Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos

Clustering aglomerativo para la desambiguación de nombres de autor

Trabajo Fin de Máster

Tipo de trabajo: Comparativa de soluciones

Presentado por: Padilla Mesa, Juan Antonio

Director/a: Lancheros Cuesta, Diana Janeth

Co-director/a: López-Herrera, Antonio Gabriel

Resumen

La ambigüedad de los nombres de autor en las bases de datos bibliográficas supone un gran problema a la hora de la recuperación de información, la realización de análisis bibliométricos y la visibilidad de la ciencia. Esto se debe a los fenómenos denominados ruido y silencio documental. Algunos autores han propuesto técnicas para paliar este problema. El objetivo principal de este trabajo es comparar la propuesta original basada en clustering aglomerativo realizada por Schulz et al. (2014), con una propuesta ampliada a partir de la original y determinar si esta propuesta ampliada ofrece mejores resultados. Para la consecución de los objetivos se ha diseñado un dataset compuesto por 60.663 registros del área de “Information Science & Library Science” provenientes de Web of Science, sobre el que se han aplicado ambas propuestas. Analizando los resultados sobre 5 autores distintos con nombres ambiguos, se ha observado que, aunque la propuesta ampliada mejore levemente los resultados, ambas propuestas ofrecen similares resultados para la resolución de este problema.

Palabras Clave: desambiguación de términos, clustering, recuperación de información, machine learning, Web of Science

Abstract

Author's name ambiguity in bibliographic databases entail a big problem to retrieving information, bibliometric analyzes and visibility of science. This is due to the phenomena called documentary noise and silence. The main objective of this work is to compare the proposal based on agglomerative clustering made by Schulz et al. (2014) with an expanded proposal from the original and determine if it offers better results. In order to achieve the objectives, a dataset has been bult that contains 60.663 records from the area of "Information Science & Library Science" from Web of Science to which both proposals have been applied. Analyzing the results on 5 different authors with ambiguous names, it has been observed that, although the extended proposal slightly improves the results, both proposals offer similar results for solving this problem.

Keywords: term disambiguation, clustering, information retrieval, machine learning, Web of Science

Índice de contenidos

Índice de tablas

Índice de figuras

Índice de ecuaciones

Índice de algoritmos

1. Introducción

La Recuperación de Información (o Information Retrieval) trata de satisfacer las necesidades de información que tienen los usuarios mediante consultas a base de datos, formuladas con un lenguaje adecuado. El objetivo es poder recuperar todos los documentos relevantes ante esa consulta, a la vez que evitamos los no relevantes. Es decir, en el proceso de recuperación de información tenemos que evitar los fenómenos (Pinto, 2018) denominados ruido y silencio documental.

El silencio documental se da cuando documentos almacenados en una base de datos son relevantes a la consulta, pero no han sido recuperados, debido a que la búsqueda ha sido demasiado específica o que no se han usado los términos correctos (Pinto, 2018).

El ruido documental se da cuando se recuperan documentos no relevantes a la consulta, debido a una búsqueda demasiado genérica normalmente (Pinto, 2018).

Estos fenómenos se dan gracias a factores tales como la ambigüedad de la firma de los autores en las bases de datos bibliográficas. Dicha ambigüedad en la firma de los autores se debe a la polisemia y homonimia de los nombres de los autores, errores en los registros de las bases de datos, publicaciones en las que participan autores de múltiples áreas del conocimiento y la falta de atención y/o formación a la hora de firmar los artículos por parte de los autores (Smalheiser & Torvik, 2009).

Todo ello, al afectar directamente a la recuperación de información, también afecta a la realización de análisis bibliométricos y a la visibilidad de la ciencia, puesto que no es posible recuperar e identificar de forma unívoca toda la producción científica de un autor y/o una institución.

La solución y/o reducción de este problema supone un gran avance en el ámbito de la recuperación de información y es de gran ayuda para la comunidad científica. Es por ello por lo que, a lo largo de los años, multitud de autores han desarrollado diferentes técnicas y metodologías para la solución de este problema, tales como los Identificadores Únicos de Autor (como el código ORCID o el ResearchID) y las técnicas de Machine Learning para la identificación en grupos de los artículos de autores.

Entre las técnicas de Machine Learning se pueden encontrar técnicas de aprendizaje supervisado, no-supervisado, semi-supervisado, basadas en grafos entre otras. Diversos autores como pueden ser (Ferreira et al., 2012), (Hussain & Asghar, 2017) y (Shoaib et al., 2020) han propuesto diferentes clasificaciones de dichas técnicas.

Este trabajo se centra en el estudio de las técnicas de aprendizaje no supervisado basadas en clustering aglomerativo para la desambiguación de las firmas de autores. Entre ellas podemos encontrar diferentes metodologías como las propuestas por (Schulz et al., 2014), (Liu et al., 2015) o (Backes, 2018).

1.1 Justificación

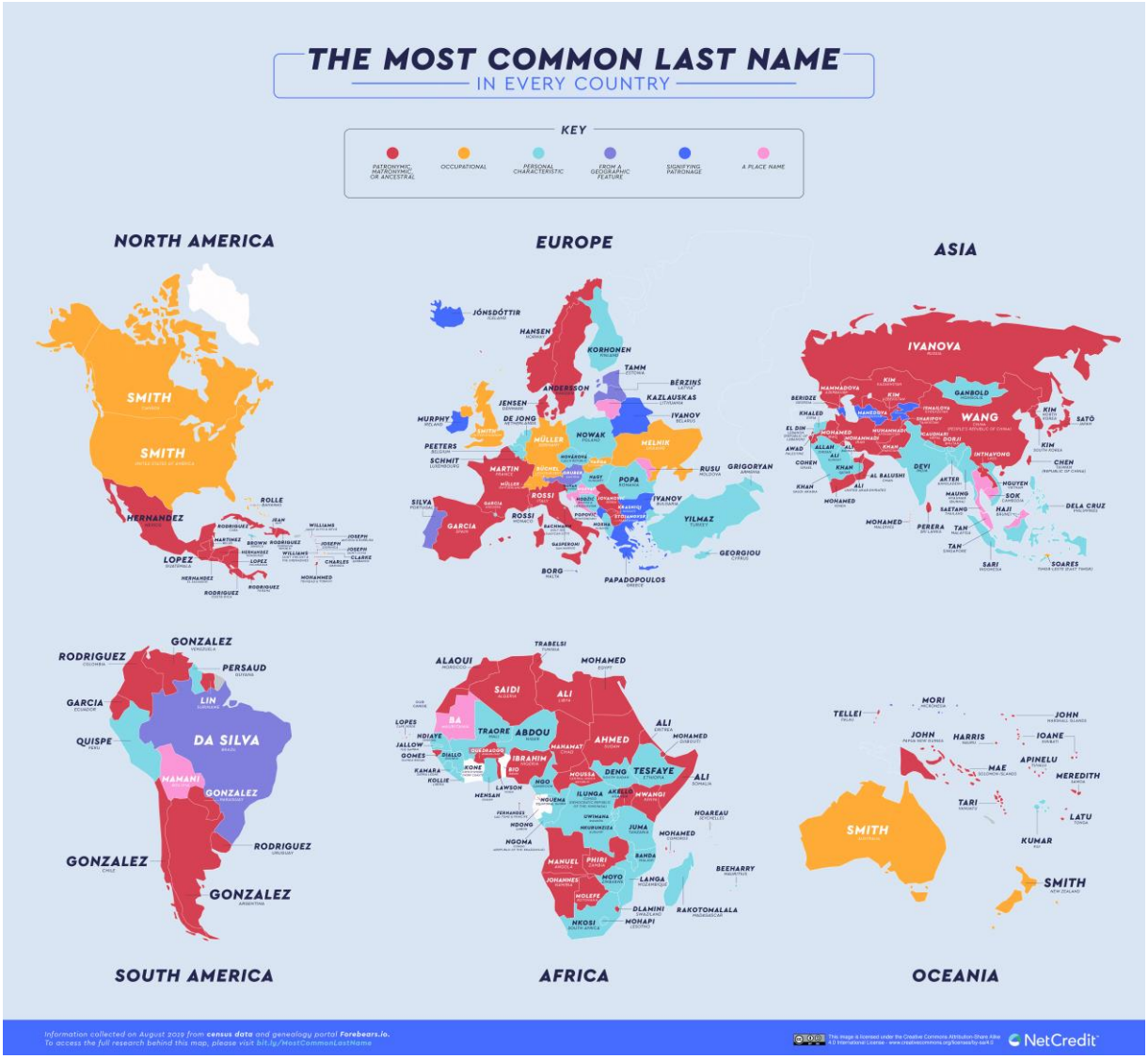
A menudo, cuando se quiere realizar una estrategia de búsqueda de información en bases de datos bibliográficas, no se logra recuperar toda la información de un autor que nos interesa (silencio documental) o, lo más habitual, se encuentra mucho ruido documental en la búsqueda (artículos que no deseamos recuperar). Esto está provocado en gran parte por el problema de la ambigüedad o falta de normalización en la firma de los autores. Puede deberse a las siguientes causas (Smalheiser & Torvik, 2009) (Canteros et al., 2018):

- Autores que firman sus trabajos bajo distintos nombres, denominado homonimia, ya sea por errores ortográficos o tipográficos, variantes en la posición de sus apellidos, cambios de sexo entre otras. Se puede ver un ejemplo de esto en la Figura 1, que muestra una ficha de autor en la base de datos bibliográfica Web of Science, en la que una autora tiene un total de 6 firmas alternativas.

Figura 1. *Ejemplo de variaciones del nombre en la base de datos Web of Science.*

- Autores distintos que firman bajo el mismo nombre, denominado polisemia. Esto ocurre sobre todo con nombres y apellidos muy comunes. En España, 1,46 millones de personas llevan el apellido García y cerca de 1 millón de personas llevan el apellido Sánchez según el Instituto Nacional de Estadística (La Razón, 2021), por lo que es normal que los autores que comparten apellidos comunes firmen de igual forma, provocando así problemas de polisemia. Al igual que en España, esto pasa en todos y cada uno de los países del mundo. En la Figura 2, se puede ver un mapa con los apellidos más comunes en países europeos elaborado por la empresa NetCredit (NetCredit, 2019).

Figura 2. *Apellidos más comunes del mundo en 2019*



Fuente: NetCredit (2019) (https://www.netcredit.com/blog/most-common-name-country/)

Por ejemplo, como puede verse en la Tabla 1, Sánchez, MJ puede referirse tanto a María José Sánchez, María Jesús Sánchez o, también, a Mark John Sánchez entre otros.

Tabla 1. *Ejemplo de polisemia en firmas de autores*

Firma del autor	Nombre real del autor	Título del artículo
Sanchez, MJ	María José Sánchez-Pérez	Cancer incidence estimation from mortality data: a validation study within a population-based cancer registry
Sanchez, MJ	María Jesús Sánchez	Code-Switching, Language Emotionality and Identity in Junot Diaz's "Invierno"
Sanchez, MJ	Mark John Sánchez	Cultures of Empire and International Solidarity

- Los metadatos de los registros en la base de datos pueden no ser suficientes o contener errores que provocarán problemas a la hora de la identificación unívoca del autor. Puede darse el caso de que la base de datos no recoja la localización del autor o la entidad a la que pertenece.

- Incremento de publicaciones multidisciplinares, lo que dificulta encasillar a los autores en ciertas publicaciones científicas. Es decir, un investigador del campo de la ciencia de datos puede colaborar en un artículo de oncología publicado en una revista médica, sin embargo, lo más habitual sería que publicase artículos en una revista de ciencias de la computación.

- Malos hábitos de los autores a la hora de firmar sus trabajos. Muchos autores no prestan atención a la hora de firmar sus publicaciones de forma normalizada, bien sea por desconocimiento o porque simplemente no están dispuestos a dedicar tiempo a ello.

La falta de normalización en bases de datos bibliográficas supone un gran problema que afecta principalmente a:

- Recuperación de información. Un ejemplo de esto sería el siguiente: se necesitan recuperar toda la producción científica realizada por un investigador durante su carrera. Si este investigador no tiene una firma normalizada o tiene un nombre muy común, será difícil poder encontrar todos sus artículos.

- Análisis bibliométricos. La falta de normalización conllevará análisis bibliométricos poco fiables. La bibliometría se encarga del estudio estadístico de la literatura científica, su evolución y autores mediante el uso de indicadores bibliométricos

(Ferreiro, 1993). Es por ello que, si no se identifican correctamente la producción de autores e instituciones, se obtendrán malos análisis bibliométricos.

- Visibilidad de la ciencia y la carrera científica de los investigadores, que viene relacionado con el problema de la recuperación de información. Si no se pueden recuperar toda la producción científica de un autor, estará afectando a la visibilidad de su carrera investigadora.

1.2 Planteamiento del trabajo

En la literatura científica, son muchos los autores que han propuesto métodos basados en técnicas de Machine Learning para tratar de solucionar este problema de la ambigüedad de los autores en bases de datos bibliográficas. Algunos de estos trabajos son (Giles et al., 2005), (Cota et al., 2007), (Tang et al., 2012), (In-Su et al., 2009), (Milojević, 2013), (Backes, 2018), (Liu et al., 2015), (Schulz et al., 2014), (Carvalho et al., 2011).

El presente trabajo pretende implementar una metodología basada en aprendizaje no supervisado, en concreto técnicas de clustering aglomerativo, para la desambiguación de autores provenientes de registros de bases de datos bibliográficas, y ampliar dicha propuesta añadiendo nuevos metadatos para, finalmente, comparar ambas soluciones y analizar si dicha ampliación ofrece mejores resultados que la propuesta original.

La identificación de un buen método para la desambiguación de autores y una posible mejora de este puede ser el punto de partida para la creación de un software que permita la elaboración de estrategias de búsqueda que eliminen por completo el ruido y silencio documental.

Para ello, en primer lugar, será necesario la formación de un dataset que al que podamos aplicar los métodos de desambiguación. En este caso, se usará la base de datos bibliográfica Web of Science para la descarga de los registros. Una vez hayamos elaborado el dataset, se examinarán los atributos que serán necesarios para la aplicación de la metodología y aquellos que se utilizarán para ampliación. Una vez creado el dataset se implementará la propuesta original y la propuesta ampliada, utilizando técnicas de clustering aglomerativo.

Una vez aplicados los algoritmos y obtenidas las métricas, se realizará una comparación exhaustiva y se decidirá cuál ofrece mejores resultados.

1.3 Estructura de la memoria

El presente trabajo comienza con una intromisión al contexto en el que se desarrolla (Capítulo 2. Contexto y estado del arte), analizando estudios relacionados e identificando los autores de referencia en este ámbito. Gracias a ello conoceremos el estado actual de la línea de investigación, problemas identificados y recomendaciones de los autores.

En el Capítulo 3, objetivos concretos y metodología de trabajo, se detallarán, en primer lugar, los objetivos generales y específicos, identificando así las aportaciones del trabajo, y a continuación, se detallará la metodología a seguir para la consecución de los objetivos planteados y el desarrollo del trabajo, es decir, los pasos que van a seguirse, las herramientas y algoritmos que van a utilizarse y cómo va a realizarse el análisis y comparación de los resultados.

En el Capítulo 4, desarrollo específico de la contribución, se detallará la creación del dataset, implementación de los algoritmos, obtención de métricas y comparación de propuestas.

Finalmente, en el Capítulo 5, conclusiones y trabajo futuro, se desarrollarán las conclusiones del trabajo, incluyendo sus limitaciones, y se propondrán las líneas futuras de la investigación.

2. Contexto y estado del arte

A continuación, se desarrolla el contexto y estado del arte del trabajo. Se produce una intromisión en materia de recuperación de información y bases de datos bibliográficas, y los problemas que supone la ambigüedad de la información almacenada en ellas. Además, se analizan las soluciones propuestas para la desambiguación de autores y estudios previos realizados en este ámbito.

2.1. Recuperación de información

La Búsqueda y Recuperación de Información (Information Retrieval) puede definirse como el proceso que tiene como punto de partida un problema o necesidad de información de una persona, que desea resolver mediante recuperación de información, y que termina cuando este problema es resuelto con la información y/o documentos que ha obtenido (Ferrer & Pérez-Montoro, 2011).

En este proceso intervienen los siguientes elementos (Ferrer & Pérez-Montoro, 2011):

- Usuarios: persona que tiene un problema que se puede resolver mediante recuperación de información. Por ejemplo, cuando buscamos reseñas de productos que deseamos adquirir en internet o queremos saber dónde se encuentra la casa de nuestro amigo.

- Necesidad de información: problema que tiene el usuario y que es posible resolver mediante la recuperación de información. Por ejemplo, no saber si el producto que quieres adquirir es bueno o malo o no saber cómo llegar a la casa de nuestro amigo.

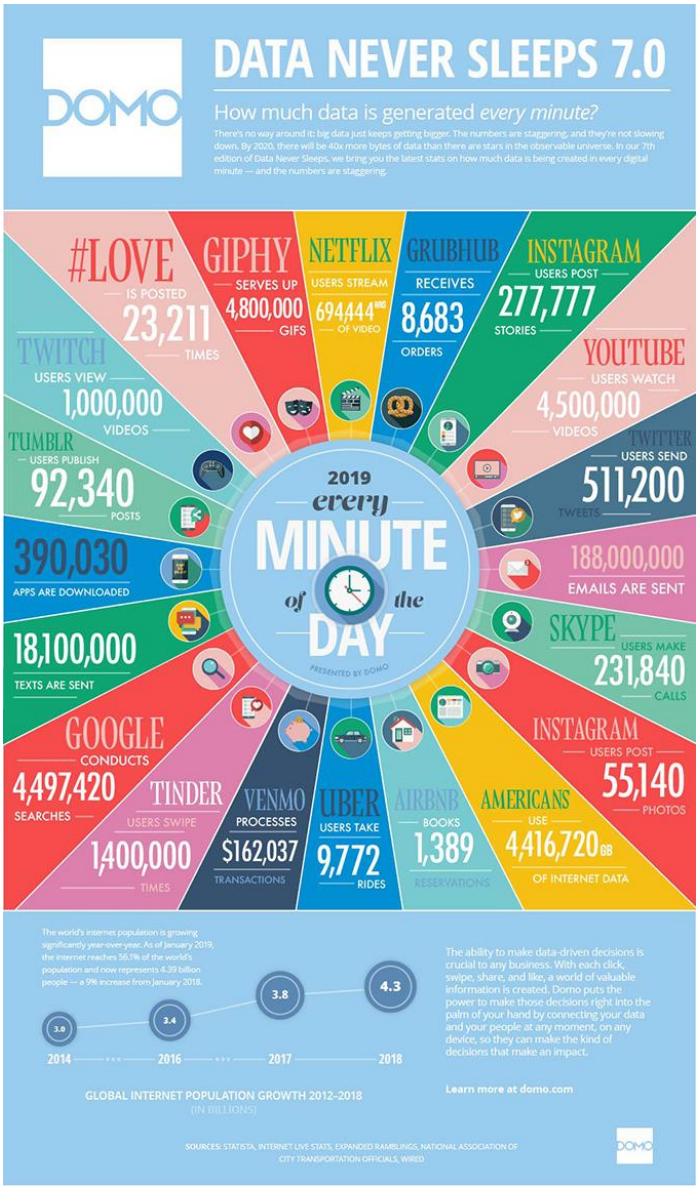
- Consulta: forma en la que se expresa la necesidad de información. Puede ser tanto oral como escrita. En importante en sistemas de recuperación formular una buena estrategia de búsqueda de modo que podamos recuperar la información adecuada a nuestro problema.

- Documentos: es la información recuperada gracias a la consulta. Pueden ser documentos físicos, páginas webs, artículos o imágenes, entre otros.

- Retroalimentación: proceso mediante el cual el usuario valida la información obtenida. Es decir, el usuario decide si la información obtenida es válida y resuelve su necesidad de información o es necesario refinar la búsqueda y volver a lazar otra consulta. Normalmente es necesario realizar más de una vez este proceso hasta encontrar una estrategia de búsqueda adecuada.

En la actualidad, debido al auge de dispositivos móviles y el fácil acceso a internet en cualquier momento y lugar del mundo, las personas estamos continuamente realizando este proceso de búsqueda y recuperación de información. Según un estudio realizado en 2019 por la compañía DOMO, cuyos resultados se pueden ver en la Figura 3, Google recibe en torno a 4,5 millones de consultas por minuto (DOMO, 2019), lo que supone un total de 6,5 billones de consultas alrededor del mundo al día. Esto es un claro ejemplo de la importancia en la actualidad que tiene la búsqueda y recuperación de información, lo importante que es indexar correctamente el contenido y la importancia de formular correctamente las estrategias de búsqueda.

Figura 3. *Infografía de datos generados en un minuto en el año 2018*



Fuente: DOMO (2019) (https://www.domo.com/learn/data-never-sleeps-7)

El objetivo final de toda estrategia de búsqueda de información es poder recuperar todos los documentos relevantes a una consulta, evitando todos aquellos que son irrelevantes. Estos dos fenómenos se denominan silencio y ruido documental (Pinto, 2018).

El silencio documental se da cuando documentos relevantes a la consulta no son recuperados. Esto suele ocurrir con consultas demasiado específicas o con términos erróneos.

El ruido documental, por otro lado, es la situación contraria a silencio documental, es decir, se recuperan documentos que son irrelevantes a una consulta. Esto suele darse con consultas demasiado generales.

