Evaluación del algoritmo de desambiguación de autores de AMiner en un metabuscador académico de Ciencias de la Computación

(Evaluation of the author disambiguation algorithm of AMiner in an academic metasearcher of Computer Science)

Ana Canteros,[1] Eduardo Zamudio[2] & Horacio Daniel Kuna[3]

Campo temático: Machine Learning.

Resumen

La desambiguación de autores es un problema de gran relevancia para los sistemas de recuperación de información del ámbito académico. El algoritmo de desambiguación de nombres de AMiner representa uno de los enfoques basados en Aprendizaje Automático con mayor impacto en la actualidad. En este trabajo, se presenta una evaluación del algoritmo de desambiguación de nombres de AMiner para la desambiguación de autores en el contexto de un metabuscador académico del área de las Ciencias de la Computación. Los resultados experimentales con datos generados por el metabuscador académico dan cuenta de un desempeño promedio similar a la referencia. Asimismo, las evaluaciones experimentales permitieron identificar casos especiales de nombres de autores en el que el algoritmo presenta un bajo desempeño en comparación con el promedio. Este hallazgo permitió identificar una asociación aparente entre el bajo desempeño del algoritmo en contextos de varios autores con un mismo nombre y con pocas publicaciones.

Palabras clave: desambiguación de autores, metabuscador académico, resolución de entidades, datos bibliográficos

[1] Universidad Nacional de Misiones. anacanteros@fceqyn.unam.edu.ar
[2] Universidad Nacional de Misiones. eduardozamudio@fceqyn.unam.edu.ar
[3] Universidad Nacional de Misiones. hdkuna@gmail.com

Abstract

Author disambiguation is a problem of considerable relevance for information retrieval systems in the academic field. The AMiner name disambiguation algorithm represents one of the approaches based on Machine Learning with greater impact in the present. In this work, an evaluation of the AMiner name disambiguation algorithm is presented for author disambiguation in the context of a Computer Science academic metasearcher. The experimental results with data generated by the academic metasearcher warn about an average performance similar to the reference. Likewise, the experimental evaluations allow the identification of special cases of author names where the algorithm presents low performance when is compared to the average. This finding allowed to identify an apparent association between the low performance of the algorithm in contexts of several authors with the same name and with few publications.

Keywords: author disambiguation, academic metasearcher, entity resolution, bibliographic data

1. Introducción

La desambiguación de autores representa uno de los componentes más importantes de un sistema de recuperación de información (SRI) del ámbito de las producciones académicas.

La desambiguación de autores es una subtarea del área de la integración de datos de un SRI académico, la cual se encarga de identificar en forma unívoca a los autores de producciones. Esta es una tarea particularmente desafiante en el contexto de la producción global de contribuciones académicas de distintas áreas de conocimiento.

La problemática de la desambiguación de autores en contextos académicos es particularmente relevante en la actualidad. Un estudio comprehensivo es realizado por Shoaib et. al. (Shoaib et al., 2020), el cual plantea una generalización del problema y describe un marco de trabajo genérico para el abordaje del mismo. En términos generales, el problema de la desambiguación de nombres de autores implica resolver casos de sinonimia y polisemia. Polisemia, en el que un mismo nombre corresponde a varios autores. Y sinonimia, en el que un autor es referido a partir de varios nombres. Asimismo, ambos problemas se pueden presentar en forma conjunta.

El Aprendizaje Automático representa una de las áreas más influyentes mediante las cuales se aborda el problema de la desambiguación de nombres de autores en contextos académicos. En este sentido, (Wang et al., 2020) presentan una estrategia que combina modelos discriminativos y generativos junto con el uso de redes de información heterogéneas. Asimismo, (W. Zhang et al., 2019) implementan un método basado en redes de información, en el que utilizan únicamente información de colaboración entre autores, con objeto de reducir la dimensionalidad de la representación.

AMiner (Wan et al., 2019) es un complejo sistema que involucra recuperación de datos de diversas fuentes en la web, para brindar una amplia variedad de servicios relacionados con la información académica.

La gran cantidad de datos utilizados por AMiner (Jie Tang, 2016b, 2016a) requiere de una arquitectura adecuada que permita extraer, integrar, almacenar y acceder a un conjunto de datos que puedan ser utilizados en modelos requeridos para brindar los servicios del sistema.

