información del ámbito académico actual, del área de las Ciencias de la Computación.

Tras haber llevado a cabo un relevamiento y análisis de varios métodos de desambiguación de nombres de autores, se han detectado métodos de desambiguación de autores que consideran otros aspectos clave para solucionar el problema:

- Santana et al. [7]: propone un método que utiliza técnicas de clustering para realizar el proceso de desambiguación en forma incremental. Toma en cuenta otros aspectos importantes como el cambio en los perfiles de autores o casos de autores nuevos que poseen pocas referencias para ser analizadas.
- Zhu et al. [11]: propone un framework multicapa dinámico que maneja distintos tipos de datos. Dicho framework es lo suficientemente flexible como para añadir o quitar capas basado en los requerimientos de la aplicación. Cada capa aplica una técnica de clustering sobre un tipo de datos específico y se detectan y corrigen los errores producidos.
- Liu et al. [12]: propone un framework de clustering de tres capas que consiste en formar fragmentos de autores. En cada capa se van agrupando dichos fragmentos de acuerdo a sus similitudes y reduciendo su número. El resultado final serían fragmentos de autores diferentes. Utiliza pocos datos para el proceso y tiene bajo costo computacional.
- Tang et al. [13]: utiliza un método probabilístico para asignar referencias a autores. Es un método implementado para un repositorio académico digital de ciencias de la computación que está en uso actualmente. Además se puede utilizar de forma incremental.

3 Consideraciones finales

Para la fase de experimentación, se buscará reproducir los métodos de desambiguación de autores mencionados en la sección anterior y hacer una comparativa entre los mismos. En base a los resultados se determinará cuáles de estos métodos resultará más adecuado para su implementación el metabuscador.

4 Referencias

1. Bhattacharya, I., Getoor, L.: Collective Entity Resolution in Relational Data. ACM Trans Knowl Discov Data. 1, (2007).
2. Fan, W., Jia, X., Li, J., Ma, S.: Reasoning About Record Matching Rules. Proc VLDB Endow. 2, 407–418 (2009).
3. Getoor, L., Machanavajjhala, A.: Entity Resolution: Theory, Practice & Open Challenges. Proc VLDB Endow. 5, 2018–2019 (2012).
4. Smalheiser, N.R., Torvik, V.I.: Author name disambiguation. Annu. Rev. Inf. Sci. Technol. 43, 1–43 (2009).
5. Diaz-Valenzuela, I., Martin-Bautista, M.J., Vila, M.-A., Campaña, J.R.: An automatic system for identifying authorities in digital libraries. Expert Syst. Appl. 40, 3994–4002 (2013).
6. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: A Brief Survey of Automatic Methods for Author Name Disambiguation. SIGMOD Rec. 41, 15–26 (2012).
7. Santana, A.F., Gonçalves, M.A., Laender, A.H.F., Ferreira, A.A.: Incremental Author Name Disambiguation by Exploiting Domain-specific Heuristics. J Assoc Inf Sci Technol. 68, 931–945 (2017).
8. Kuna, H., Martin, R., Martini, E., Solonezen, L.: Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación. Rev. Latinoam. Ing. Softw. 2, 107 (2014).
9. Kuna, H.D., Rey, M., Martini, E., Canteros, A., Rambo, A.R., Biale, C.O., Zamudio, E.: Avances en la construcción de un Sistema de Recuperación de Información para información científica en Ciencias de la Computación. Presented at the XVIII Workshop de Investigadores en Ciencias de la Computación (Entre Ríos, Argentina) (2016).
10. Kuna, H.D., Martini, E., Rey, M.: Evolución de un algoritmo de ranking para documentos científicos del área de las ciencias de la computación. Presented at the XX Congreso Argentino de Ciencias de la Computación (Buenos Aires, Argentina) (2014).
11. Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G.P., Tang, Y.: A Novel Multiple Layers Name Disambiguation Framework for Digital Libraries Using Dynamic Clustering. Scientometrics. 114, 781–794 (2018).
12. Liu, Y., Li, W., Huang, Z., Fang, Q.: A fast method based on multiple clustering for name disambiguation in bibliographic citations. J. Assoc. Inf. Sci. Technol. 66, 634–644 (2015).
13. Tang, J., Fong, A.C.M., Wang, B., Zhang, J.: A Unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE Trans. Knowl. Data Eng. 24, 975–987 (2012).