2.2. Bases de datos bibliográficas

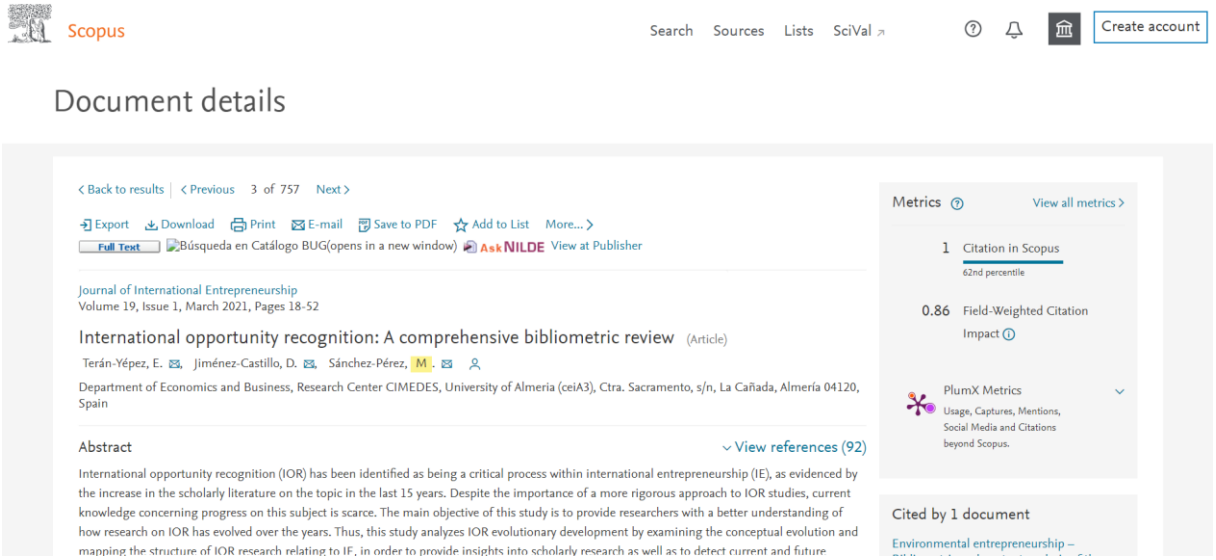
Las bases de datos pueden definirse como (Marqués, 2011) un conjunto de datos que son almacenados y organizados de forma estructurada de manera que pueda satisfacer las necesidades de los usuarios. Estas no pueden ser aprovechadas sin los llamados Sistemas Gestores de Bases de Datos (Somoza, 2015), software que nos permite gestionar y acceder a la base de datos, y los índices, que permiten el acceso a los datos de manera rápida.

En una base de datos, la información está estructurada de la siguiente forma (Somoza, 2015):

- Registro. Cada registro es la representación de un documento o el documento en sí. Todos los registros comparten la misma estructura, por lo que tendrán los mismos campos de datos.
- Campos. Son cada una de las columnas en las que se fracciona un registro, es decir, cada uno de los distintos tipos de datos que componen un registro.
- Datos. Son la unidad mínima de información que se almacena en los campos de los registros.

Las bases de datos bibliográficas están centradas en el almacenamiento de referencias de publicaciones científicas de una o varias áreas de conocimiento, a las que se puede acceder mediante un lenguaje de consulta estructurado. Ejemplos de ellas son Web of Science, Scopus y PubMed entre otras. En la Figura 4, se puede observar un ejemplo de registro almacenado en la base de datos Scopus.

Figura 4. Ejemplo de registro en Scopus.



Estas bases de datos son muy utilizadas para el estudio o investigación sobre cualquier campo de la ciencia, debido a que podemos encontrar artículos de referencia, y para la realización de estudios bibliométricos. Pero, al igual que en todas las bases de datos, podemos encontrarnos con los fenómenos de ruido y silencio documental. Es decir, al realizar una consulta podemos obtener documentos no relevantes o, por el contrario, no recuperar todos los documentos relevantes a la consulta.

Esto está provocado por diferentes factores, divididos principalmente en dos grupos:

- Comportamiento de los usuarios. Muchos usuarios de bases de datos bibliográficas no saben elaborar correctamente estrategias de búsqueda, es decir, no saben representar mediante un lenguaje estructurado su necesidad de información. A menudo, confunden entre los operadores booleanos AND y OR, y no conocen el uso de paréntesis, truncamiento y máscaras.
- Contenido de la base de datos. Es muy usual encontrar registros con campos de datos vacíos, erróneos, ambiguos o no normalizados.

2.3. Ambigüedad en nombres de autores

La ambigüedad o falta de normalización es uno de los grandes problemas, ya que favorece los fenómenos de silencio y ruido documental. Sobre todo, se da en nombres de autores y afiliación de estos.

Uno de los principales factores que favorece la falta de normalización en la firma de los autores es que existen (1) autores que firman sus trabajos de diversas formas (sinonimia), debido a la omisión o variación en la posición de los apellidos y nombres (sobre todo en personas con nombres compuestos), errores y variaciones ortográficas o tipográficas (como el uso de guion para unir apellidos o no), cambio de apellidos al contraer matrimonio en algunos países, cambios de nombre al cambiar de religión o en casos de cambios de género (Smalheiser & Torvik, 2009). Puede darse el caso de que un autor, con nombre compuesto pueda llegar a tener hasta un total de 6 firmas distintas (Figura 1, página 13).

Otro de los grandes problemas son los nombres muy comunes (Smalheiser & Torvik, 2009). En todos los países del mundo existen nombres y apellidos que son muy repetidos entre sus habitantes (Figura 3), por lo que es bastante frecuente encontrar (2) autores distintos que firman bajo el mismo nombre (polisemia). Nombres como María García, Harry Murphy o Thomas Müller son ejemplos de ello.

Esto se suma a que, a menudo, los (3) autores no prestan demasiada atención a la hora de firmar todos sus trabajos de la misma forma, consiguiendo así una firma ambigua. Estos malos hábitos a la hora de firmar sus trabajos pueden ser por falta de conocimiento de las reglas de firma de trabajos o porque no quieren dedicar tiempo a ello.

En la actualidad, existe un (4) incremento de publicaciones multidisciplinares y de multiautoría (Smalheiser & Torvik, 2009). Se dan cada vez más casos de autores publicando en revistas de un área totalmente distinta a la suya, como puede ser un informático publicando en revistas de historia debido a la publicación de un estudio histórico que se ha visto apoyado por la informática en una colaboración puntual, por ejemplo. En cualquier caso, esto dificulta la correcta identificación de las publicaciones de autores.

Por último, podemos encontrarnos, en registros de las bases de datos, (5) metadatos que están erróneos o incompletos (Smalheiser & Torvik, 2009), lo que dificulta la identificación unívoca del autor. Hay casos en los que la base de datos no recoge el primer nombre del autor, la localización de este o la entidad a la que pertenece.

La ambigüedad de autores en las bases de datos bibliográficas supone un gran problema para la recuperación de información, visibilidad de la ciencia y carrera científica de los investigadores y análisis bibliométricos. Si la necesidad de información, por ejemplo, se resume en recuperar todas las publicaciones de un autor concreto en los últimos 5 años, si este tiene una firma ambigua provocará que no se recuperen todos los documentos relevantes a la consulta. Esto afectará directamente a la visibilidad de la ciencia y la carrera

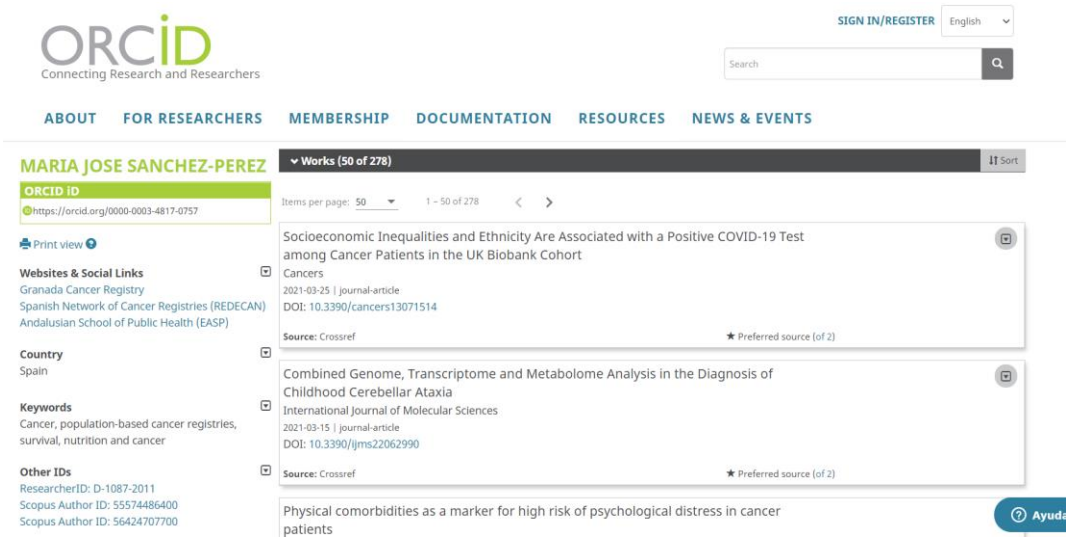
científica de los investigadores. Además, también afectará a los análisis bibliométricos porque provocará análisis poco fiables.

2.4. Identificador Único de Autor como solución al problema

A priori, la solución parece sencilla: crear un Identificador Único de Autor, como si de un DNI se tratara. Esto pondría solución al problema de la ambigüedad, ya que cada autor tendrá su identificador único e intransferible, el cual incluirá asociadas todas sus publicaciones. Al hacer una consulta por este identificador, automáticamente obtendríamos todas sus publicaciones, además, si realizamos consultas por otro tipo de campos, podemos identificar rápidamente si hemos obtenido ruido documental o no, poniendo atención a este campo.

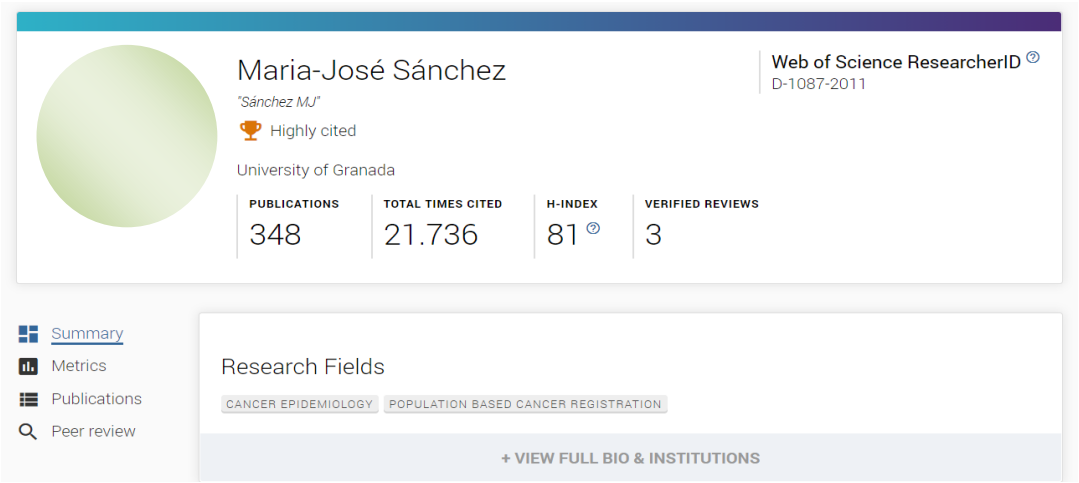
Existen varias propuestas, la más extendida e interesante es el código ORCID. ORCID (Open Researcher & Contributor ID) es una organización internacional, interdisciplinar y sin ánimo de lucro, creada para poner solución al problema de la ambigüedad de los nombres de autores (Haak et al., 2012). Esta organización lanzó en 2009 el código ORCID, un código numérico de 16 dígitos (por ejemplo 0000-1111-2222-3333), con la misión (Haak et al., 2012) de proveer a los investigadores un registro único de autor persistente. Para la creación de este código no se necesita información sensible, basta con el nombre y email, aunque sí es posible añadir información como formación académica, experiencia profesional o proyectos de investigación entre otras cosas si se desea (Sobrido et al., 2016). Reconoce que los investigadores son los dueños de su registro, no la organización, por lo que son estos quienes deciden acerca de la configuración de privacidad de sus datos y qué información se comparte. Entre las características más destacadas (ORCID, s. f.) se encuentran que es un sistema global, los propios investigadores pueden registrarse para conseguir su código ORCID, pueden realizar el mantenimiento de los registros asociados a su código (añadir, actualizar, eliminar registros entre otras funciones) y pueden desactivar su código si ellos así lo desean. En la Figura 5, se puede observar un perfil de autor en ORCID.

Figura 5. *Ejemplo de registro código ORCID.*



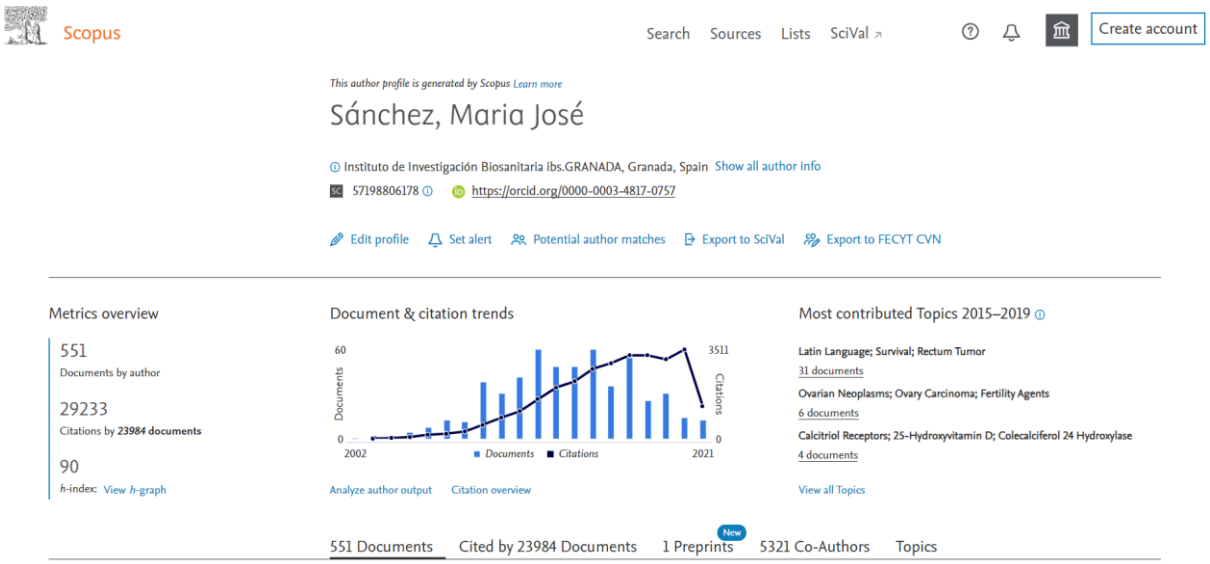
Otra propuesta interesante es el ResearchID de Thomson Reuters (Sobrido et al., 2016), lanzado en 2008. Al igual que el código ORCID y el resto de Identificadores Únicos de Autor, su finalidad es que cada investigador tenga su identificador unívoco con el que se puedan realizar búsquedas sin problemas de ambigüedad. Este código está compuesto por una letra, cuatro dígitos y el año en el que se registra el código, por ejemplo: A-0123-2021. Este código depende del autorregistro y auto identificación de los registros por parte del investigador, por lo que este tiene control total acerca de su código. No existe ningún mecanismo para evitar que un investigador se registre varias veces y pueda obtener varios códigos. Un ejemplo de este código se puede ver en la Figura 6.

Figura 6. *Ejemplo de registro ResearchID.*



La base de datos Scopus también cuenta con su propio Identificador Único de Autor, llamado AuthorID. Fue lanzado en 2006 con el objetivo de proporcionar un identificador único a los investigadores. A diferencia de los anteriores (Sobrido et al., 2016), el código numérico se asigna a cada autor mediante un algoritmo que tiene en cuenta el nombre de autor, coautores, afiliación, dirección, área temática, título de la fuente y fecha de citas. A pesar de ello, puede ocurrir que a un autor se le pueda asignar varios identificadores. Aunque esté automatizado, Scopus facilita mecanismos de verificación, validación y modificación del perfil a los investigadores. En la Figura 7, se puede ver un ejemplo de AuthorID de Scopus.

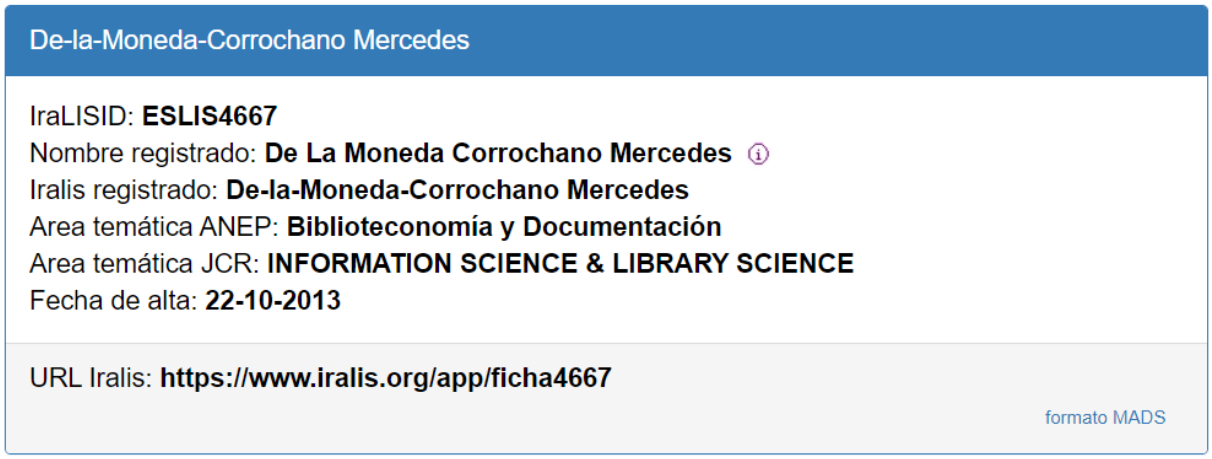
Figura 7. *Ejemplo de registro Scopus Author ID.*



Han ido surgiendo más como puede ser IraLIS (International Registry for Authors in Library and Information Science), proyecto que surge en 2006 para poner solución al problema de la ambigüedad en autores de biblioteconomía, documentación y archivística mediante dos tareas principales (Baiget et al., 2007): crear un registro de nombres de autor que incluya tanto las firmas que el autor haya usado como las firmas que se interprete que pueda usar, y concienciar a los autores para que firmen sus trabajos de forma normalizada. Para obtener el código IraLIS es necesario que los propios autores se registren en la base de datos rellenando un formulario en el que será necesario indicar el nombre completo. IraLIS automáticamente mostrará al usuario las diferentes variantes de su firma y este decidirá la que él prefiera. Cuando se ha completado el registro, el sistema usará tanto la firma decidida como las variantes para identificar al autor. Además, se generará automáticamente un código IraLIS formado por: XX (código del país), ZZZ (especialidad), 00000 (número

automático). Un ejemplo de identificador sería ESLIS4667, como se puede ver en la Figura 8.

Figura 8. *Ejemplo de registro IraLIS.*



Sin embargo, aunque los Identificadores Únicos de Autor sean interesantes, tienen varias desventajas (Smalheiser & Torvik, 2009) que hacen que no sea la solución definitiva al problema de la ambigüedad. La mayoría de los códigos, como hemos visto, dependen del autorregistro y del mantenimiento por parte de los autores, esto obliga a los autores a recordar sus contraseñas y actualizar y mantener sus datos y su producción por el resto de su carrera. Esto supone un gran problema para personas que no están dispuestas a ello. Por otro lado, los códigos son una suma de bastantes dígitos y caracteres, pudiendo llegar hasta 16 dígitos como es el caso del código ORCID y es responsabilidad del autor recordarlo y firmar las publicaciones con dicho código. Además, algunos no tienen mecanismos que eviten que un autor pueda tener varios identificadores, por lo que puede darse el caso de que un autor tenga asignados varios identificadores y cada uno recoja distintas publicaciones, siendo un problema para la normalización. Al ser una actualización manual, también se pueden dar errores tipográficos y ortográficos. Por último, (Shoaib et al., 2020) algunos autores, en especial los más veteranos, no se han adaptado a las nuevas tecnologías y no darán el paso a ellas, por lo que no obtendrán estos identificadores. En conclusión, los Identificadores Únicos de Autor conllevan un comportamiento activo y responsable por parte de los autores, los cuales, a menudo, no están dispuestos a ello.

2.5. Métodos de desambiguación de nombres de autores

Como alternativa a los Identificadores Únicos de Autor, surgen diversos métodos o enfoques para la desambiguación de nombres de autor. Varios estudios han realizado comparativas entre ellos. A continuación, se analizan los más relevantes.