La integración de datos en AMiner implementa una estrategia de desambiguación de nombres (Y. Zhang et al., 2018), como uno de sus componentes principales. Esta estrategia es presentada como un método de aprendizaje de representación, con resultados superadores a otros métodos de referencia en términos de eficacia y eficiencia.

Actualmente, se dispone de un SRI de producciones científico tecnológicas del área de las Ciencias de la Computación (Kuna et al., 2019). El mismo consiste en un metabuscador, el cual utiliza datos generados por los servicios de recuperación de información de buscadores académicos, algunos específicos del área de

conocimiento mencionada como ACM e IEEE, y otros más generales como Google Scholar y Microsoft Academics.

El desarrollo de este Metabuscador Académico de las Ciencias de la Computación (MACC) permitió contribuir al estudio de problemas específicos de la recuperación de información como la generación de algoritmos de ranking de resultados y la expansión de consultas. Asimismo, se han desarrollado estrategias de generación de perfiles de entidades (Kuna et al., 2017), las cuales requieren la implementación de métodos adecuados para la resolución de dichas entidades. En particular, una tarea específica de la resolución de entidades requerida en la implementación del metabuscador es la de la desambiguación de autores.

En este trabajo, se presenta una evaluación del método de desambiguación de nombres desarrollado en AMiner para su implementación en el MACC.

El objetivo general del mismo es determinar la adecuación del algoritmo de desambiguación de nombres de autores en el contexto del MACC.

Este documento se organiza de la siguiente manera. En la sección 2 se presenta una descripción de distintas alternativas representativas al problema de la desambiguación de nombres de autores en contextos académicos. Asimismo, se describen los aspectos principales del algoritmo de desambiguación de nombres de autores de AMiner. En la sección 3 se presenta una descripción del contexto experimental para la evaluación del algoritmo en los datos del MACC. En la sección 4 se describen los resultados experimentales y en la sección 5 se presentan las discusiones en base a dichos resultados. Finalmente, en la sección 6 se presentan las conclusiones del trabajo.

2. Alternativas y antecedentes de AMiner

En esta sección se presenta el relevamiento de métodos de desambiguación de nombres de autores seleccionados de acuerdo a sus características distintivas. Entre ellas se menciona la adaptabilidad a distintos escenarios y cambios en los perfiles o información de autores, datos que utilizan en el proceso de desambiguación, coste computacional y corrección de errores durante el proceso.

Existe una amplia variedad de enfoques que se dedican a resolver esta problemática. En (Ferreira et al., 2012) se presenta una clasificación de los métodos más representativos.

A partir de dicha clasificación se hizo un relevamiento de métodos de desambiguación más actuales, y se estudiaron sus características, implementación y efectividad de solución. Entre esos métodos se consideraron algunos para mayor análisis:

- (Santana et al., 2017): su enfoque hace énfasis en tratar el problema de incrementación de datos y cómo ello afecta al proceso de desambiguación

en sí.

- (Zhu et al., 2018): proponen un framework dinámico multicapa. Cada capa corresponde a un atributo de un documento académico tales como título, co-autores, lugares de publicación, entre otros. El método posee adaptabilidad a distintos escenarios considerando que el modelo propuesto permite añadir o sustraer capas, con los atributos que mejor se adapten a cada situación.
- (Liu et al., 2015): es un método de aprendizaje no supervisado que se enfoca en un proceso de desambiguación que utilice la menor cantidad de información posible y con bajo costo computacional.
- (J. Tang et al., 2012): es una propuesta que hace varios años fue implementada en un repositorio académico digital, ArnetMiner.

Es destacable el método presentado en (J. Tang et al., 2012). Es un método propuesto hace años que fue utilizado en ArnetMiner, ahora conocido como AMiner. Se trata de un modelo probabilístico basado en Campos Ocultos Aleatorios de Markov que captura dependencias entre publicaciones científicas.

Sin embargo, en los últimos años dicho modelo se fue dejando atrás debido a su ineficacia en escenarios con un volumen masivo de datos. En cambio, AMiner propone un framework de desambiguación que incorpora aprendizajes tanto a nivel global como a nivel local. Además, presenta un método para estimar el número de clusters. El código fuente del framework está disponible para su experimentación con datos de prueba propios de AMiner.

2.1 El framework de desambiguación de autores de AMiner

El framework de desambiguación consiste en una primera etapa de aprendizaje global, seguido de aprendizaje a nivel local.