Shoaib et al., en su artículo “Author Name Disambiguation in Bibliographic Databases: A Survey”, analizan diferentes artículos en los que se desarrollan métodos de desambiguación de autores y los categorizan en cinco tipos (Shoaib et al., 2020):

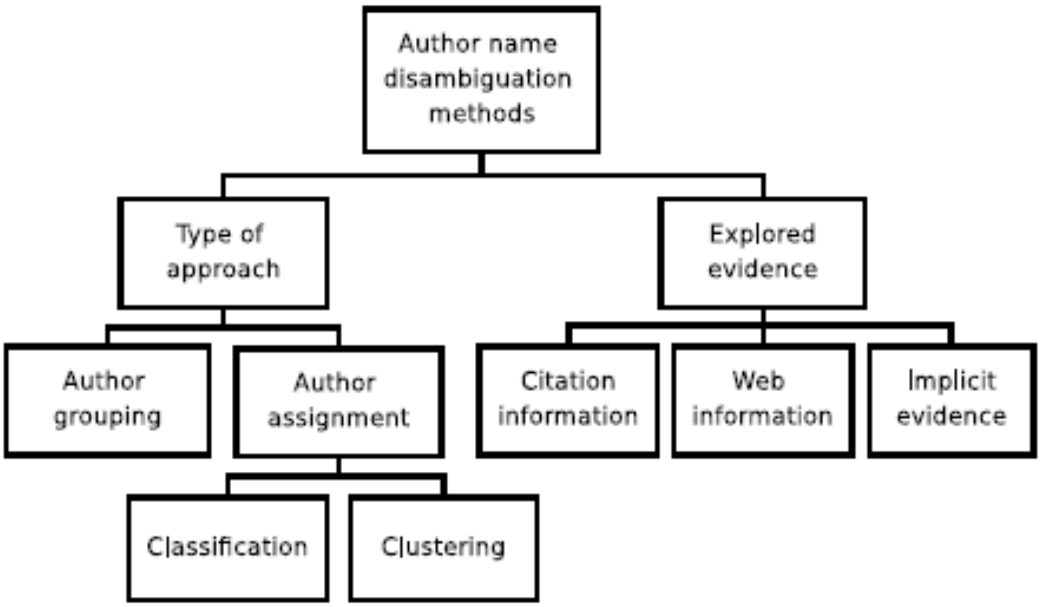
- Aprendizaje supervisado. El objetivo de los métodos de aprendizaje supervisado es encontrar la clase mediante la explotación de información relacionada. Estos métodos pueden usarse para eliminar la ambigüedad de las publicaciones de un autor o para predecir el nombre de un autor en una cita.
- Aprendizaje no supervisado. Los métodos de aprendizaje no supervisado tienen como objetivo realizar agrupaciones. No necesitan de unas clases predefinidas
- Aprendizaje semi-supervisado. Combina las características de los métodos de aprendizaje supervisado y del no supervisado.
- Métodos basados en grafos. Su objetivo es la explotación de relaciones. Es habitual el empleo de redes de co-autoría para ver la similaridad entre dos entidades o nombres de autor.
- Métodos basados en ontologías. Tienen como objetivo la explotación de relaciones entre entidades o autores basadas en semántica.

Todos tienen el objetivo final de separar las publicaciones de un autor en una clase o un grupo único de manera que se puedan identificar todas sus publicaciones unívocamente. En cada uno de ellos identifica: el problema que resuelve, la herramienta y/o método que usa, los atributos y características seleccionados, comparaciones que realiza, dataset utilizado, hallazgos y limitaciones.

Ferreira et al. también presentaron un breve estudio (Ferreira et al., 2012) de los métodos de desambiguación de nombres de autores y propuso una taxonomía (Figura 9) en la que se clasifican y se jerarquizan los métodos. También realiza una descripción general de aquellos más representativos. Incluye métodos tanto de aprendizaje supervisado como no supervisado. Clasifica los métodos según el tipo de enfoque en agrupación de autores (agrupar las referencias dirigidas a un mismo autor mediante la similitud entre atributos) y

asignación de autores (asignar las referencias a sus autores). También hace otro tipo de clasificación de métodos según la evidencia explorada: información de la cita, información de la web o datos implícitos.

Figura 9. *Taxonomía propuesta por Ferreira et al. (2012)*



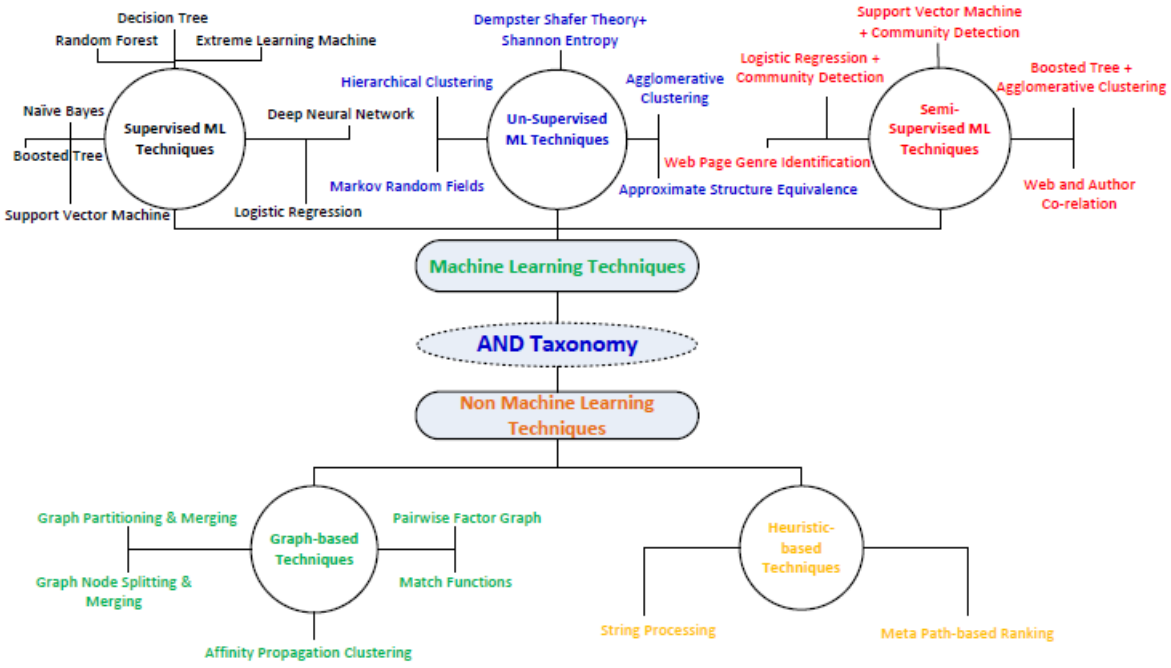
Fuente: Ferreira et al. (2012)

Tekles y Bornmann, en su estudio (Tekles & Bornmann, 2019), se centran en la comparación de métodos de aprendizaje no supervisado para la desambiguación de autores. Analizan el rendimiento general de los métodos, el papel de la parametrización y la dependencia de los resultados de la complejidad de la tarea de desambiguación. Para la evaluación de los métodos, usan un conjunto de menciones de autor que contienen ResearcherID, Identificador Único de Autor de Web of Science. De los métodos utilizados en su conjunto de datos, el propuesto por Caron y van Eck (2014) es el que arroja mejores resultados.

Canteros et al. analiza en su trabajo (Canteros et al., 2018) cuatro métodos distintos de desambiguación e identifica las características en relación con los datos que usan, requisitos de implementación, herramientas de evaluación e impacto en los sistemas de recuperación de información. Con los resultados obtenidos de la comparación de los cuatro métodos, diseña e implementa una estrategia para la desambiguación en un metabuscador del área de las Ciencias de la Computación.

Hussain y Asgha realizan un estudio (Hussain & Asghar, 2017) de las técnicas de desambiguación de nombres de autor entre el periodo 2010 y 2016. El objetivo principal del estudio es analizar los avances realizados en la desambiguación de nombres de autor y la clasificación de las técnicas en supervisadas, no supervisadas, semi-supervisadas, basadas en gráficos y basadas en heurísticas. Realiza una revisión de las técnicas, las compara a nivel abstracto y analiza las limitaciones. La taxonomía propuesta se muestra en la Figura 10.

Figura 10. *Taxonomía propuesta por Hussein y Asghar (2017).*



Fuente: Hussain & Asghar (2017)

Las metodologías para la desambiguación de nombres de autor basadas en técnicas de aprendizaje no supervisado son una gran opción ya que no necesitan de unas clases preestablecidas para realizar la clasificación de los artículos y desambiguar los nombres de autor. En la Tabla 2, se puede observar un conjunto de trabajos relevantes que proponen este tipo de metodologías, el dataset y los atributos que estas utilizan para tal propósito.

Tabla 2. *Métodos de aprendizaje no supervisado para la desambiguación de nombres de autor*

Referencia		(Giles et al., 2005)	(Cota et al., 2007)	(In-Su et al., 2009)	(Carvalho et al., 2011)	(Tang et al., 2012)	(Milojević, 2013)	(Schulz et al., 2014)	(Liu et al., 2015)	(Backes, 2018)
Autores		C. Lee Giles; Hui Han; Hongyuan Zha	Ricardo G. Cota; Marcos André Gonçalves; Alberto H. F. Laender	In-Su Kang; Seung-Hoon Na; Seungwoo Lee; Hanmin Jung; Pyung Kim; Won-Kyung Sung; Jong-Hyeok Lee	Ana Paula de Carvalho; Anderson A. Ferreira; Alberto H. F. Laender; Marcos André Gonçalves	Jie Tang; A.C.M. Fong; Bo Wang; Jing Zhang	Staˇsa Milojevic	Christian Schulz; Amin Mazloumian; Alexander M Petersen; Orion Penner; Dirk Helbing	Yu Liu; Weijia Li; Zhen Huang; Qiang Fang	Tobias Backes
Año		2005	2007	2009	2011	2012	2013	2014	2015	2018
Dataset	**Tamaño**	+400.000 registros	4.239 registros	8.675 registros	+4.500 registros	2,074 registros	45.119 registros	47 millones registros	250.000 registros	1.5 millones ítems
	Disciplina	Variada	Variada	Variada	Variada	Variada	Astronomía, matemáticas, robótica, ecología y economía	Variada	Variada	Variada
Atributos	**Autores**	✔	✔	✔	✔	✔	✔	✔	✔	✔
	Título	✔	✔		✔	✔			✔	
	Revista		✔							
	Afiliación									✔
	Lugar publicación	✔			✔	✔			✔	
	Año				✔	✔				✔
	Resumen					✔				
	Referencias					✔		✔		✔
	Categoría WOS									✔
	Keywords									✔
	Términos Frecuentes									✔
	Mails									✔
	Citas							✔		

Giles et al. proponen (Giles et al., 2005) un modelo de aprendizaje no supervisado que utiliza como técnica de Machine Learning el clustering espectral K-way, y, como atributos para la desambiguación, solamente los nombres de co-autores, título y lugar de la publicación. Como datasets de experimentación utiliza registros descargados de la base de datos DBLP y evalúa los resultados mediante una matriz de confusión.

Dos años más tarde, Cota et al. propusieron (Cota et al., 2007) el clustering jerárquico basado en la heurística como método para la desambiguación de nombres de autor. Este utiliza los nombres de co-autores, el título del artículo y el nombre de la revista como atributos para la desambiguación. También usa registros de DBLP como dataset de experimentación.

In-Su et al. también proponen un método (In-Su et al., 2009) basado en técnicas de clustering, en este caso clustering aglomerativo, y solo utiliza información sobre co-autores como atributo para la desambiguación.

Tang et al. en su propuesta (Tang et al., 2012) utilizan un marco probabilístico basado en el Campo Aleatorio de Markov (Markov Random Field) para afrontar el problema de la desambiguación. Mediante un enfoque dinámico tratan de estimar el número de personas K y mediante un algoritmo de dos pasos tratan de estimar los parámetros. Los atributos utilizados son el título, lugar de publicación, año de publicación, resumen, nombres de autores y referencias.

Carvalho et al. proponen un método (Carvalho et al., 2011) basado en clustering jerárquico basado en heurística y que afronta el problema de la ambigüedad en el momento de insertar nuevos registros en la base de datos, evitando así tener que eliminar la ambigüedad en todo el conjunto de registros una vez insertados. Para ello utilizan como atributos los nombres de autor, título, lugar de publicación y año de publicación.

Milojevic´, ante la dificultad que supone a menudo implementar métodos avanzados de desambiguación, propone un método (Milojević, 2013) más simple que solo utilice los nombres de autor como atributo para resolver el problema. Este método tiene tres vertientes. Una primera que solo utiliza el apellido y la primera inicial del autor, la segunda vertiente que utiliza apellido y todas las iniciales del nombre y una tercera que es un híbrido de las dos anteriores.

Schulz et al., en su estudio (Schulz et al., 2014), proponen un nuevo método de desambiguación basado en clustering aglomerativo en dos pasos, el cual primero conecta artículos individuales y, a continuación, conecta grupos similares. Aunque solo usen como atributos co-autores, referencias y citas, es posible usar este método con cualquier atributo.

Liu et al. también proponen (Liu et al., 2015) un método de desambiguación basado en clustering aglomerativo. En este caso se trata de clustering múltiple en tres pasos usando solamente los atributos más comunes y de fácil acceso: nombres de co-autores, título y lugar de publicación. El primer paso es agrupar los artículos por co-autores. A continuación, en el segundo paso, agrupar los fragmentos anteriores por título. Por último, agrupar los fragmentos resultantes del paso anterior por lugar de publicación.

Backes centra su propuesta (Backes, 2018), al igual que anteriores autores, en la técnica de clustering aglomerativo, pero usando 8 atributos distintos de registros de WoS: términos frecuentes, afiliación, categoría de WoS, palabras clave, co-autores, referencias, emails y año de publicación. Una vez realizado el clustering, evalua los resultados usando el Research ID de WoS.

2.6. Conclusiones

La búsqueda y recuperación de información se ha convertido en una de las actividades principales en nuestro día a día gracias al auge de dispositivos móviles y las facilidades de acceso a la web en cualquier momento y lugar del mundo. Por ello, la formulación de buenas estrategias de búsqueda y la indexación correcta del contenido es muy importante para evitar los fenómenos del silencio y ruido documental. Estos fenómenos provocan que no encontremos todos los documentos relevantes en nuestra consulta (silencio documental) o, bien, que recuperemos documentos no relevantes (ruido documental).

Este problema se da en todos los tipos de bases de datos, pero se convierte en un problema grave en el caso de las bases de datos bibliográficas. Estas bases de datos son muy usadas en investigación y la existencia del problema del ruido y silencio documental provoca errores especialmente en la recuperación de información, la visibilidad de la ciencia y los análisis bibliométricos. Este problema se ve más claro con el siguiente ejemplo: si no se recuperan todos los artículos relevantes sobre el tema de investigación en el que estemos trabajando, estaremos perdiendo información muy valiosa para el estudio.

En este caso, estos problemas no solo vienen dados por el comportamiento de los usuarios a la hora de establecer sus estrategias de búsqueda, también están muy afectados por la falta de normalización del contenido de las bases de datos. Esta falta de normalización conlleva a la ambigüedad del contenido, que afecta especialmente a los nombres de autores. La ambigüedad de nombres de autores está provocada por los fenómenos de la sinonimia y polisemia, el comportamiento de los autores a la hora de firmar sus trabajos, el

auge de publicaciones multidisciplinares y la existencia de metadatos erróneos o incompletos en los registros bibliográficos.

Para poner solución a este problema de ambigüedad de nombres de autor, se han propuesto los Identificadores Únicos de Autor, que son códigos alfanuméricos que identifican unívocamente a los autores en las bases de datos bibliográficas. Se han propuesto a lo largo del tiempo varias soluciones como son el código ORCID, el ResearchID o el código IraLIS, sin embargo, no han supuesto una solución total al problema debido a que la mayoría dependen de la gestión y el mantenimiento por parte de los propios autores, los cuales, por diversos motivos como falta de tiempo o poca soltura con las nuevas tecnologías, mantienen sus perfiles desactualizados, duplicados o, simplemente, no disponen de estos perfiles.

Surgen entonces, los métodos de desambiguación de nombres de autor para paliar este problema. Existen muchos métodos de desambiguación basados en aprendizaje supervisado, aprendizaje no supervisado, basados en ontologías entre otros. Algunos autores, en sus estudios, comparan diferentes metodologías y realizan clasificaciones. Algunos de estos autores son: (Shoaib et al., 2020) clasifica los métodos de desambiguación en aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi-supervisado, basados en ontologías y basados en grafos; (Ferreira et al., 2012) propone una taxonomía en la que divide los métodos según el tipo de enfoque y según la evidencia explorada; (Hussain & Asghar, 2017) propone una taxonomía en la cual separa las técnicas basadas en Machine Learning y las no basadas en Machine Learning, y dentro de cada una de ellas las separa en aprendizaje supervisado, no supervisado y semi-supervisado en el primer caso, y en técnicas basadas en grafos y basadas en ontologías en el segundo caso.

Los métodos basados en aprendizaje no supervisado son una gran opción debido a que no necesitan de unas clases preestablecidas para realizar la desambiguación de los autores. Es por ello por lo que son el objeto de estudio de este trabajo. Existen gran cantidad de metodologías basadas en aprendizaje no supervisado. Algunas de ellas solo utilizan los nombres de los autores como atributo para su aplicación mientras que otras pueden llegar a utilizar hasta ocho atributos distintos. Las propuestas más recientes son las realizadas por (Schulz et al., 2014), (Liu et al., 2015) y (Backes, 2018). Todas ellas utilizan el clustering aglomerativo como técnica de desambiguación, aunque su aplicación es diferente y no utilizan los mismos atributos.

3. Objetivos concretos y metodología de trabajo

A continuación, se presentan los objetivos generales y específicos de la investigación, y se detalla la metodología usada para el desarrollo de esta.

3.1. Objetivo general

El objetivo principal de este trabajo es la comparación de la propuesta original realizada por Schulz et al. (2014) para la desambiguación de nombres de autor basada en técnicas de clustering aglomerativo, con una propuesta ampliada a partir de la original para determinar si esta ampliación ofrece mejores resultados.

3.2. Objetivos específicos

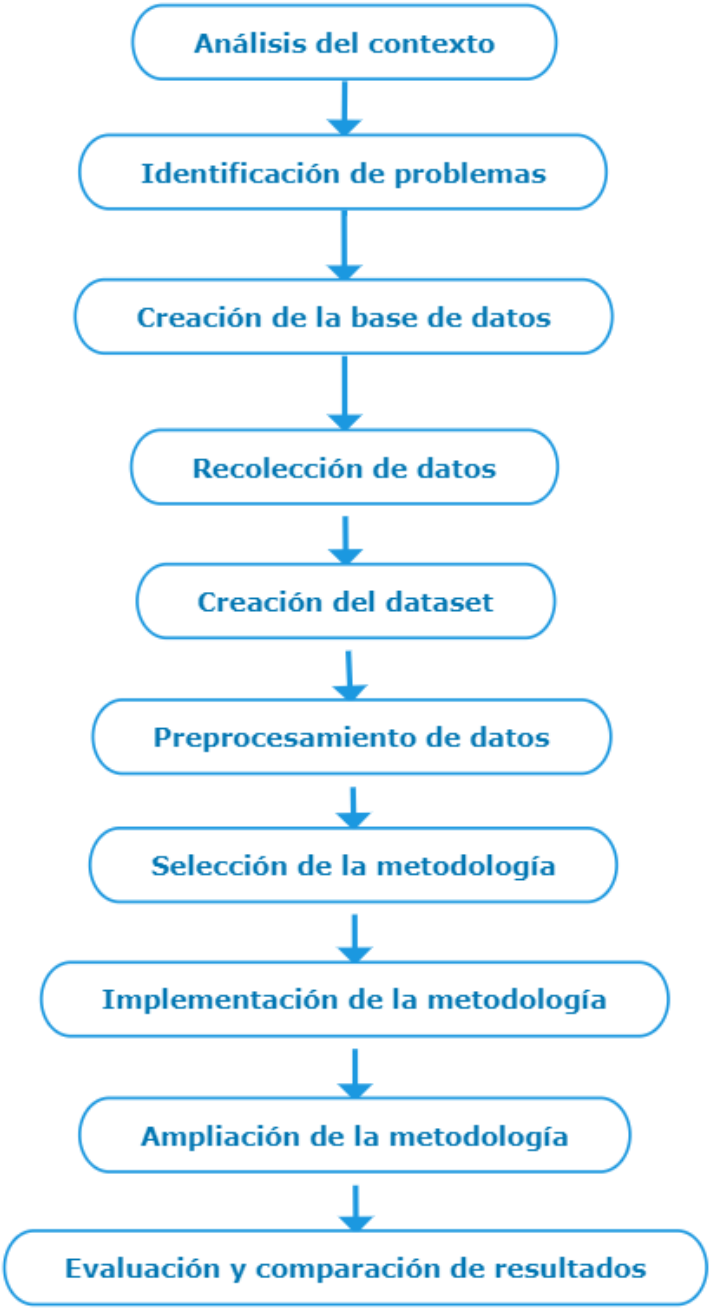
Los objetivos específicos son los siguientes:

- Analizar los estudios actuales sobre la desambiguación de autores para adentrarnos en el contexto.
- Identificar los problemas reales de la falta de normalización de autores en bases de datos bibliográficas.
- Explorar las metodologías propuestas para la desambiguación de autores.
- Construir un dataset de registros bibliográficos para la aplicación de técnicas de aprendizaje no supervisado.
- Implementación de la propuesta original de Schulz et al. (2014) para la desambiguación de nombres de autor.
- Implementación de una propuesta ampliada para la desambiguación de nombres de autor basada en la propuesta original de Schulz et al. (2014).
- Comparar los resultados obtenidos en cada una de las propuestas.
- Determinar qué propuesta arroja mejores resultados y, por lo tanto, es más adecuada para la resolución del problema de la ambigüedad de los autores en bases de datos bibliográficas.

3.3. Metodología del trabajo

En la Figura 11, se pueden observar los pasos a seguir propuestos para la realización de este trabajo y la consecución de los objetivos.

Figura 11. *Metodología propuesta para la realización del trabajo.*



Para la realización de este trabajo se comienza con una revisión del contexto en el que se desarrolla y de los trabajos previos relacionados, que va a permitir identificar correctamente

los problemas que supone la ambigüedad de los autores y posibles soluciones. Para ello se realiza una revisión sistemática de la literatura sobre recuperación de información, bases de datos bibliográficas, ambigüedad de autores y métodos propuestos para la resolución de este problema.

Una vez analizado el contexto e identificados los problemas y posibles soluciones, es necesario crear una base de datos que servirá de almacén en la descarga y creación del dataset. Para ello se utiliza phpMyAdmin, (phpMyAdmin, s. f.) software gratuito, escrito en PHP, el cual permite, mediante una interfaz web fácil e intuitiva, la administración de bases de datos MySQL y MariaDB. Para su creación se realiza un análisis exhaustivo de los atributos que van a ser almacenados y se desarrolla el modelo entidad relación que servirá de estructura.

Creada la base de datos, se comienza con la recolección de datos. Se recolectan datos procedentes de registros de Web of Science (WoS). Web of Science es una base de datos multidisciplinar que incluye artículos desde el año 1900 de más de 12.000 revistas (Gasparyan et al., 2013). Es una de las bases de datos bibliográficas más importantes junto a Scopus, PubMed/Medline o The Cochrane Library, en concreto, WoS es la más reconocida y prestigiosa, e instituciones académicas de todo el mundo fomentan la publicación en revistas indexadas en WoS para aumentar su reconocimiento y posicionamiento en los rankings (Gasparyan et al., 2013). Para la descarga de los registros, se programa un script que conecta con WoS y hace descargas masivas de registros bibliográficos. Se descargan artículos del área de investigación “Information Science & Library Science” publicados en revistas indexadas en la colección principal de Web of Science. El número total de registros descargados es 63.937. Se utiliza esta área de investigación por los siguientes motivos: (1) es una temática suficientemente grande y abordable para la realización de un trabajo de estas características; (2) la temática está directamente relacionada con los estudios universitarios de grado realizados, por lo que se tiene un conocimiento de la categoría en cuestión; (3) como podemos observar en la Tabla 3, cuyos datos han sido extraídos de la herramienta Co-author Index del Grupo de Investigación EC3 de la Universidad de Granada (EC3 - Grupo de Evaluación de la Ciencia y la Comunicación Científica, 2016), el número medio de autores por disciplina en artículos publicados en Web of Science es contenido, no siendo demasiado elevado como en disciplinas de Ciencias de la Salud (lo que provocaría análisis y evaluaciones demasiado complejas), ni tan pequeño como en disciplinas de Artes y Literatura.

Tabla 3. *Número medio de autores por artículo en diferentes disciplinas.*

Disciplina	Nº medio de autores por artículo	
	Nacional	Internacional
Artes	1,2	1,6
Biología	3,6	7,9
Biomedicina	5,6	9,8
Ciencia Política	1,5	2,4
Derecho	1,2	2
Documentación	1,8	3,1
Farmacia	4	7,6
Informática	3,4	4,1
Ingenierías	3,2	6,8
Literatura	1,1	1,1

Fuente: Elaboración propia con datos extraídos de la herramienta Co-author Index del Grupo de Investigación EC3 de la Universidad de Granada (EC3 - Grupo de Evaluación de la Ciencia y la Comunicación Científica, 2016) (http://www.coauthorindex.info/)

Una vez descargados los registros, se procede al preprocesamiento de los datos. Este preprocesamiento consiste en la selección de registros objeto de estudio, selección de atributos que se van a utilizar para la aplicación de las propuestas de desambiguación, descarte de registros que contengan alguno de los atributos vacíos, ya que añaden ruido y no son relevantes, y normalización de los datos.

A continuación, se procede con la selección de la metodología de desambiguación que será implementada. Este trabajo se centra en técnicas de desambiguación de nombres de autor que estén basadas en técnicas de aprendizaje no supervisado, en concreto técnicas de clustering jerárquico aglomerativo. Se implementa la metodología de desambiguación propuesta por (Schulz et al., 2014), basada en clustering jerárquico aglomerativo en dos pasos.

Seleccionada la metodología que se va a implementar, se aplica siguiendo los pasos y se realiza otra propuesta añadiendo nuevos metadatos a la propuesta original.

Finalmente, se realiza una comparativa de resultados entre la propuesta original y la ampliada utilizando la precisión y el recall o exhaustividad. Para ello se seleccionan 5 autores distintos cuyos nombres de firma son ambiguos y se le aplican ambas propuestas. Una vez se han obtenido los resultados del clustering, con ayuda de los perfiles oficiales de los autores en Web of Science, se obtienen los listados de sus publicaciones y, mediante el

código de identificación del artículo, se realiza una comparativa con la base de datos descargada para identificar cuántos de ellos han sido obtenidos.

Ahora, estando en disposición del listado de artículos de cada autor en la base de datos descargada y el listado de artículos de cada clúster, se realiza una revisión manual entre los listados para cada uno de los autores en cada propuesta.

Para el cálculo de la precisión se examina la cantidad de publicaciones del clúster que se han clasificado correctamente (Ecuación 1). Es decir, el número total de publicaciones del clúster correctamente clasificadas entre el número total de publicaciones del clúster.

Ecuación 1. *Ecuación para el cálculo de precisión.*

$$Precisión = \frac{publicaciones\ correctamente\ clasificadas}{total\ de\ publicaciones\ del\ clúster}$$

Por otro lado, para calcular la exhaustividad se observa el número total de publicaciones clasificadas correctamente en el clúster y se divide entre el total de publicaciones del autor que contiene el dataset. Se utiliza para ello la Ecuación 2.

Ecuación 2*. Ecuación para el cálculo del recall o exhaustividad.*

$$Recall = \frac{publicaciones\ correctamente\ clasificadas}{total\ de\ publicaciones\ del\ autor}$$

Obtenidas las métricas de cada uno de los autores en ambas propuestas, se realiza una comparación de estas y se establece qué propuesta arroja mejores resultados.

4. Desarrollo específico de la contribución

4.1. Creación de la base de datos

Previo a la recolección de datos y creación del dataset, se crea una base de datos en la que se almacenarán los registros descargados de Web of Science. En este caso, se crea una base de datos de tipo relacional debido principalmente a dos razones: (1) los registros de Web of Science tienen una estructura fija, por lo que se conocen todos los campos que estos contienen permitiendo así mantener la uniformidad de los registros; (2) las bases de datos relacionales garantizan la integridad de los registros y la no duplicidad de estos usando un identificador único para cada uno de ellos.

Para tal propósito se usa phpMyAdmin. Este software gratuito, escrito en PHP, permite administrar bases de datos MySQL y MariaDB a través de una interfaz de usuario fácil e intuitiva a través de la web (phpMyAdmin, s. f.)

El primer paso para la creación de la base de datos es el análisis de la estructura de los registros de Web of Science y el desarrollo del modelo de datos que será utilizado.

La colección principal de Web of Science, de la cual serán descargados los registros, contiene un total de 73 campos distintos que podemos ver en la Tabla 4 (Clarivate Analytics, 2020). Algunos de los campos más relevantes y que nos servirán posteriormente para la desambiguación de los autores son *Autores, Citas, Referencias, Keywords y Área de Investigación.*

No todos los campos son almacenados en la base de datos debido a que existen gran cantidad de ellos que son irrelevantes e innecesarios para la finalidad del dataset. En este caso, el dataset está centrado únicamente en la tipología de artículos de revista, por lo que automáticamente se descartan todos aquellos campos que están relacionados con libros y conferencias.

Tabla 4. *Etiquetas de campo de la colección principal de Web of Science.*

Etiquetas de campo de la Colección principal Web of Science					
FN	Nombre de archivo	RP	Dirección para petición de copias	PN	Identificador de subdivisión de publicación
VR	Número de versión	EM	Dirección de correo electrónico	SU	Suplemento
PT	Tipo de publicación (J=Revista; B=Libro; S=Colección; P=Patente)	RI	Número de ResearcherID	MA	Abstract de reunión
AU	Autores	OI	Identificador ORCID (Open Researcher and Contributor ID)	BP	Página de inicio
AF	Nombre completo de autor	FU	Entidad financiadora y número de concesión	EP	Página final
BA	Autores del libro	FX	Texto de financiación	AR	Número de artículo
BF	Nombre completo de autores del libro	CR	Referencias citadas	DI	Identificador digital de objeto (DOI)
CA	Autoría conjunta	NR	Número de referencias citadas	D2	Identificador digital de objeto de libro (DOI)
GP	Autoría conjunta del libro	TC	Número de veces citado de la Colección principal de Web of Science	EA	Fecha de acceso anticipado
BE	Editores	Z9	Número total de veces citado	EY	Año de acceso anticipado
TI	Título de documento	U1	Recuento de uso (Últimos 180 días)	PG	Número de páginas
SO	Nombre de publicación	U2	Recuento de uso (Desde 2013)	P2	Número de capítulos (Book Citation Index)
SE	Título de colección	PU	Editorial	WC	Categorías de Web of Science
BS	Subtítulo de colección	PI	Ciudad de la editorial	SC	Áreas de investigación
LA	Idioma	PA	Dirección de la editorial	GA	Número de entrega de documento
DT	Tipo de documento	SN	Número Internacional Normalizado de Publicaciones Seriadas (ISSN)	PM	ID de PubMed
CT	Título de la conferencia	EI	Número Electrónico Internacional Normalizado de Publicaciones Seriadas (eISSN)	UT	Número de acceso
CY	Fecha de la conferencia	BN	Número Estándar Internacional de Libros (ISBN)	OA	Indicador de acceso abierto
CL	Ubicación de la conferencia	J9	Abreviatura de la fuente de 29 caracteres	HP	Artículo popular de ESI
SP	Patrocinadores de la conferencia	JI	Abreviatura de la fuente ISO	HC	Artículo muy citado de ESI
HO	Organizador de la conferencia	PD	Fecha de publicación	DA	Fecha en la que se generó este informe.
DE	Palabras clave de autor	PY	Año de publicación	ER	Fin del registro
ID	KeyWords Plus®	VL	Volumen	EF	Fin del archivo
AB	Abstract	IS	Número		
C1	Dirección de autor	SI	Número especial		

Fuente: Elaboración propia con datos obtenidos de la página web de Web of Science (https://images.webofknowledge.com/WOKRS517B4/help/es_LA/WOK/hs_wos_fieldtags.html)

Después de un análisis exhaustivo, los campos finalmente almacenados son los siguientes (Clarivate Analytics, 2020):

- Autores. El campo autores contiene todos los nombres de los autores que firman el documento. Es decir, incluye tanto el autor principal del documento como los coautores.
- Citas. Este campo contiene el número total de citas. Además, también se puede acceder mediante él a los artículos citantes, es decir, aquellos artículos que citan en sus referencias al artículo en cuestión.
- Referencias. Este campo contiene el número de referencias del artículo y los artículos que referencia, es decir, aquellos artículos que el artículo en cuestión nombra o cita.
- Keywords de autor. El campo keywords recoge las palabras clave del artículo, es decir, los términos que representan el contenido del artículo. En Web of Science existen dos tipos, las keywords de autor y las keywords de la base de datos o keywords plus. Las keywords de autor son las palabras clave que están especificadas por los autores mientras que las keywords plus son definidas por Web of Science mediante la recolección de palabras provenientes del título. En este caso se centra en las keywords de autor.
- Identificador WOS. El identificador de Web of Science es un código de acceso único que tiene asociado cada uno de los registros de la base de datos. Es decir, es como el DNI/NIF de los registros.
- Año de publicación. Contiene el año de publicación del artículo.

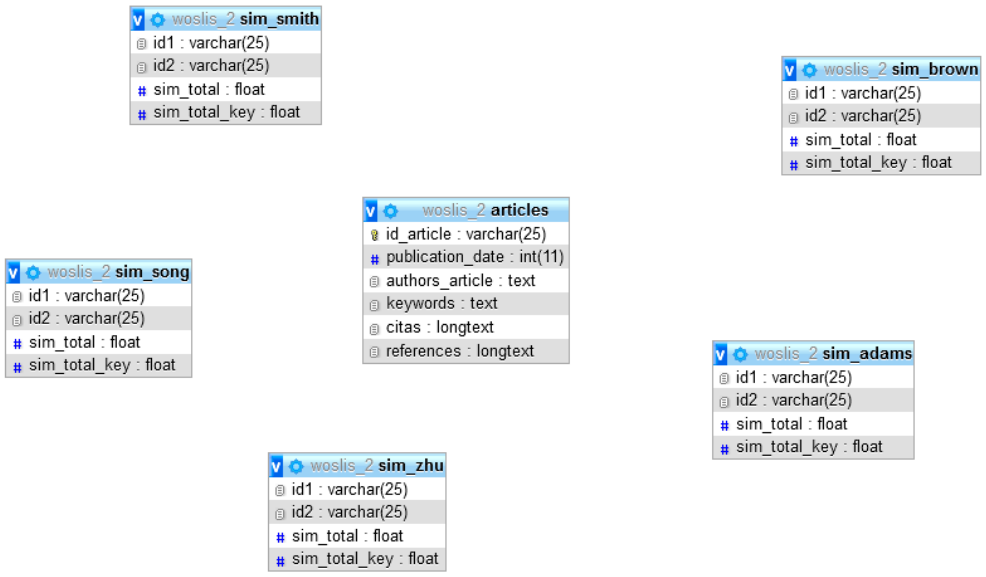
Con estos campos se desarrolla la tabla principal de la base de datos, llamada articles, que almacenará toda la información descargada de los registros y que está compuesta por 6 columnas distintas: id_article, publication_date, authors_article, keywords, cites, references. La clave primaria es el identificador de WOS y el resto de los atributos son el año de publicación, los autores del artículo, las palabras clave de autor, las citas y las referencias

Además, se crean 5 tablas más, uno por cada autor sobre los cuales se aplicarán las propuestas para evaluar los resultados y que incluirán los cálculos de similaridad por cada par de artículos de cada uno de ellos: sim_smith, sim_brown, sim_song, sim_adams, sim_zhu. Todas ellas están compuestas por cuatro columnas: el identificador del artículo 1,

el identificador del artículo 2, la similaridad total de la propuesta original y la similaridad total de la propuesta ampliada.

De este modo, el modelo de datos final está compuesto por 6 tablas distintas, tal y como puede verse en la Figura 12.

Figura 12. *Estructura de la base de datos.*



4.2. Creación del dataset

Para la realización de este trabajo se crea un dataset propio sobre el que se aplican los métodos de desambiguación de nombres de autor. Como característica distintiva de la mayoría de los trabajos analizados en la Tabla 2 (página 30), se ha optado por descargar los registros de la base de datos Web of Science. Solo dos de los nueve trabajos analizados, (Schulz et al., 2014) y (Backes, 2018), utilizan registros de dicha base de datos. Además, el dataset se centra exclusivamente en el área de investigación de Web of Science "Information Science & Library Science", por lo que se trata de un dataset unidisciplinar, al contrario del resto de trabajos analizados en la Tabla 2 (página 30) que son multidisciplinares. El uso de esta disciplina se debe principalmente a tres motivos:

1. Disciplina suficientemente grande y abordable para este trabajo.
2. Amplio conocimiento y manejo de la disciplina.

3. El número medio de autores por artículo en esta disciplina es contenido tal y como observamos en los datos recogidos en la herramienta Co-author Index (EC3 - Grupo de Evaluación de la Ciencia y la Comunicación Científica, 2016).

Para la descarga de los registros, se programa un script (Algoritmo 1) en lenguaje de programación Python. Este script, conecta con Web of Science, ejecuta una consulta o query y hace descargas masivas de registros bibliográficos.

La estrategia de búsqueda o query para la descarga de los registros es sencilla. Como se quieren recolectar registros publicados de la Colección principal de Web of Science, cuya área de investigación sea “Information Science & Library Science” y cuyo tipo de documento sea artículo, la estrategia de búsqueda es la siguiente: SU=Information Science & Library Science AND DT=(Article).

Formulada la consulta, el script va lanzando la consulta y comienza la obtención masiva de registros, los cuales se van descargando en ficheros CSV año por año. Además, si el registro contiene citas, se descarga un fichero CSV el cual contiene todos y cada uno de los artículos que lo citan. A fecha 1 de agosto de 2021, se han descargado un total de 160.888 registros distintos.

Algoritmo 1. *Descarga de registros de Web of Science*

```
Paso 1: Conexión con Web of Science a través de la API

Paso 2: Definición de la query

Paso 3: Lanzamiento de la consulta a través de la API

Paso 4: Descarga de registros en ficheros CSV año por año.

Paso 5: Si el registro contiene citas, descargar los registros
de los artículos citantes en un fichero CSV.

Paso 6: Cierre conexión Web of Science
```

Posteriormente, se realiza una limpieza de los ficheros CSV, eliminando aquellos registros cuyos autores no sean conocidos (es decir, cuyo nombre en la base de datos es desconocido y es etiquetado como Anonymous en ella) y/o que no contengan palabras clave de autor.

Una vez realizada la limpieza de los ficheros CSV, comienza la carga estos en la base de datos creada previamente. Para ello se lanza un script en el que, en primer lugar, se comprueba si los registros se encuentran ya en la base de datos y de no ser así se incluye

el registro con las columnas seleccionadas. Una vez importados todos los registros únicos en la base de datos, por cada uno de ellos, se comprueba si contiene citas, y de ser así, se importan los identificadores de cada artículo citante en el campo citas de la tabla de la base de datos. Los pasos de este algoritmo pueden verse en el Algoritmo 2.

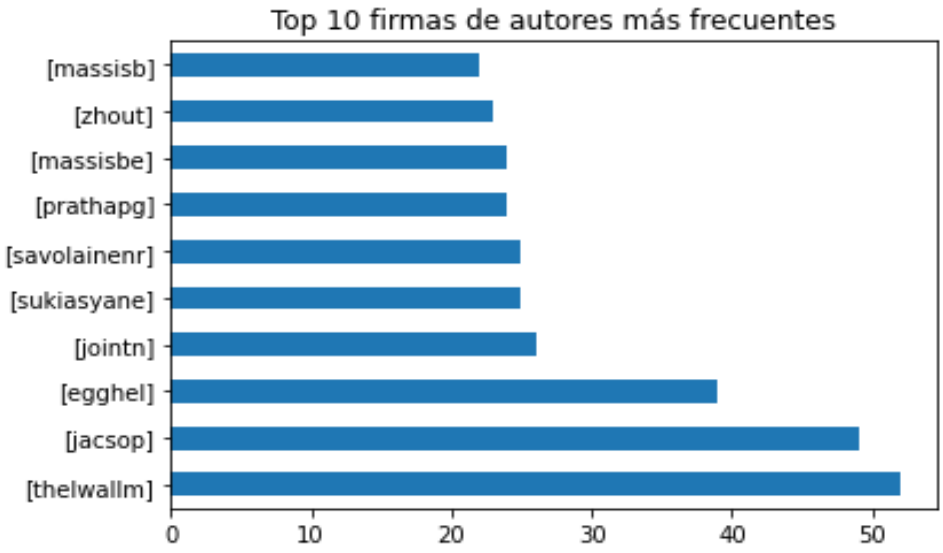
Algoritmo 2*. Importación de registros a la base de datos MySQL*

```
Paso 1: Conexión con la base de datos MySQL

Paso 2: Comprobar si el registro existe en la base de datos: si
el registro no existe, importar el registro seleccionando las
columnas necesarias.

Paso 3: Finalización de importación de todos los registros

Paso 4: Importación de las citas: si el registro contiene un
fichero con los artículos citantes, importar los identificadores
de cada uno de los registros del fichero en el campo citas de la
tabla de la base de datos.

Paso 5: Finalización de importación de citas.

Paso 6: Cierre de conexión con la base de datos MySQL
```

Finalmente, la base de datos contiene un total de 60.663 registros de artículos científicos del área de investigación “Information Science & Library Science”, cuyos campos de autores y palabras clave de autor no están vacíos.

La base de datos contiene un total de 80.696 firmas de autor distintas. Se pueden observar las 10 firmas de autores más frecuentes junto con el número de publicaciones en las que aparecen en la Figura 13.

Figura 13. *Top 10 firmas de autores más frecuentes*



4.3. Preprocesamiento de datos

Una vez creado el dataset, se comienza la preparación o preprocesamiento de datos para su posterior uso en la metodología de desambiguación seleccionada y su ampliación. Debido a que la propuesta ampliada usa un atributo extra para la desambiguación de nombres de autor, el preprocesamiento difiere un poco en ambas propuestas.

Para la implementación de ambas propuestas serán necesarios los atributos autores, citas y referencias, además del atributo keywords de autor que será usado exclusivamente en la propuesta ampliada. Han sido seleccionadas las keywords de autor y no las keywords plus porque dependen de la actuación y decisión de los autores, y esto puede relacionarse con un autor en concreto. Es decir, un mismo autor, es probable que utilice las mismas palabras clave y de la misma forma siguiendo unos patrones, así que será más útil a la hora de la desambiguación que las keywords plus, ya que estas las asigna WoS automáticamente mediante las palabras del título del artículo y no son controladas por los autores. El tratamiento y preprocesamiento de los atributos comunes (autores, citas y referencias) será el mismo en las dos propuestas.