El modelo de aprendizaje global consiste en tomar todo el conjunto de artículos y analizar sus atributos. Se llama atributo a cada característica que posee el artículo, información que es útil en el proceso de desambiguación. Por ejemplo, nombre, abstract, coautores, año, lugar de publicación, entre otros. A partir de los atributos del conjunto de documentos en su totalidad se toma la información de sus atributos y se forma un espacio de embeddings, con el objetivo de agrupar documentos en base a sus similitudes. Dicho agrupamiento se realiza en forma de pares positivos, si presentan un alto grado de similitud o pares negativos si el grado de similitud es baja. Además, el framework introduce triples en el proceso de desambiguación. Un triple consiste en un documento base, su par positivo y su par negativo. La inclusión de triples en esta etapa permite que los documentos pertenecientes al mismo autor se mantengan a una distancia cercana, mientras que las publicaciones de distintos autores deben mantenerse lo más alejados posible (Y. Zhang et al., 2018).

Por otro lado, el modelo de aprendizaje local toma los resultados del aprendizaje global con el objetivo de refinar el proceso o corregir posibles errores. En este caso ya no se toma todo el conjunto de documentos, sino que se toma un conjunto de documentos por nombre de autor. Se construye un grafo con el objetivo de enlazar los documentos con el mayor grado de similitud entre sí. En esta etapa se aplica un algoritmo de clustering aglomerativo jerárquico, formándose grupos de documentos pertenecientes a distintos autores que comparten el mismo nombre.

Además de los nos modelos de aprendizaje mencionados, el framework presenta un método para estimar el número de clusters por nombre de autor, basado en redes neuronales recurrentes.

3. Experimentación

Se realizaron pruebas con un conjunto de datos obtenidos del MACC. La base de datos del MACC almacena tanto consultas como resultados de dichas consultas. Entre esos resultados figuran nombres de autores con sus publicaciones.

Como se menciona en la sección anterior, el código fuente del algoritmo de desambiguación de AMiner se encuentra disponible para realizar pruebas, junto con un conjunto de datos propio [4]. El mismo consiste en un archivo en formato JSON con datos de publicaciones académicas, una lista de autores, y una lista que asocia publicaciones a sus respectivos autores para validar los resultados de las pruebas.

Los datos del MACC cuentan con información suficiente como para realizar pruebas con el framework de desambiguación. Sin embargo, el formato de dichos datos debe ser modificado para adaptarse al que necesita el framework.

En primer lugar, del conjunto de datos del MACC se toma un conjunto de nombres de autores que presenten ambigüedades. Se tomaron aquellos nombres que fueran compartidos por dos o más autores, y las publicaciones de dichos autores. En total se tomaron 40 nombres de autores y sus publicaciones, que suman 4317 documentos.

Luego se obtiene un archivo en formato JSON con todos los documentos y sus atributos, el cual es utilizado en la etapa de aprendizaje global y se forma el espacio de *embeddings. También se elaboran las listas que asocian las publicaciones a sus autores. Existen dos listas en este formato, dado que una es utilizada para el entrenamiento y otra para las pruebas. Se toman los datos de 28 autores y sus publicaciones para el entrenamiento, correspondiente al 70 % del total. El 30 % restante es utilizado en las pruebas.*

4. Resultados

El Cuadro 1 presenta los resultados de la experimentación. Los datos que se

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

muestran en la misma son: nombre de autor; número total de publicaciones; número de clusters determinado por el algoritmo; seguido del verdadero número de clusters o cantidad real de individuos que comparten el mismo nombre.

Para medir la efectividad del algoritmo se utilizan las métricas *precision, recall y f1-score. Precision es la proporción de documentos relevantes sobre el total de documentos recuperados. Mientras que recall es el número de documentos relevantes recuperados sobre el total de documentos relevantes en todo el conjunto de datos (Van Rijsbergen, 1979).*

El f1-score es una métrica que balancea los puntajes obtenidos por precision y recall. Dicho de otra manera, es la media armónica entre ambas métricas (Sasaki, 2007).

Dichas métricas se calculan al finalizar la etapa de aprendizaje local. El framework forma clusters que corresponden a distintos autores que comparten el mismo nombre y asigna las publicaciones a dichos clusters. Luego se evalúa si las asignaciones fueron correctas.

Cuadro 1. Resultados de desambiguación en el conjunto de datos propio (Elaboración propia).