El primer paso del pre-procesamiento es la importación en Python de los atributos necesarios en forma de dataframe. Para ello, se llaman las librerías necesarias para la importación y tratamiento de los datos y se definen las credenciales que permitirán conectar con la base de datos MySQL. A continuación, se crea la conexión con la base de datos y un cursor que permite realizar las operaciones CRUD (Create, Read, Update, Delete). Una vez

creada la conexión con la base de datos, se definen las consultas SQL que reportarán los datos necesarios para la implementación de las propuestas. En este caso, se definen tres consultas distintas, una para el atributo autor, otra para el atributo citas y otra para el atributo keywords de autor. Los resultados de las consultas son almacenados en un dataframe para su posterior procesamiento.

Los dataframes resultantes son los siguientes:

- Dataframe de autores. Está compuesto por dos columnas, el identificador del artículo y la lista de los autores del artículo. Contiene tantas filas como artículos únicos.
- Dataframe de citas. Está compuesto por dos columnas, el identificador del artículo y la lista de los identificadores de los artículos que lo citan. Al igual que en el dataframe de autores, contiene tantas filas como artículos únicos.
- Dataframe de keywords. Este dataframe está compuesto por dos columnas, una primera que incluye el identificador del artículo, y una segunda columna que incluye el listado de keywords de autor. Como en los dataframes anteriores, contiene tantas filas como artículos tiene el dataset.

Con respecto al dataframe de referencias, no ha sido posible obtenerlo. Durante la descarga de registros, la API oficial utilizada para la descarga de estos ha sufrido una actualización en la que ha eliminado la opción de obtener dicho campo desde Web of Science. Debido a esto, solo se han obtenido las referencias de unos pocos registros, por lo que, para evitar resultados no fiables, se ha optado por suprimir dicho atributo.

Formuladas las consultas e importados los dataframes, se realiza una limpieza y normalización de los datos. En el caso de los autores, se pasan a minúscula, y se eliminan signos de puntuación, guiones y espacios, dando como resultado una cadena de caracteres por autor. Por ejemplo, Sánchez-Pérez, MJ queda como sanchezperezmj.

Por otro lado, en el caso de las keywords, además de pasar todo a minúscula y eliminar signos de puntuación, guiones y espacios, es necesario realizar un stemming para llevar la palabra a su raíz y que de esta forma se tomen los términos que se encuentren en plural, singular, verbos y sustantivos entre otras formas.

El dataframe de citas no es necesario normalizarlo debido a que solo incluye campos de identificadores de artículos (es decir, los códigos WOS), por lo que ya están normalizados.

Por último, con los datos contenidos en los dataframes iniciales, se preparan los dataframes que finalmente se utilizarán para realizar la implementación de las propuestas. Para ello, en

primer lugar, se extrae el listado de artículos únicos que son objeto de estudio. Con ayuda de este listado, se itera en los dataframes iniciales para realizar los cálculos necesarios y conseguir un dataframe final en el que un artículo esté representado únicamente por una fila. Los dataframes finales resultantes son los siguientes:

- Dataframe de autores. Este dataframe está compuesto por tres columnas: identificador de artículo, número de autores y listado de autores. Es decir, solo es necesario el recuento del listado de autores para cada uno de los artículos y añadirlos al dataframe en una nueva columna, dando como resultado el dataframe final.

- Dataframe de citas. Está formado por tres columnas: identificador artículo, número total de citas recibidas y lista de los artículos que lo citan. El proceso para la obtención de este dataframe es el mismo que el realizado para la obtención del dataframe de autores, es decir, se realiza el recuento de la lista de artículos citantes de cada artículo y se añada al dataframe inicial en una nueva columna, obteniendo así el dataframe final.

- Dataframe de keywords. Este dataframe contiene, al igual que el resto, tres columnas: identificador de artículo, número de keywords de autor y lista de keywords. Al igual que en los dataframes anteriores, solo será necesario realizar el recuento de elementos de la lista de keywords del artículo y añadir el resultado en una tercera columna. De esta forma, se obtiene el nuevo dataframe compuesto por las tres columnas necesarias.

Obtenidos todos los dataframes, se comienza con la implementación de la metodología y la propuesta de ampliación.

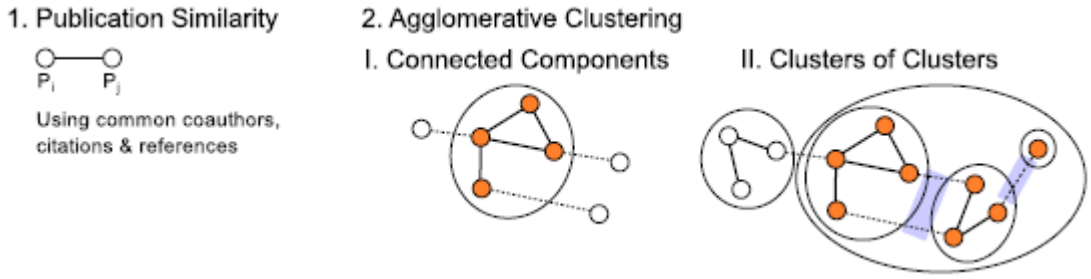
4.4. Aplicación de métodos de desambiguación

La metodología de desambiguación seleccionada para su implementación y ampliación, y posterior comparación, ha sido la propuesta por (Schulz et al., 2014). Esta metodología utiliza el clustering aglomerativo para intentar resolver el problema de la ambigüedad de los nombres de autor en las bases de datos bibliográficas.

4.4.1. Propuesta original realizada por Schulz et al. (2014)

Schulz et al. proponen un enfoque aglomerativo en 2 pasos (Figura 14) para resolver el problema de la ambigüedad de los autores en las bases de datos bibliográficas (Schulz et al., 2014).

Figura 14. *Metodología original propuesta por Schulz et al. (2014)*



Fuente: Schulz et al. (2014)

En primer lugar, el objetivo consiste en determinar si dos artículos comparten un mismo autor, es decir, si tienen como autor a la misma persona. Para ello calculamos la similitud por pares de todos los artículos del dataset utilizando como atributos las referencias, autores y artículos que citan utilizando la Ecuación 3. En esta ecuación de similaridad, primero se detectan los autores compartidos por ambos artículos, a continuación, se detectan posibles autocitas mediante el índice de inclusión, en tercer lugar, se detectan las referencias comunes y, por último, se detectan los artículos que citan ambas publicaciones.

Ecuación 3. *Cálculo de similaridad por pares de artículos.*

$$s_{ij} = \alpha_A \left(\frac{|A_i \cap A_j|}{\min(|A_i|, |A_j|)} \right) + \alpha_S \left(|p_i \cap R_j| + |p_j \cap R_i| \right)$$
$$+ \alpha_R \left(|R_i \cap R_j| \right) + \alpha_C \left(\frac{|C_i \cap C_j|}{\min(|C_i|, |C_j|)} \right).$$

Fuente: Schulz. et al. (2014)

A cada una de las variables descritas, se le asigna un peso distinto, siendo la variable citas la que más peso tiene y la variable referencias la que menor peso tiene. De este modo la variable autor tiene un peso de $\alpha_A = 0.6$, la variable autocitas tiene un peso de $\alpha_S = 0.75$, la

variable referencias tiene un peso de $\alpha_R = 0.15$ y la variable citas un peso de $\alpha_C = 1.1$. Debido a la imposibilidad de obtener las referencias esta variable se normaliza a 0. Además, aunque originalmente, para obtener la variable autocitas se utilizan las referencias, en este caso sí que es posible utilizar las citas para su obtención. Si un artículo a_1 referencia a un artículo a_2 quiere decir que el artículo a_2 recibe una cita del artículo a_1. Siguiendo esta misma lógica, si un artículo a_2 recibe una cita del artículo a_1 y comparten autores, es muy posible que se trate de una autocita.

Una vez obtenida la similaridad de los artículos, se comienza el proceso de clustering aglomerativo en dos pasos. El primer paso consiste en vincular cada par de artículos utilizando la similaridad calculada anteriormente, es decir, cuando la similaridad de un par de artículos supera un umbral $\beta1 = 0.6$, quiere decir que están relacionados y se etiquetan como grupo. El siguiente paso es calcular de nuevo la similaridad con los nuevos grupos creados en el paso anterior mediante la Ecuación 4, y se repite el proceso de agrupación. Para ello se suma la similaridad por cada par de artículos que superen un umbral, $\beta2 = 0.2$, y se divide entre el producto del número de publicaciones de cada clúster. Si la similitud entre dos grupos supera un umbral $\beta3 = 0.01$, quedan conectados.

Ecuación 4. *Cálculo de similaridad entre clústeres.*

$$S_{\gamma,\kappa} = \sum_{i \in \gamma j \in \kappa} \frac{s_{ij} \Theta(s_{ij} > \beta_2)}{|\gamma||\kappa|}.$$

Fuente: Schulz et al. (2014)

Los artículos restantes que no han podido agregarse a ningún grupo se unen al grupo en el que superen el umbral de similitud $\beta4 = 0.45$ con alguno de los artículos que componen el grupo.

El resultado final es N clústeres, en el que, idealmente, cada uno contiene los artículos publicados por un autor.

Los pasos del script para su aplicación están representados en el Algoritmo 3.

Algoritmo 3. *Algoritmo desarrollado para la aplicación de la propuesta original de Schulz et al. (2014)*

```
Paso 1: Importación de librerías necesarias

Paso 2: Conexión con la base de datos y creación del cursor

Paso 3: Definición de las querys necesarias para la importación
de los datos

Paso 4: Importación de los datos en dataframes

Paso 5: Definición de las funciones para la limpieza y
normalización de los datos

Paso 6: Aplicación de las funciones de limpieza y normalización a
los dataframes

Paso 7: Creación de los dataframes finales con los datos
necesarios para el cálculo de las variables necesarias para el
cálculo de similaridad

Paso 8: Unión de los dataframes finales en un único dataframe

Paso 9: Filtrar todos los artículos en los que aparezca la firma
del autor a desambiguar del dataframe resultante

Paso 10: Definición de las funciones necesarias para obtener las
variables de la función de similaridad

Paso 11: Inicio del bucle para el cálculo de la similaridad de
cada par de artículos asignando los pesos a las variables. Guardar
los resultados en la base de datos si estos no están ya añadidos

Paso 12: Definir las querys para la importación desde la base de
datos de la tabla con los cálculos de similaridad obtenidos en el
paso anterior, en un dataframe con la similaridad final por cada
par de artículos

Paso 13: Establecer la matriz de distancias

Paso 14: Agrupación de pares de artículos usando la matriz de
distancias

Paso 15: Cálculo de similaridad de los grupos obtenidos en el paso
anterior

Paso 16: Agrupación de los grupos obtenidos

Paso 17: Unión de artículos restantes al grupo con el que tengan
más similaridad

Paso 18: Obtención de clústeres final

Paso 19: Cierre del cursor y cierre de conexión con la base de
datos
```

El primer paso es la importación de las librerías necesarias para la implementación del código, es decir, las librerías que van a permitir conectar con la base de datos, la creación de dataframes, cálculo de variables y similaridad y agrupaciones.

Importadas las librerías, es necesario crear la conexión con la base de datos de la cual se importan los registros objeto de estudio. Se definen las credenciales, se conecta con la base de datos y se crea un cursor que va a permitir realizar las operaciones CRUD.

El siguiente paso es definir las consultas SQL que serán lanzadas a la base de datos para la selección de los datos y, posteriormente, ejecutarlas para importar los datos en forma de dataframes. En este caso, los atributos que tenemos que importar para la aplicación de esta metodología son las citas, las referencias y los autores.

A continuación, se definen las funciones para la limpieza y normalización de los dataframes anteriormente importados y se aplican dando como resultado otros dataframes con todos los registros normalizados.

Posteriormente, se crean los nuevos dataframes finales que, finalmente, se usarán para el cálculo de las variables necesarias para calcular la similaridad, y se unen en un único dataframe.

Una vez obtenido el dataframe final con todos los datos necesarios para el cálculo de la similaridad, se filtra de tal manera que se obtenga un dataframe con todos aquellos registros en los que aparezca la firma que se quiere desambiguar.

Además, se definen las funciones que permitirán la obtención del cálculo de cada una de las variables de la función de similaridad. En este caso las variables a calcular son los autores compartidos, autocitas y artículos que citan comunes.

Con todo esto, se está en disposición de iniciar un bucle que permita calcular por pares de artículos todas y cada una de las variables necesarias, asignarle el peso correspondiente a cada una y obtener el cálculo de similaridad final.

El funcionamiento del bucle es el siguiente. Para cada artículo único, se compara con todos y cada uno de resto de artículos del dataframe utilizado y se realiza el cálculo de las variables. A cada una de las variables resultantes se le asigna el peso definido por la metodología y se suman para obtener la similaridad total. Este cálculo es almacenado en un dataframe, el cual, una vez el bucle haya terminado de comparar dicho artículo con el resto, se almacena en una tabla de la base de datos para su posterior uso. En el caso de que los cálculos de este artículo ya estén incluidos en la base de datos, no se almacena y se pasa al siguiente.

Realizados todos los cálculos de similaridad, se definen las consultas que van a permitir la importación de todos estos ellos en un dataframe y se lanzan a la base de datos. Con este nuevo dataframe se establece la matriz de distancias y se realiza la primera agrupación,

obteniendo, de este modo, los primeros clústeres de artículos y quedando conectados todos aquellos que superen el umbral establecido.

Con estos primeros clústeres, se crea un nuevo dataframe que incluye las variables necesarias para realizar el cálculo de la similaridad por grupos. Las variables necesarias son la suma de similaridad por cada par de artículos de ambos grupos y el total de combinaciones posibles entre ambos grupos.

Con este dataframe se calcula la similaridad entre grupos, se establece de nuevo la matriz de distancias y se realiza la agrupación de los grupos que superen el umbral establecido. Aquellos artículos que no están unidos a ningún grupo se unen al grupo en el que superen el umbral establecido con alguno de los artículos que forman parte de este.

Finalmente, se obtienen los clústeres finales, los cuales, idealmente, contendrán los artículos publicados por un autor, y se cierran el cursor aún abierto y la conexión con la base de datos.

4.4.2. Propuesta ampliada: uso de keywords como atributo para desambiguación.

Aunque en la propuesta original realizada por Schulz et al. (2014) utilizan coautores, citas y referencias como atributos para la desambiguación de nombres de autor, en su estudio indican que es posible ampliar su propuesta con cualquier otro metadato.

Para estudiar si las keywords son útiles para la desambiguación, se ha decidido ampliar la propuesta original con este atributo y comparar posteriormente los resultados con la propuesta original para ver cuál arroja mejores resultados.

La elección de este atributo es por las siguientes razones: (1) un autor, normalmente, centra sus trabajos en un área concreta, por lo que las keywords que este utilizará serán similares en todos los artículos; (2) las personas tenemos patrones de redacción distintivos que pueden verse reflejados a la hora de redactar y establecer las keywords del artículo. Es por ello por lo que las keywords pueden ser de ayuda para la identificación de los artículos de un autor.

Para la aplicación de esta propuesta ampliada se sigue la misma lógica que en la propuesta original, pero añadiendo la variable keywords a la ecuación de similaridad entre pares de artículos. Es decir, calculamos la similaridad por pares de artículos utilizando como atributos las referencias, autores, artículos que citan y keywords de autor utilizando la Ecuación 5. En esta ecuación de similaridad, primero se detectan los autores compartidos por ambos

artículos, a continuación, se detectan posibles autocitas mediante el índice de inclusión, en tercer lugar, se detectan las referencias comunes, en cuarto lugar, se detectan los artículos que citan ambas publicaciones y, por último, se detectan las keywords de autor compartidas.

Ecuación 5. *Cálculo de similaridad por pares de artículos ampliada.*

$$Sij = \propto A\left(\frac{|Ai\ \cap Aj|}{\min(|Ai|,|Aj|)}\right) + \propto s\ (|pi\ \cap Rj| + |pj\ \cap Ri|) +$$
$$\propto R(|Ri\ \cup Rj|) + \propto C\left(\frac{|Ci\ \cap Cj|}{\min(|Ci|,|Cj|)}\right) + \propto K\left(\frac{|Ki\ \cap Kj|}{\min(|Ki|,|Kj|)}\right)$$

Al igual que en la propuesta original, se fijan unos pesos distintos a cada variable. Las variables utilizadas en la propuesta original mantienen los mismos pesos, es decir, la variable autor tiene un peso de $\alpha_A = 0.6$, la variable autocitas tiene un peso de $\alpha_S = 0.75$, la variable referencias tiene un peso de $\alpha_R = 0.15$ y la variable citas un peso de $\alpha_C = 1.1$. Además, se fija el peso de la variable α_K en 0.15, al igual que la variable referencias. Pero, al igual que en la propuesta original, como no ha sido posible obtener las referencias, se normaliza dicha variable a 0.

Calculados los pesos, se comienza con el clustering aglomerativo en dos pasos. En primer lugar, se vinculan los pares de artículos que superen un umbral β1 = 0.6. Obtenidos los grupos iniciales, se calcula de nuevo la similaridad haciendo uso de la Ecuación 4 (página 49), y se vuelve a realizar un proceso de agrupación. Tal y como se realiza en la propuesta original, se suma la similaridad por cada par de artículos que superen un umbral, β2 = 0.2, y se divide entre el producto del número de publicaciones de cada clúster. Quedarán conectados los grupos cuya similitud superen el umbral β3 = 0.01. Los artículos restantes se unirán al grupo con el que superen una similaridad de β4 = 0.45 con alguno de los artículos de dicho grupo.

Finalmente, se obtienen N clústeres que, idealmente, incluirán los artículos realizados por un autor.

Los pasos a seguir en este algoritmo están descritos, al igual que en la propuesta original, en el Algoritmo 3 (página 50). Aunque los cálculos sean distintos, la estructura del algoritmo es la misma. Es decir, la principal diferencia es la importación de un nuevo atributo, en este caso las keywords de autor, para el cálculo de una nueva variable (coincidencia de keywords) la cual se añade al cálculo de la similaridad.

4.5. Evaluación de resultados

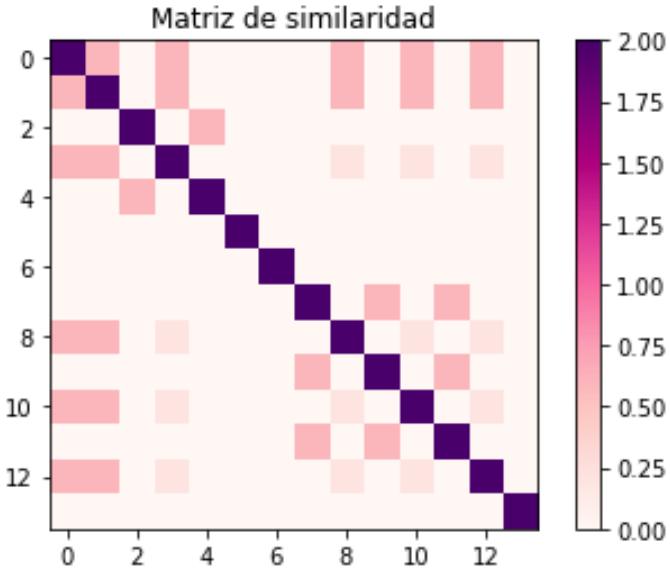
Debido a que el tamaño del dataset es de 60.663 artículos, calcular la similaridad para cada par de artículos conllevaría un total de 1.839.969.453 millones de combinaciones y un tiempo de cálculo de varios meses. Por ello, y debido a limitaciones de tiempo, para evaluar la efectividad de las propuestas y realizar una comparación de ellas, se ha acotado el dataset y se han seleccionado 5 autores cuyos nombres de firma son ambiguos. Los autores son: Brett Smith, Jonathan Adams, Yongjun Zhu, Carol V. Brown y Min Song. En las siguientes secciones se muestran y se comparan los resultados de ambas propuestas.

4.5.1. Brett Smith

4.5.1.1. Propuesta original

Para la aplicación de la propuesta se realiza un filtrado del dataset inicial, quedando únicamente los artículos en los que aparezca la firma Smith, B. En total, 14 artículos contienen dicha firma. Este resultado se almacena en un dataframe y se realiza el cálculo de la similaridad entre los artículos seleccionados, dando como resultado una matriz de similaridad de 14 filas y 14 columnas que puede verse en la Figura 15.

Figura 15. *Matriz de similaridad. Propuesta original sobre Smith, B.*



Obtenida esta matriz de similaridad, se realiza la primera agrupación asignando los umbrales establecidos, siendo 11 artículos los que superan el primer umbral de similaridad

con algún otro artículo y dando como resultado tres clústeres distintos como puede verse en la Figura 16. El primer clúster compuesto por 3 artículos, el segundo, al igual que el primero, por 3 artículos y el tercero por 5 artículos.

Figura 16. *Listado de artículos y clúster al que pertenecen. Propuesta original sobre Smith, B.*

Índice	data_index	cluster
0	WOS:000211993900008	2
1	WOS:000211997300003	1
2	WOS:000240630900009	2
3	WOS:000268719500005	0
4	WOS:000276851100012	1
5	WOS:000333322000008	2
6	WOS:000346646000008	0
7	WOS:000385616700004	1
8	WOS:000426063000004	0
9	WOS:000438385000006	0
10	WOS:000542246800009	0

Realizada la primera agrupación, se realiza la segunda agrupación en la que se unen aquellos clústeres que superen el umbral de similaridad establecido, dando lugar a la agrupación de los clústeres 1 y 2, y quedando de esta forma dos clústeres como puede observarse en la Figura 17, uno de ellos con 6 artículos y el otro compuesto por 5 artículos.

Figura 17. *Segunda agrupación de clústeres. Propuesta original sobre Smith, B.*

id_cluster	list_cluster	n_cluster	lis_articles
1	[0]	1	['WOS:000268719500005', 'WOS:000346646000008', 'WOS:000426063000004', 'WOS:000438385000006', 'WOS:000542246800009']
0	[1, 2]	2	['WOS:000211997300003', 'WOS:000276851100012', 'WOS:000385616700004', 'WOS:000211993900008', 'WOS:000240630900009', 'WOS:000333322000008']

A continuación, se realiza una tercera agrupación en la que se añade a cada clúster los artículos restantes que superen un umbral establecido. En este caso ninguno de los artículos restantes supera el umbral por lo que el resultado final son dos clústeres, uno

compuesto por 5 artículos y otro compuesto por 6 artículos, tal y como puede verse en la Figura 18.

Figura 18. *Resultado final. Propuesta original sobre Smith, B.*

id_cluster	n_articles	lis_articles
0	6	['WOS:000211997300003', 'WOS:000276851100012', 'WOS:000385616700004', 'WOS:000211993900008', 'WOS:000240630900009', 'WOS:000333322000008']
1	5	['WOS:000268719500005', 'WOS:000346646000008', 'WOS:000426063000004', 'WOS:000438385000006', 'WOS:000542246800009']

Previo al cálculo de la precisión y exhaustividad es necesario comprobar el número de artículos del dataset en los cuales participa Brett Smith como autor. Para ello, haciendo uso del perfil del autor en Web of Science, se obtiene el listado de identificadores de artículos asociados a su perfil y se realiza una consulta de dichos identificadores en la base de datos, dando como resultado 2 artículos.

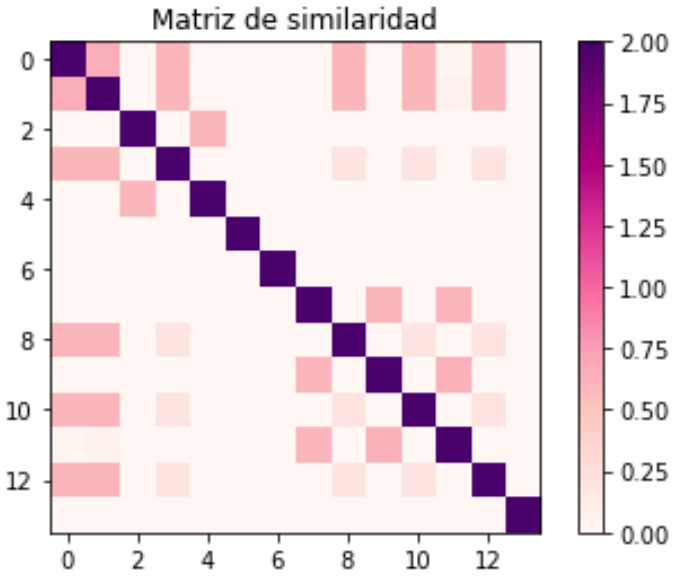
De los 5 artículos que componen el primer clúster, 2 de ellos corresponde a Brett Smith y, debido a que solo existen estos 2 artículos en todo el dataset de dicho autor, se observa que todos los artículos del autor han sido clasificados en el mismo clúster. Es decir, la exhaustividad o recall de esta propuesta para la clasificación de los artículos del autor Brett Smith es total, consiguiendo clasificar todos sus artículos en el mismo clúster (Recall = 2/2 = 1), pero su precisión es baja, ya que, junto a los dos artículos del autor, agrupa 3 artículos más que no pertenecen al mismo (Precisión = 2/5 = 0.4).

4.5.1.2. Propuesta ampliada

Para la aplicación de la propuesta ampliada se sigue el mismo proceso que en el caso anterior, pero con una diferencia, el valor de la similaridad entre pares de artículos incluye la información de las keywords de autor.

En primer lugar, se realiza el filtrado del dataset para obtener los artículos en los que aparezca Smith, B dando como resultado un dataset reducido de 14 artículos que lo almacenamos en forma de dataframe. A continuación, se realiza el cálculo de similaridad incluyendo las keywords de autor como metadato para la desambiguación y se obtiene la matriz de similaridad con 14 filas y 14 columnas y los valores de similaridad de cada par de artículos que se puede ver en la Figura 19.

Figura 19. *Matriz de similaridad. Propuesta ampliada sobre Smith, B.*



El siguiente paso es realizar la primera agrupación haciendo uso de la matriz de similaridad anterior, quedando 11 artículos que superan el umbral de similaridad con algún otro artículo. El resultado son tres clústeres distintos, como puede verse en la Figura 20. El primer clúster compuesto por 5 artículos y el segundo y tercer clúster por 3 artículos cada uno.

Figura 20. *Listado de artículos y clúster al que pertenece. Propuesta ampliada Smith, B.*

Índice	data_index	cluster
0	WOS:000211993900008	2
1	WOS:000211997300003	1
2	WOS:000240630900009	2
3	WOS:000268719500005	0
4	WOS:000276851100012	1
5	WOS:000333322000008	2
6	WOS:000346646000008	0
7	WOS:000385616700004	1
8	WOS:000426063000004	0
9	WOS:000438385000006	0
10	WOS:000542246800009	0

A continuación, se realiza la segunda agrupación en la que se unen aquellos clústeres que superen el umbral de similaridad establecido, dando lugar a la agrupación de los clústeres 0 y 2, quedando así dos clústeres como puede observarse en la Figura 21, uno de ellos con 8 artículos y el otro compuesto por 3 artículos

Figura 21. *Segunda agrupación de clústeres. Propuesta ampliada sobre Smith, B.*

id_cluster	list_cluster	n_cluster	lis_articles
0	[0, 2]	2	['WOS:000268719500005', 'WOS:000346646000008', 'WOS:000426063000004', 'WOS:000438385000006', 'WOS:000542246800009', 'WOS:000211993900008', 'WOS:000240630900009', 'WOS:000333322000008']
1	[1]	1	['WOS:000211997300003', 'WOS:000276851100012', 'WOS:000385616700004']

En la última agrupación, se añaden a cada clúster los artículos restantes que superen el umbral establecido, pero ninguno de ellos supera dicho umbral en este caso, por lo que el resultado final son dos clústeres, uno compuesto por 8 artículos y el otro por 3 artículos, como puede observarse en la Figura 22.

Figura 22. *Resultado final de la agrupación. Propuesta ampliada sobre Smith, B.*

id_cluster	n_articles	lis_articles
0	8	['WOS:000268719500005', 'WOS:000346646000008', 'WOS:000426063000004', 'WOS:000438385000006', 'WOS:000542246800009', 'WOS:000211993900008', 'WOS:000240630900009', 'WOS:000333322000008']
1	3	['WOS:000211997300003', 'WOS:000276851100012', 'WOS:000385616700004']

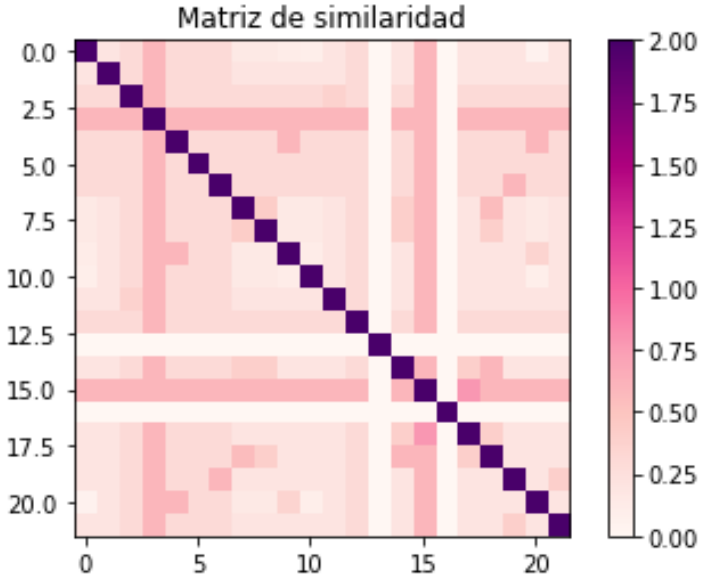
El primer clúster, compuesto por 8 artículos, contiene los 2 artículos existentes en el dataset completo que corresponden al autor Brett Smith. Esto nos da una exhaustividad o recall de 1, ya que esta propuesta ha sido capaz de agrupar todos los artículos del autor en el mismo clúster. Sin embargo, la precisión de esta propuesta es baja, ya que, junto a los 2 artículos del autor, ha agrupado 6 artículos más correspondientes a otros autores, es decir, una precisión de 0.25.

4.5.2. Jonathan Adams

4.5.2.1. Propuesta original

En primer lugar, se filtra el dataset inicial para obtener los artículos en los que aparezca la firma Adams, J, y almacenarlos en un dataframe. En total, 22 artículos contienen dicha firma con los cuales se realiza el cálculo de similaridad y se obtiene la matriz simétrica de 22 filas por 22 columnas que puede verse en la Figura 23.

Figura 23. *Matriz de similaridad. Propuesta original sobre Adams, J.*



Se realiza la primera agrupación, quedando 20 artículos que superan el umbral de similaridad con algún otro artículo, y se obtienen cuatro clústeres que pueden verse en la Figura 24. El primer clúster con 6 artículos, el segundo con 12 artículos, y el tercer y cuarto clúster con 1 artículo cada uno.

Figura 24. *Listado de artículos y clúster al que pertenecen. Propuesta original sobre Adams, J.*

Índice	data_index	cluster
0	WOS:000188229100011	1
1	WOS:000233359300002	1
2	WOS:000272195000003	1
3	WOS:000277204400003	3
4	WOS:000281681700009	1
5	WOS:000290586300004	1
6	WOS:000306547000017	1
7	WOS:000315242200012	1
8	WOS:000329319200034	1
9	WOS:000330931200001	0
10	WOS:000331263600019	1
11	WOS:000336258100003	0
12	WOS:000376273700015	1
13	WOS:000442670600024	0
14	WOS:000451074800005	2
15	WOS:000460550800017	1
16	WOS:000469058000017	1
17	WOS:000529892100025	0
18	WOS:000542420400001	0
19	WOS:000559623600001	0

A continuación, se realiza la segunda agrupación, uniendo los clústeres que superen un umbral establecido. De este modo, quedan unidos los clústeres 0, 1 y 3, dando como resultado final dos clústeres, uno de ellos formado por 1 artículo y otro formado por 19 artículos, tal y como puede verse en la Figura 25.

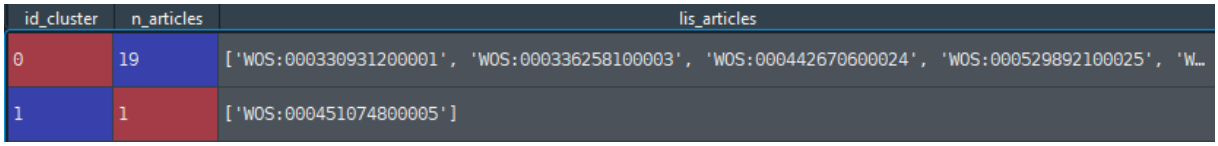
Figura 25. *Agrupación de clústeres. Propuesta original sobre Adams, J.*

_clust	it_clust	_cluste	lis_articles
0	[0, 1, 3]	3	['WOS:000330931200001', 'WOS:000336258100003', 'WOS:000442670600024', 'WOS:000529892100025', 'WOS:000542420400001', 'WOS:000559623600001', 'WOS:000281681700009', 'WOS:000290586300004', 'WOS:000306547000017', 'WOS:000315242200012', 'WOS:000329319200034', 'WOS:000331263600019', 'WOS:000376
1	[2]	1	['WOS:000451074800005']

Finalmente, se realiza una tercera agrupación de los artículos que han quedado independientes a los clústeres resultantes, pero como ninguno supera el umbral establecido,

no se realiza ninguna agrupación más. De este modo, el resultado final son dos clústeres, uno compuesto por 19 artículos y otro compuesto únicamente por 1 artículo, tal y como puede verse en la Figura 26.

Figura 26. *Resultado final de la agrupación. Propuesta original sobre Adams, J.*



id_cluster	n_articles	lis_articles
0	19	['WOS:000330931200001', 'WOS:000336258100003', 'WOS:000442670600024', 'WOS:000529892100025', 'W…
1	1	['WOS:000451074800005']

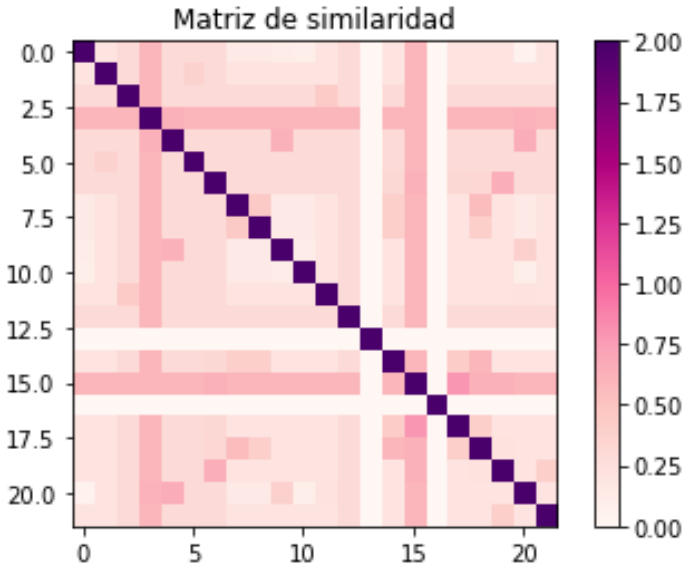
Para realizar los cálculos de precisión y exhaustividad, se hace uso perfil del autor en Web of Science para obtener el listado de los identificadores de artículos asociados a su perfil y realizar una consulta de dichos identificadores en la base de datos, dando como resultado 9 artículos del autor Jonathan Adams en el dataset original.

De los 9 artículos de dicho autor existentes en el dataset original, 8 de ellos se han clasificado en el primer clúster, el cual está compuesto por 19 artículos, y el restante está clasificado individualmente en el clúster 2. Esto da como resultado una precisión de 0.42 y una exhaustividad de 0.89. Esta propuesta, aunque consigue identificar todos los artículos del autor, no los agrupa todos correctamente.

4.5.2.2. Propuesta ampliada

Se realiza el filtrado del dataset original para obtener los artículos en los que firme Adams, J y se obtiene un dataset reducido de 22 artículos que es almacenado en un dataframe para realizar los cálculos de similaridad, en este caso incluyendo la variable keywords de autor. Realizados los cálculos se establece la matriz de similaridad de los artículos que se muestra en la Figura 27.

Figura 27. *Matriz de similaridad. Propuesta ampliada sobre Adams, J.*



A continuación, se realiza la primera agrupación con la matriz de similaridad anterior y se obtienen como resultado cuatro clústeres que se pueden ver en la Figura 28, que solo incluyen los 20 artículos que superan el umbral de similaridad establecido. Dos de ellos con 9 artículos cada uno y los dos restantes con un artículo cada uno.

Figura 28. *Listado de artículos y clúster al que pertenecen. Propuesta ampliada sobre Adams, J.*

Índice	data_index	cluster
0	WOS:000188229100011	2
1	WOS:000233359300002	2
2	WOS:000272195000003	2
3	WOS:000277204400003	3
4	WOS:000281681700009	2
5	WOS:000290586300004	2
6	WOS:000306547000017	2
7	WOS:000315242200012	2
8	WOS:000329319200034	2
9	WOS:000330931200001	0
10	WOS:000331263600019	0
11	WOS:000336258100003	0
12	WOS:000376273700015	0
13	WOS:000442670600024	0
14	WOS:000451074800005	1
15	WOS:000460550800017	0
16	WOS:000469058000017	2
17	WOS:000529892100025	0
18	WOS:000542420400001	0
19	WOS:000559623600001	0

Obtenidos los resultados de la primera agrupación, se realiza la segunda en la que se unen los clústeres que superen el umbral establecido. En este caso, quedan unidos los clústeres 0 y 2 en un clúster formado por 18 artículos, y los clústeres 1 y 3 en otro clúster formado por 2 artículos. Esto puede verse en la Figura 29.

Figura 29. *Agrupación de clústeres. Propuesta ampliada sobre Adams, J.*

id_cluster	list_cluster	n_cluster	lis_articles
1	[0, 2]	2	['WOS:000330931200001', 'WOS:000331263600019', 'WOS:000336258100003', 'WOS:000376273700…
0	[1, 3]	2	['WOS:000451074800005', 'WOS:000277204400003']

En la última agrupación, se añaden a cada clúster los artículos restantes que superen el umbral establecido, pero al no superarlo ninguno, el resultado final son dos clústeres, uno compuesto por 18 artículos y el otro por 2 artículos, como puede observarse en la Figura 30.

Figura 30. *Resultado final de la agrupación. Propuesta ampliada sobre Adams, J.*

id_cluster	n_articles	lis_articles
1	18	['WOS:000330931200001', 'WOS:000331263600019', 'WOS:000336258100003', 'WOS:000376273700015', 'W…
0	2	['WOS:000451074800005', 'WOS:000277204400003']

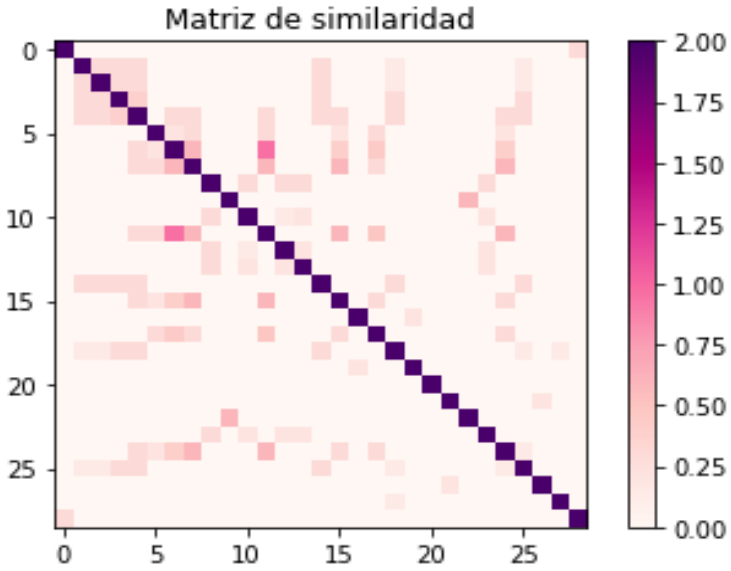
De los 18 artículos que componen el primer clúster, 8 corresponden al autor Jonathan Smith, por lo que esta propuesta tiene una precisión de 0.44. En cuanto a la exhaustividad, esta propuesta consigue clasificar 8 de los 9 artículos del autor en el mismo clúster, mientras que el artículo restante es clasificado en el otro clúster. De este modo, la exhaustividad de esta propuesta es de 0.89.

4.5.3. Yongjun Zhu

4.5.3.1. Propuesta original

Para comenzar la aplicación de la propuesta, se filtra el dataset para obtener aquellos artículos en los que aparezca la firma Zhu, Y, y se obtiene un dataframe el cual contiene 29 artículos. A continuación, se calcula la similaridad para cada par de artículos y se crea la matriz de similaridad con 29 filas por 29 columnas que puede verse en la Figura 31.

Figura 31. *Matriz de similaridad. Propuesta original sobre Zhu, Y.*



El siguiente paso, es realizar la primera agrupación, con la cual se obtienen 3 clústeres distintos que pueden verse en la Figura 32, utilizando los 7 artículos que superan el umbral de similaridad establecido con algún otro artículo. El primer clúster está formado por 4 artículos, el segundo clúster por 2 artículos y el tercer clúster por un único artículo.

Figura 32. *Listado de artículos y clúster al que pertenece. Propuesta original sobre Zhu, Y.*

Índice	data_index	cluster
0	WOS:000382914200026	0
1	WOS:000389548900015	1
2	WOS:000399871500001	0
3	WOS:000403857200023	2
4	WOS:000442737700032	0
5	WOS:000510245500003	1
6	WOS:000519275900001	0

Realizada la primera agrupación, se ejecuta la segunda agrupación que une los clústeres que superen el umbral establecido, uniendo en este caso el clúster 0 y el clúster 1. De este modo, el resultado de esta segunda agrupación son dos clústeres, uno compuesto por 6 artículos y otro compuesto por un solo artículo, tal y como puede observarse en la Figura 33.

Figura 33. *Agrupación de clústeres. Propuesta original sobre Zhu, Y.*

Índice	id_cluster	list_cluster	n_cluster	lis_articles
0	0	[0, 1]	2	['WOS:000382914200026', 'WOS:000399871500001', 'WOS:000442737700032', 'WOS:000519275900001', 'WOS:000389548900015', 'WOS:000510245500003']
1	1	[2]	1	['WOS:000403857200023']

Por último, se realiza una tercera agrupación de los artículos que han quedado independientes a los clústeres resultantes, obteniendo como resultado final un clúster con 6 artículos y otro clúster con 2 artículos, tal y como puede verse en la Figura 34.