Nombre	n_pubs	n_clusters	true_n_clusters	precision	recall	f1
Ananth Grama	111	2	2	0.78913	0.54037	0.64148
Satchidananda Dehuri	64	2	2	0.65625	0.69789	0.67643
Joseph Sifakis	179	2	2	0.73670	0.55144	0.63075
Joydeep Ghosh	127	2	2	0.60522	0.51292	0.55526
Paolo Rosso	162	2	2	0.89760	0.59294	0.71413
Yehuda Koren	62	2	2	0.79330	0.54791	0.64816
M. K. Tiwari	83	2	2	0.65106	0.54687	0.59443
Narayanaswamy Balakrishnan	93	2	2	0.81827	0.49858	0.61962
Jiebo Luo	207	2	3	0.50329	0.99142	0.66765
Chang Liu	171	4	11	0.59364	0.36452	0.45169
Lidan Wang	45	2	4	0.90767	0.87710	0.89212
C. Mohan	78	2	9	0.87863	0.52691	0.65877
Promedio				**0.7359**	**0.60407**	**0.6635**

Al observar la tercer y cuarta columna, se puede ver que la cantidad de clusters estimados por el algoritmo coincide con la cantidad real de clusters. En cambio, en las últimas cuatro filas, se pueden observar diferencias entre la cantidad de clusters estimados por el framework de desambiguación y la cantidad real de clusters.

En esos casos, lo que ocurrió es que, en un punto del proceso de desambiguación, cuando se preparan los datos para el entrenamiento a nivel local, se descartan autores con menos de 5 publicaciones.

La última fila del Cuadro 1 presenta el promedio global de precision, recall y f1-score para el conjunto de datos de prueba del MACC.

En el Cuadro 2 se describen los promedios de precision, recall y f1-score del framework de desambiguación aplicado en el conjunto de datos de AMiner y en el del MACC.

Cuadro 2. Resultados de desambiguación en el conjunto de datos propio y el de referencia (Elaboración propia).

Dataset	Precision	Recall	f1-score
MACC	0.7359	0.60407	0.66350
AMiner	0.7685	0.61661	0.68423
Diferencia	0.0326	0.01254	0.02073

Según las métricas observadas, los resultados del algoritmo aplicado sobre el conjunto de datos de AMiner son mayores en promedio que los resultados de MACC. La diferencia entre ambas operaciones es de 0.0326, 0.01254 y 0.02073 en Precision, Recall y F1 score respectivamente.

5.Discusión

En el Cuadro 2 se comparan los resultados del framework de desambiguación aplicados a dos conjuntos de datos diferentes. Uno provisto por AMiner y otro basado en el MACC. Los promedios globales en ambos escenarios muestran que las tres métricas presentan una diferencia menor a 0.05, lo cual resulta poco significativo.

Se puede decir, por lo tanto, que este enfoque de desambiguación es adaptable al MACC.

El Framework de desambiguación posee varias ventajas: entre ellas se puede mencionar el formato de los datos que utiliza para el procesamiento. Si bien fue necesario transformar los datos del metabuscador para que se adapten al algoritmo esto se hizo en la estructura y orden de los datos y no hubo pérdida de información.

Otra ventaja a mencionar está relacionada con la información que necesita el framework para su procesamiento. Existen atributos que son obligatorios para el modelo de aprendizaje como título y coautores, mientras que otros atributos se pueden omitir o añadir. Esta es una de las características que indica que el framework es flexible y adaptable a otros contextos.

Por otro lado, se pueden observar ciertas desventajas. Estimar el número correcto de clusters continúa siendo un desafío, a pesar de que este método fue diseñado para procesar grandes cantidades de datos.

Además, se hallaron ciertas arbitrariedades en el código fuente del algoritmo. Entre ellas se destaca el hecho de que se descartan los autores con menos de cinco publicaciones. Ello afecta significativamente el proceso de desambiguación, debido a que existen datos que no se incluyen en el proceso de clustering.

El algoritmo utilizado para la experimentación incluye un módulo para la estimación de cantidad de clusters por nombre de autor, basado en los resultados del aprendizaje global y local. Sin embargo, los resultados de la ejecución del módulo no se integran con los resultados de la etapa anterior. Esto se suma a las dificultades halladas durante la experimentación sobre si es posible mejorar la precisión del algoritmo.