Figura 34*. Resultado final de la agrupación. Propuesta original sobre Zhu, Y.*

id_cluster	n_articles	lis_articles
0	6	['WOS:000382914200026', 'WOS:000399871500001', 'WOS:000442737700032', 'WOS:000519275900001', 'WOS:000389548900015', 'WOS:000510245500003']
1	2	['WOS:000403857200023', 'WOS:000463255900001']

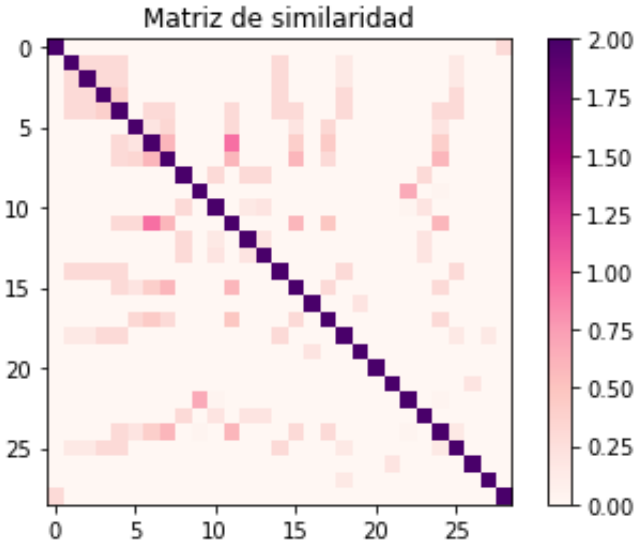
Para la evaluación de resultados, se realiza una consulta de los artículos del autor Yongjun Zhu en el dataset original, haciendo uso del listado de artículos asociados a su perfil en Web of Science. En total, 8 artículos del dataset inicial corresponden a dicho autor.

De estos 8 artículos, esta propuesta ha conseguido clasificar correctamente en el mismo clúster a 4 de ellos, mientras que, de los 4 restantes, 2 de ellos han sido clasificados incorrectamente en el otro clúster y otros 2 no han conseguido ser clasificados. Esto nos da una exhaustividad de 0.5 y una precisión de 0.67.

4.5.3.2. Propuesta ampliada

Se filtra el dataset inicial y se obtienen los artículos en los que aparezca Zhu, Y, obteniendo un dataset reducido de 29 artículos que es almacenado en un dataframe. Se calcula la similaridad de dichos artículos, en este caso añadiendo las keywords de autor, y se realiza la matriz de similaridad de 29 filas y 29 columnas que se puede ver en la Figura 35.

Figura 35. *Matriz de similaridad. Propuesta ampliada sobre Zhu, Y*



Una vez obtenida la matriz de similaridad, se realiza la primera agrupación cuyo resultado son 3 clústeres distintos con los 7 artículos que superan el umbral de similaridad establecido, uno compuesto por 4 artículos, otro compuesto por 2 artículos y otro compuesto por un solo artículo tal y como puede verse en la Figura 36.

Figura 36. *Listado de artículos y clúster al que pertenecen. Propuesta ampliada sobre Zhu, Y*

Índice	data_index	cluster
0	WOS:000382914200026	0
1	WOS:000389548900015	1
2	WOS:000399871500001	0
3	WOS:000403857200023	2
4	WOS:000442737700032	0
5	WOS:000510245500003	1
6	WOS:000519275900001	0

A continuación, se realiza la segunda agrupación en la que se unen los clústeres que superen el umbral establecido. En este caso, se unen los clústeres 1 y 2 en un clúster formado por 4 artículos, mientras que el clúster 0, formado por 3 artículos, queda independiente. Esto puede verse en la Figura 37.

Figura 37. *Agrupación de clústeres. Propuesta ampliada sobre Zhu, Y*

Índice	id_cluster	list_cluster	n_cluster	lis_articles
0	1	[0]	1	['WOS:000382914200026', 'WOS:000399871500001', 'WOS:000442737700032', 'WOS:000519275900001']
1	0	[1, 2]	2	['WOS:000389548900015', 'WOS:000510245500003', 'WOS:000403857200023']

Finalmente, se añaden a cada clúster los artículos restantes que superen el umbral establecido, obteniendo como resultado final dos clústeres, cada uno de ellos compuesto por 4 artículos como puede observarse en la Figura 38.

Figura 38. *Resultado final de la agrupación. Propuesta ampliada sobre Zhu, Y*

id_cluster	n_articles	lis_articles
1	4	['WOS:000382914200026', 'WOS:000399871500001', 'WOS:000442737700032', 'WOS:000519275900001']
0	4	['WOS:000389548900015', 'WOS:000510245500003', 'WOS:000403857200023', 'WOS:000463255900001']

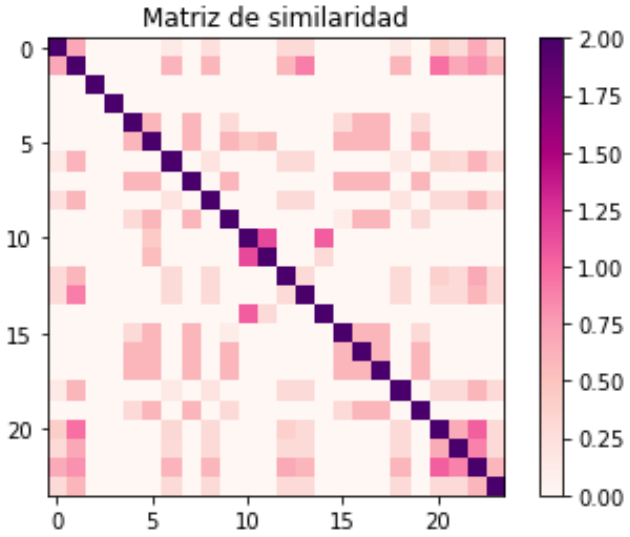
De los 8 artículos existentes en el dataset inicial que corresponden al autor Yongjun Zhu, esta propuesta ha conseguido identificar 6 artículos. Estos 6 artículos han sido divididos en dos clústeres distintos, clasificando en cada clúster 3 artículos del autor, de los 4 artículos que los forman. De este modo, esta propuesta arroja una precisión de 0.75, ya que 3 de cada 4 artículos son del autor, pero presenta una exhaustividad de 0.375, debido a que solo consigue clasificar correctamente 3 de los 8 artículos del autor.

4.5.4. Carol V. Brown

4.5.4.1. Propuesta original.

Se comienza con un filtrado del dataset original para obtener los artículos en los que aparezca la firma Brown, C y se obtiene un dataset reducido de 24 artículos que se almacena en un dataframe. Se realizan los cálculos de similaridad y se crea la matriz simétrica compuesta por 24 filas y 24 columnas, que puede verse en la Figura 39.

Figura 39. *Matriz de similaridad. Propuesta original sobre Brown, C.*



A continuación, se realiza la primera agrupación, quedando 22 artículos que superan el umbral de similaridad establecido con algún otro artículo. Como resultado se obtienen 6 clústeres distintos que pueden verse en la Figura 40. Los clústeres 0, 1, 3 y 4 están formados por 3 artículos, el clúster 2 por 9 artículos y el clúster 5 por un único artículo.

Figura 40. *Listado de artículos y clúster al que pertenece. Propuesta original sobre Brown, C*

Índice	data_index	cluster
0	WOS:000073867200004	0
1	WOS:000084192700007	3
2	WOS:000218097100002	0
3	WOS:000226327100002	3
4	WOS:000232639800001	0
5	WOS:000238431900007	2
6	WOS:000246807000003	0
7	WOS:000251730500003	0
8	WOS:000259771700004	3
9	WOS:000283037100003	2
10	WOS:000333022200003	0
11	WOS:000346857700005	4
12	WOS:000374150600007	1
13	WOS:000383781800011	0
14	WOS:000420628800007	4
15	WOS:000437520900005	1
16	WOS:000481174700001	2
17	WOS:000491348300005	0
18	WOS:A1994PX16200004	0
19	WOS:A1996WZ68700006	1
20	WOS:A1997WU72900005	5
21	WOS:A1997XH51900004	4

A continuación, se realiza una segunda agrupación que une los clústeres que superen el umbral establecido, obteniendo de este modo dos clústeres distintos. Uno de ellos formado por la unión de los clústeres 0, 1, 2, 3 y 4, compuesto por 21 artículos, y otro clúster correspondiente al clúster 5, formado por un solo artículo, tal y como puede verse en la Figura 41.

Figura 41. *Agrupación de clústeres. Propuesta original sobre Brown, C*

d_cluste	list_cluster	_cluste	
0	[0, 1, 2, 3, 4]	5	['WOS:000073867200004', 'WOS:000218097100002', 'WOS:000232639800001', 'WOS:000246 000437520900005', 'WOS:A1996WZ68700006', 'WOS:000238431900007', 'WOS:000283037100
1	[5]	1	['WOS:A1997WU72900005']

Finalmente, se unen a los clústeres aquellos artículos que no han sido unidos anteriormente y que superen el umbral establecido, pero, debido a que ninguno lo supera, los clústeres finalmente se quedan como la agrupación anterior, es decir, uno formado por 21 artículos y otro formado por un solo artículo (Figura 42).

Figura 42. *Resultado final de la agrupación. Propuesta original sobre Brown, C*

id_cluster	n_articles	lis_articles
0	21	['WOS:000374150600007', 'WOS:000437520900005', 'WOS:A1996WZ68700006', 'WOS:000346857700005', 'WOS…
1	1	['WOS:A1997WU72900005']

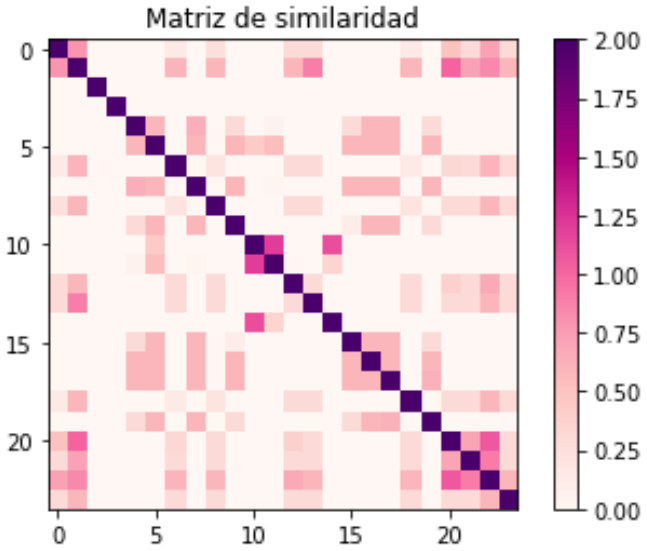
Para la evaluar los resultados de esta propuesta, se realiza una consulta sobre el dataset inicial de los artículos de la autora Carol V. Brown, haciendo uso del listado de artículos que se encuentran asociados a su perfil en Web of Science. En total, 11 artículos corresponden a dicha autora.

Esta propuesta ha identificado los 11 artículos de la autora, pero no los ha clasificado todos en el mismo clúster. De estos 11 artículos, 10 han sido clasificados correctamente en un clúster, mientras que el artículo restante ha sido clasificado erróneamente en otro clúster. Así pues, esta propuesta ha obtenido una exhaustividad de 0.91 a la hora de clasificar los artículos de esta autora. Sin embargo, ha añadido mucho ruido al clúster, clasificando erróneamente junto a los artículos de la autora un total de 11 artículos que no le corresponden. Por ello, su precisión ha sido de 0.48.

4.5.4.2. Propuesta ampliada

Para comenzar, se filtra el dataset inicial y se obtienen los artículos en los que aparece Brown, C como autor, obteniendo un dataset reducido de 24 artículos que se almacena en un dataframe para realizar los cálculos de similaridad incluyendo las keywords de autor como variable extra para la desambiguación. Hecho esto, se realiza la matriz de similaridad de 24 filas y 24 columnas que se puede ver en la Figura 43.

Figura 43. *Matriz de similaridad. Propuesta ampliada sobre Brown, C*



A continuación, se realiza la primera agrupación, obteniendo un total de 6 clústeres distintos con los 22 artículos que superan el umbral de similaridad establecido con algún otro artículo, tal y como puede verse en la Figura 44. El clúster 0 está formado por 9 artículos, los clústeres 1, 2, 3 y 4 están formados por 3 artículos cada uno y el clúster 5 está formado por un único artículo.

Figura 44. *Listado de artículos y clúster al que pertenecen. Propuesta ampliada sobre Brown, C*

Índice	data_index	cluster
0	WOS:000073867200004	0
1	WOS:000084192700007	3
2	WOS:000218097100002	0
3	WOS:000226327100002	3
4	WOS:000232639800001	0
5	WOS:000238431900007	2
6	WOS:000246807000003	0
7	WOS:000251730500003	0
8	WOS:000259771700004	3
9	WOS:000283037100003	2
10	WOS:000333022200003	0
11	WOS:000346857700005	4
12	WOS:000374150600007	1
13	WOS:000383781800011	0
14	WOS:000420628800007	4
15	WOS:000437520900005	1
16	WOS:000481174700001	2
17	WOS:000491348300005	0
18	WOS:A1994PX16200004	0
19	WOS:A1996WZ68700006	1
20	WOS:A1997WU72900005	5
21	WOS:A1997XH51900004	4

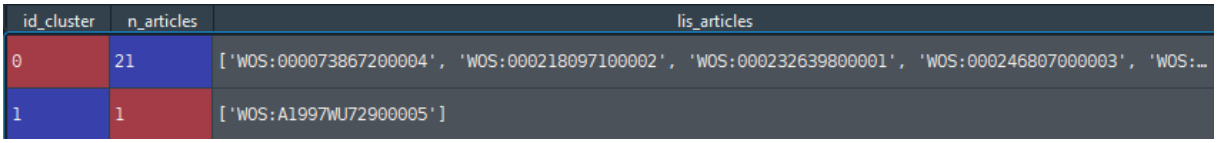
El siguiente paso agrupa aquellos clústeres que superan el umbral establecido, quedando unidos los clústeres 0, 1, 2, 3, y 4 en un único clúster formado por 21 artículos, mientras que el clúster 5, formado por un artículo queda independiente, tal y como puede verse en la Figura 45.

Figura 45. *Agrupación de clústeres. Propuesta ampliada sobre Brown, C*

id_cluster	list_cluster	n_cluster	lis_articles
0	[0, 1, 2, 3, 4]	5	['WOS:000073867200004', 'WOS:000218097100002', 'WOS:000232639800001', 'WOS:000246807000003', 'WOS:00025…
1	[5]	1	['WOS:A1997WU72900005']

Finalmente, a cada clúster se unen aquellos artículos que han quedado independientes y que superen un umbral establecido, pero como ninguno cumple los criterios, el resultado final es el mismo que el resultado de la agrupación anterior. Es decir, el resultado final son dos clústeres, uno formado por 21 artículos y otro formado por un único artículo, tal y como se muestra en la Figura 46.

Figura 46. *Resultado final de la agrupación. Propuesta ampliada sobre Brown, C*



id_cluster	n_articles	lis_articles
0	21	['WOS:000073867200004', 'WOS:000218097100002', 'WOS:000232639800001', 'WOS:000246807000003', 'WOS:…
1	1	['WOS:A1997WU72900005']

La evaluación de los resultados de esta propuesta ampliada comienza con una consulta sobre el dataset inicial de los artículos de la autora Carol V. Brown. Para ello, se hace uso del listado de artículos que se encuentran asociados a su perfil en Web of Science y se obtienen un total de 11 artículos correspondientes a dicha autora.

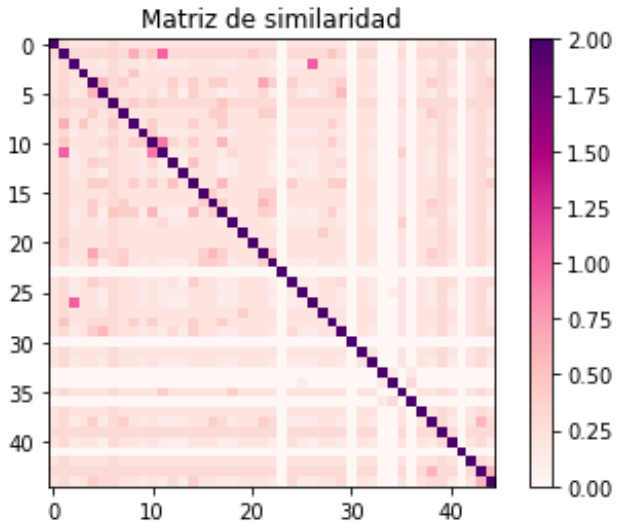
En esta propuesta ha conseguido identificar los 11 artículos de Carol V. Brown, y, a excepción de un artículo que ha sido clasificado erróneamente en otro clúster, el resto de los 10 artículos ha sido clasificados correctamente en un mismo clúster. Por ello, como 10 de 11 artículos ha sido clasificados correctamente, esta propuesta tiene una exhaustividad de 0.91. Por otro lado, aunque 10 de los 11 artículos hayan sido clasificados en el mismo clúster, en este mismo clúster se han incluido otros 11 artículos correspondientes a otros autores. Es por ello, que la precisión de esta propuesta es de 0.48, una precisión algo baja.

4.5.5. Min Song

4.5.5.1. Propuesta original.

Para la aplicación de la propuesta, se filtra el dataset inicial quedando solo aquellos artículos en los que aparece la firma Song, M, obteniendo un total de 45 artículos que son almacenados en un dataframe. A continuación, se realiza el cálculo de similaridad entre ellos y se obtiene la matriz de similaridad que está compuesta de 45 filas y 45 columnas como se puede ver en la Figura 47.

Figura 47. *Matriz de similaridad. Propuesta original sobre Song, M*



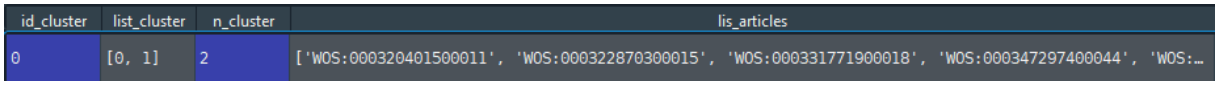
A continuación, se realiza la primera agrupación y se obtienen 2 clústeres distintos entre los que se dividen los 12 artículos que superan el umbral de similaridad establecido con alguno del resto de artículos, tal y como puede verse en la Figura 48. El clúster 0 está compuesto por 7 artículos y el clúster 1 está compuesto por 5 artículos.

Figura 48. *Listado de artículos y clúster al que pertenece. Propuesta original sobre Song, M*

Índice	data_index	cluster
0	WOS:000320401500011	0
1	WOS:000322870300015	0
2	WOS:000331771900018	0
3	WOS:000342228300019	1
4	WOS:000347297400044	0
5	WOS:000348324000003	1
6	WOS:000368338400014	0
7	WOS:000389548900005	1
8	WOS:000409715100001	1
9	WOS:000528948000004	0
10	WOS:000567839000003	1
11	WOS:000568825100001	0

Obtenida la primera agrupación, se realiza una segunda en la que se unen los clústeres 0 y 1 ya que superan el umbral establecido, obteniendo de este modo un solo clúster que incluye los 12 artículos como puede verse en la Figura 49.

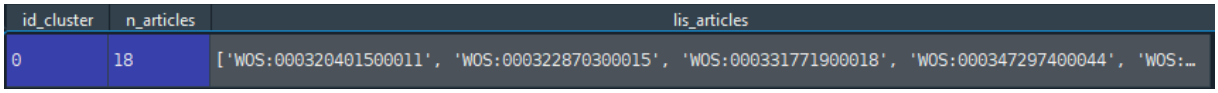
Figura 49. *Agrupación de clústeres. Propuesta original sobre Song, M*



id_cluster	list_cluster	n_cluster	lis_articles
0	[0, 1]	2	['WOS:000320401500011', 'WOS:000322870300015', 'WOS:000331771900018', 'WOS:000347297400044', 'WOS:…

Por último, se unen al clúster los artículos que previamente no se han unido a ningún clúster y que superen el umbral establecido, quedando finalmente un único clúster compuesto por 18 artículos, como puede verse en la Figura 50.

Figura 50. *Resultado final de la agrupación. Propuesta original sobre Song, M*



id_cluster	n_articles	lis_articles
0	18	['WOS:000320401500011', 'WOS:000322870300015', 'WOS:000331771900018', 'WOS:000347297400044', 'WOS:…

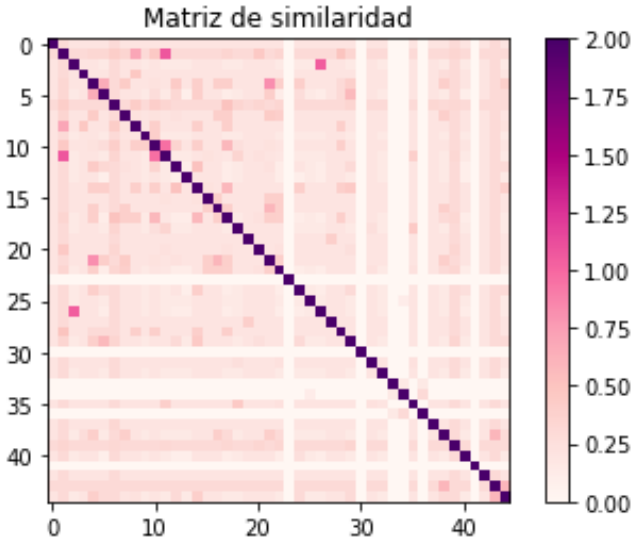
Para la evaluación de los resultados, haciendo uso del perfil del autor en Web of Science, se obtiene el listado de identificadores de artículos asociados a su perfil y se realiza una consulta de dichos identificadores en la base de datos, dando como resultado 29 artículos.