6. Conclusiones

Este trabajo constituye un estudio sobre enfoques de desambiguación de autores para su aplicación en un Metabuscador Académico de las Ciencias de la Computación (MACC).

Mediante este estudio se hallaron distintas propuestas que tratan de resolver la problemática de ambigüedad en nombres de autores. Asimismo, se analizaron sus características, tales como la información que utilizan para el proceso de desambiguación, aspectos a tener en cuenta a la hora de diseñar una solución eficaz y su adaptabilidad a otros contextos.

La evaluación el enfoque de desambiguación de AMiner en el contexto del MACC permitió la realización de pruebas con datos propios para evaluar su efectividad aplicado a dicho sistema.

El enfoque seleccionado posee ciertas características útiles para su aplicación en contextos diferentes como lo es en el MACC. Entre dichas características se puede mencionar el formato de datos que utiliza y el hecho que el desempeño del algoritmo con un conjunto de datos basado en el MACC es similar a la referencia.

Sin embargo, la evaluación del algoritmo con datos del MACC presenta a una asociación aparente entre el bajo desempeño del algoritmo en contextos de varios autores con un mismo nombre y con pocas publicaciones.

Finalmente, estas características deberían ser estudiadas en mayor profundidad y en contextos de mayor heterogeneidad de datos.

Trabajos futuros en esta línea apuntan a extender las condiciones de evaluación del algoritmo de desambiguación de AMiner en conjuntos de datos más amplios. Asimismo, se prevé la adaptación del algoritmo para adecuarlo a condiciones de autores con un número reducido de publicaciones.

Referencias

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Rec., 41(2), 15–26. https://doi.org/10.1145/2350036.2350040*

Kuna, H., Cantero, A., Canteros, A., Rey, M., Zamudio, E., Rambo, A., Martini, E., Pautsch, G., Biale, C., Krujoski, S., & Rauber, F. (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, 198-202. http://www.wicc2019.unsj.edu.ar/descargas/Libro_WICC2019.pdf

Kuna, H., Rey, M., Zamudio, E., Olivas, J. A., Rambo, A., Cantero, A., Canteros, A., Martini, E., & Biale, C. (2017). An Entity Profile Schema for Data Integration in an Academic Metasearch Engine. *Proceedings of the 2017 International Conference on Artificial Intelligence*, 281–285. http://csce.ucmss.com/cr/books/2017/ConferenceReport?ConferenceKey=ICA

Liu, Y., Li, W., Huang, Z., & Fang, Q. (2015). A fast method based on multiple clustering for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology*, *66*(3), 634-644. https://doi.org/10.1002/asi.23183

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2017). Incremental Author Name Disambiguation by Exploiting Domain-specific Heuristics. *J. Assoc. Inf. Sci. Technol.*, *68*(4), 931–945. https://doi.org/10.1002/asi.23726

Shoaib, M., Daud, A., & Amjad, T. (2020). Author Name Disambiguation in Bibliographic Databases: A Survey. *arXiv preprint arXiv:2004.06391*.

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2012). A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 975-987. https://doi.org/10.1109/TKDE.2011.13

Tang, Jie. (2016a). AMiner: Mining deep knowledge from big scholar data. *Proceedings of the 25th international conference companion on world wide web*, 373–373.

Tang, Jie. (2016b). AMiner: Toward understanding big scholar data. *Proceedings of the ninth ACM international conference on web search and data mining*, 467–467.

Wan, H., Zhang, Y., Zhang, J., & Tang, J. (2019). Aminer: Search and mining of academic social networks. *Data Intelligence*, *1*(1), 58–76.

Wang, H., Wang, R., Wen, C., Li, S., Jia, Y., Zhang, W., & Wang, X. (2020). Author Name Disambiguation on Heterogeneous Information Network with Adversarial Representation Learning. *arXiv preprint arXiv:2002.09803*.

Zhang, W., Yan, Z., & Zheng, Y. (2019). Author Name Disambiguation Using Graph Node Embedding Method. *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 410–415.

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

Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G. P., & Tang, Y. (2018). A Novel Multiple Layers Name Disambiguation Framework for Digital Libraries Using Dynamic Clustering. *Scientometrics*, *114*(3), 781–794. https://doi.org/10.1007/s11192-017-2611-8

Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*.

Van Rijsbergen, C. (1979). *Information Retrieval | Guide books*. https://dl.acm.org/doi/book/10.5555/539927