Esta propuesta aplicada sobre el autor Min Song, ha sido capaz de identificar 18 de los 29 artículos que contiene el dataset inicial del autor. Por ello, la exhaustividad de esta propuesta sobre este autor es de 0.62. Además, estos 18 artículos los ha conseguido clasificar en un mismo clúster sin añadir ningún artículo más de ningún otro autor, por lo que cuenta con una precisión de 1 a la hora de clasificar los artículos de este autor.

4.5.4.2. Propuesta ampliada

Se comienza filtrando el dataset inicial para obtener los artículos en los que firma el autor Song, M y se obtiene un dataset reducido de 45 artículos que es almacenado en un dataframe. Se calcula la similaridad de dichos artículos y se obtiene la matriz de similaridad que está compuesta por 45 filas y 45 columnas que puede verse en la Figura 51.

Figura 51. *Matriz de similaridad. Propuesta ampliada sobre Song, M*



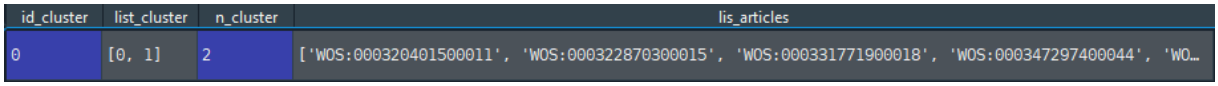
El siguiente paso es realizar la primera agrupación con la que se obtienen dos clústeres distintos con los 13 artículos que superan el umbral de similaridad establecido con alguno de los demás artículos. El clúster 0 compuesto por 7 artículos y el clúster 1 compuesto por 6 artículos como puede verse en la Figura 52.

Figura 52. *Listado de artículos y clúster al que pertenecen. Propuesta ampliada sobre Song, M*

Índice	data_index	cluster
0	WOS:000320401500011	0
1	WOS:000322870300015	0
2	WOS:000331771900018	0
3	WOS:000340479500006	1
4	WOS:000342228300019	1
5	WOS:000347297400044	0
6	WOS:000348324000003	1
7	WOS:000368338400014	0
8	WOS:000389548900005	1
9	WOS:000409715100001	1
10	WOS:000528948000004	0
11	WOS:000567839000003	1
12	WOS:000568825100001	0

A continuación, se agrupan los clústeres que superan el umbral de similaridad establecido, uniendo así ambos clústeres en un único clúster que contiene todos los artículos, como puede verse en la Figura 53.

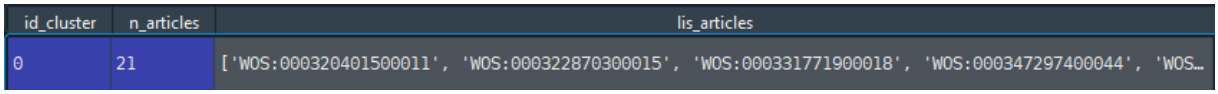
Figura 53. *Agrupación de clústeres. Propuesta ampliada sobre Song, M.*



id_cluster	list_cluster	n_cluster	lis_articles
0	[0, 1]	2	['WOS:000320401500011', 'WOS:000322870300015', 'WOS:000331771900018', 'WOS:000347297400044', 'WO…

Finalmente, al clúster obtenido en la agrupación anterior, se le unen los artículos que superan un umbral de similaridad establecido y que no han sido unidos con anterioridad a ningún clúster. Así pues, el resultado final es un único clúster formado por 21 artículos, tal y como puede verse en la Figura 54.

Figura 54. *Resultado final de la agrupación. Propuesta ampliada sobre Song, M*



id_cluster	n_articles	lis_articles
0	21	['WOS:000320401500011', 'WOS:000322870300015', 'WOS:000331771900018', 'WOS:000347297400044', 'WOS…

Para realizar la evaluación de los resultados de esta propuesta ampliada, es necesario realizar una consulta sobre el dataset inicial de los artículos del autor Min Song. Haciendo uso del listado de artículos asociados a su perfil de Web of Science, se realiza dicha consulta y se obtienen un total de 29 artículos de dicho autor en el dataset.

Se observa que de los 29 artículos del dataset correspondientes a este autor, esta propuesta ha sido capaz de identificar 21 de ellos, por lo que esta propuesta ampliada sobre este autor tiene una exhaustividad de 0.72. Además, los 21 artículos que componen el clúster son de dicho autor, por lo que su precisión con esta propuesta sobre este autor es de 1.

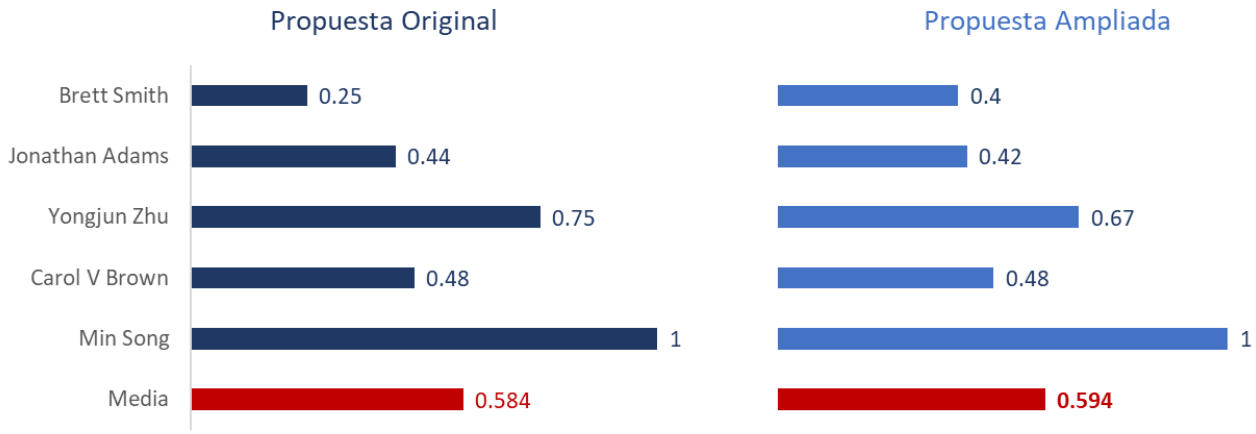
4.6. Discusión y análisis de resultados

Tras estudiar, implementar y experimentar con ambas propuestas sobre las firmas de los 5 autores elegidos, no se observan grandes diferencias entre los resultados arrojados.

En cuanto a la precisión, tal y como puede observarse en la Figura 55, la propuesta original obtiene una media de 0.584 mientras que la propuesta ampliada obtiene una media de precisión de 0.594. Esto quiere decir que la propuesta original es capaz de clasificar correctamente el 58% los elementos del clúster, o, dicho de otra forma, de todos los

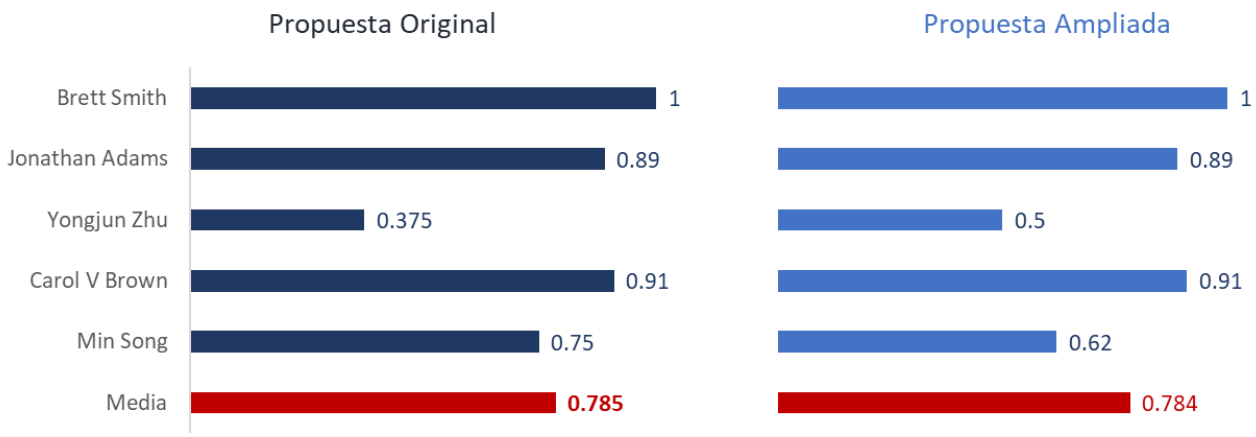
artículos del clúster, el 58% de ellos han pertenecen al autor que se pretende desambiguar. Sin embargo, la propuesta ampliada mediante el uso de las keywords ha sido capaz de mejorar levemente este resultado consiguiendo clasificar correctamente el 59% de los elementos del clúster, es decir, de todos los artículos del clúster, el 59% pertenecen al autor objeto de estudio. Cabe destacar, que, aunque la propuesta ampliada mejore levemente los resultados, en la mayoría de los autores ambas propuestas se comportan prácticamente igual a excepción de la desambiguación del autor de Brett Smith, en la que la propuesta ampliada destaca obteniendo una precisión del 15% mayor que la propuesta original. También hay que señalar que en la desambiguación de los autores Jonathan Adams y Yongjun Zhu, la propuesta original supera a la propuesta ampliada, pero esta mejora es prácticamente insignificante.

Figura 55. *Comparación de precisión entre la propuesta original y propuesta ampliada.*



Con respecto a la exhaustividad, la propuesta original obtiene una media de 0.785 y la propuesta ampliada una media de 0.784, por lo que no se observan diferencias en cuanto al uso de una u otra propuesta, tal y como puede observarse en la Figura 56. Esto quiere decir que ambas propuestas son capaces de identificar el 78% de los artículos del autor correctamente, unos datos bastante buenos. Bien es cierto que, aunque, en general no se observen diferencias entre ambas propuestas, a la hora de desambiguar a los autores Yongjun Zhu y Min Song se observan leves diferencias. La propuesta original es más exhaustiva a la hora de desambiguar al autor Min Song, con una diferencia del 13%, es decir, es capaz de identificar un 13% más de artículos del autor con respecto a la propuesta ampliada. Pero en el caso del autor Yongjun Zhu la propuesta ampliada obtiene mejores resultados, con un aumento de exhaustividad del 12,5% con respecto a la propuesta original, por lo que es capaz de identificar un 12,5% más de artículos del autor.

Figura 56. *Comparación de exhaustividad entre la propuesta original y propuesta ampliada.*



Con estos resultados, podemos observar que la propuesta ampliada con las keywords de autor, aunque suponga una leve mejoría en los resultados a la hora de desambiguar nombres de autor, no supone una mejora suficientemente significativa para afirmar que sea mejor que la original.

Entre las ventajas y desventajas que se pueden destacar de la propuesta ampliada frente a la original se encuentran las siguientes. Como principal desventaja, la propuesta ampliada disponiendo de un atributo más para la desambiguación de nombres de autor, alcanza unos resultados similares a la original, lo que supone un mayor procesamiento de datos para obtener prácticamente los mismos resultados.

Otra de las desventajas de la propuesta ampliada es el tratamiento de las keywords para su utilización como atributo desambiguador. A menudo, los autores combinan keywords en plural y singular, usan guiones, palabras compuestas entre otras variantes. Esto provoca que sea necesario prestar gran atención a la limpieza y normalización de este atributo. Aun así, es difícil conseguir que dicho atributo quede totalmente normalizado para su utilización.

Por otro lado, la principal ventaja es que, al usar un atributo más, al encontrar coincidencias entre artículos y etiquetarse dentro de un mismo autor, los resultados son más fiables puesto que se suma otro atributo más con coincidencias al conjunto. La propuesta ampliada producirá resultados más certeros en igualdad de condiciones.

Por último, otra de las ventajas de la propuesta ampliada es que, aunque bien es cierto que los resultados son similares y conlleva más procesamiento, sí que mejora levemente los resultados. Esto podría verse más claro con el análisis de más autores, pero no ha sido posible debido a limitaciones de tiempo.

5. Conclusiones y trabajo futuro

A continuación, se exponen las conclusiones que se han obtenido con la realización de este trabajo y las limitaciones que se han producido a lo largo de la realización de este. Además, se presentan una serie de posibles líneas de trabajo futuras que podrían suponer una mejora al trabajo ya realizado.

5.1. Conclusiones

A lo largo de este trabajo, se ha podido observar que la ambigüedad de las firmas de los autores en las publicaciones científicas supone un gran problema para la recuperación de información, análisis bibliométricos y la visibilidad de la ciencia debido a los fenómenos denominados silencio y ruido documental. Estos fenómenos provocan que no se recupere toda la literatura científica de un autor o, bien, que se recuperen documentos irrelevantes a la consulta. Entre los factores que favorecen la falta de normalización de las firmas de los autores se encuentra la sinonimia y polisemia de los nombres de autor, los malos hábitos de los autores al firmar sus trabajos, el incremento de publicaciones multidisciplinares y los fallos en los metadatos en las bases de datos bibliográficas. Así pues, se dan por resueltos el primer y segundo objetivos específicos.

Como solución a este problema, surgen los Identificadores Únicos de Autor que se comportan como un DNI para el autor el cual tendría asociado todas sus publicaciones, pero no ponen una solución real al problema por diferentes desventajas como pueden ser la falta del mantenimiento de su código por parte de los autores o la no adaptación a las nuevas tecnologías de los más veteranos. Surgen como alternativa, además, las técnicas de desambiguación de nombres de autor, las cuales tienen como objetivo identificar todas las publicaciones de un autor en un dataset. Existen muchas técnicas de desambiguación de diferentes tipologías como por ejemplo las basadas en aprendizaje supervisado, semi-supervisado o no supervisado. Estas últimas son unas de las más utilizadas puesto que no necesitan de unas clases preestablecidas. Mediante esta exploración de las diversas soluciones propuestas para el problema de la ambigüedad se da por resuelto el tercer objetivo específico.

El cuarto objetivo específico se cumple mediante la creación de un dataset de registros bibliográficos provenientes de la base de datos Web of Science. Este dataset contiene los metadatos necesarios para poder aplicar técnicas de desambiguación basadas en aprendizaje no supervisado principalmente. Actualmente, el dataset cuenta con un total de

60.663 registros, pero puede ser enriquecido posteriormente mediante la inclusión de más registros, ya sea de la propia Web of Science o de otras bases de datos como Scopus. Este dataset se pretende que sirva de ayuda para futuras investigaciones en este ámbito y que puede servir como punto de partida para investigadores que no cuenten con un dataset propio.

Con este dataset, se ha implementado la propuesta realizada por Schulz et al. (2014) para la desambiguación de los nombres de autor. Esta propuesta está basada en un clustering aglomerativo en dos pasos usando como atributos para la desambiguación los nombres de autor, las citas y las referencias, pero debido a la imposibilidad de la obtención de las referencias de Web of Science, este atributo ha tenido que ser suprimido. Además, se ha propuesto una ampliación de dicha propuesta que pretende mejorar los resultados de la propuesta original. A esta propuesta se ha añadido un nuevo atributo, las keywords de autor. Después de un análisis de los diferentes atributos ofrecidos por Web of Science, se ha llegado a la conclusión de que dichas keywords al depender de la redacción del autor y de la temática del trabajo, pueden estar muy relacionadas entre las diferentes publicaciones de un mismo autor. Con la implementación de ambas propuestas se dan por conseguidos el quinto y sexto objetivo específico.

Implementadas ambas propuestas, se ha realizado una comparación de las métricas de precisión y exhaustividad obtenidas por cada una de ellas aplicadas a un total de 5 autores distintos para evaluar la calidad de los agrupamientos. Se puede observar que ambas propuestas pueden ofrecer una solución óptima al problema, pero no se producen grandes diferencias entre el uso de la propuesta original y la propuesta ampliada con las keywords de autor, ofreciendo en ambos casos resultados similares. Es por ello, que no se puede afirmar que la ampliación de la propuesta con el uso de las keywords de autor mejore claramente los resultados. Así pues, se llega a la conclusión de que ambas propuestas son igualmente adecuadas para la resolución del problema. De esta forma, los objetivos específicos séptimo y octavo también se dan por conseguidos.

Con la consecución de todos los objetivos específicos, se puede dar por conseguido el objetivo general de este trabajo.

5.2. Limitaciones

Durante la realización de este trabajo, han sido varias las limitaciones que se han ido encontrando:

- La principal limitación que se ha encontrado ha sido la imposibilidad de obtener las referencias de los registros bibliográficos a través de la API oficial utilizada para la descarga. Durante la ejecución de los scripts, esta API ha sufrido una actualización al igual que la propia Web of Science, dejando de permitir la descarga de estas. Al disponer solo de unos pocos registros con las referencias, se ha optado por la eliminación de este atributo para la desambiguación para evitar así análisis poco fiables. Esto se debe a que si dos artículos de los que se tienen las referencias contienen alguna coincidencia entre ellas obtendrían mayor puntuación de similaridad que otros que contengan más coincidencias entre las referencias, pero cuyo valor sería 0 por no haber podido ser descargadas.

- Relacionada con la anterior limitación, al disponer de un atributo menos, los umbrales de similaridad establecidos para la agrupación de los artículos por el autor de la propuesta han tenido que ser modificados a medida que se ha ido experimentando con el conjunto de datos hasta llegar a un umbral óptimo que ofrece buenos resultados.

- Por último, otra de las limitaciones que se ha encontrado a la hora de implementar las propuestas, ha sido la falta de indicaciones por parte de los autores originales para la construcción de los algoritmos. La aplicación de técnicas de desambiguación es una tarea compleja, por lo que la falta de estas indicaciones a la hora de construir los algoritmos ha supuesto repetir en varias ocasiones desde el principio todo el proceso por algunos errores de comprensión del proceso. Además, esta falta de indicaciones supone un gran problema para el uso de estas propuestas por parte de los usuarios.

5.3. Líneas de trabajo futuro

Este trabajo es posible ampliarlo y mejorarlo con las siguientes líneas de trabajo futuras:

- Investigación y/o desarrollo de una nueva API que permita la descarga de las referencias de los registros de Web of Science. De esta forma se tendrían todas las herramientas para la obtención de un dataset completo al que poder aplicar las propuestas incluyendo dicho atributo.

- También se pueden formular varias propuestas de ampliación de la metodología añadiendo nuevos metadatos hasta encontrar la combinación de atributos más útiles para la desambiguación de nombres de autor. Por ejemplo, crear una nueva propuesta en la que se añada el atributo título. El título, al igual que las keywords de

autor, dependen de la redacción del propio autor del artículo. Las personas, por lo general, tienen unos patrones de redacción y siguen una temática similar en todos sus trabajos, así que es posible que, añadiendo el título a los atributos ya utilizados, podría ser una combinación muy útil para mejorar la desambiguación de los nombres de autor.

- Además, sería interesante ampliar el dataset con nuevas áreas de investigación de Web of Science y, también, con registros de otras bases de datos bibliográficas como pueden ser Scopus o Cochrane. De este modo, se conseguiría un gran dataset que podría utilizarse para realizar diferentes experimentaciones, no solo centradas en la desambiguación de nombres de autor.

- Debido al escaso tiempo disponible para la realización de este trabajo en la titulación, se ha tenido que reducir la aplicación de las propuestas a 5 autores. Lo ideal sería aumentar el número de autores analizados para, de este modo, tener una visión más clara de los resultados de ambas propuestas.

- Por último, sería interesante poder crear un software centrado en la desambiguación de nombres de autor que disponga de una interfaz gráfica y que ayude a la elaboración de estrategias de búsqueda de información. La implementación de las técnicas de desambiguación es muy compleja y requiere de conocimientos de programación, por lo que este software facilitaría el acceso al usuario de dichas técnicas. El funcionamiento sería el siguiente:

    1. El usuario posee un dataset con una serie de registros bibliográficos del cual quiere eliminar los problemas de ruido y silencio documental.

    2. Este usuario importa dicho dataset en el software y selecciona la técnica de desambiguación a utilizar y el autor del que quiere obtener los registros.

    3. El software analiza el dataset y le devuelve dos listados de registros distintos, uno con los documentos que contiene el dataset y que realmente pertenecen al autor y otro con los documentos del autor que no se encuentran en el dataset.

6. Bibliografía

Backes, T. (2018). Effective Unsupervised Author Disambiguation with Relative Frequencies. *arXiv:1808.04216 [cs, stat]*. https://doi.org/10.1145/3197026.3197036

Baiget, T., Rodríguez-Gairín, J.-M., Peset, F., Subirats, I., & Ferrer-Sapena, A. (2007). Normalización de la información: La aportación de IraLIS. *El Profesional de la Informacion*, *16*(6), 636-643. https://doi.org/10.3145/epi.2007.nov.10

Canteros, A., Zamudio, E., & Kuna, H. D. (2018, septiembre). *Desambiguación de autores para un sistema de recuperación de expertos en un contexto académico*. XIX Simposio Argentino de Inteligencia Artificial (ASAI) - JAIIO 47 (CABA, 2018). http://sedici.unlp.edu.ar/handle/10915/70697

Carvalho, A. P. de, Ferreira, A. A., Laender, A. H. F., & Gonçalves, M. A. (2011). Incremental unsupervised name disambiguation in cleaned digital libraries. *Journal of Information and Data Management*, *2*(3), 289-304.

Clarivate Analytics. (2020, marzo 15). *Colección principal de Web of Science—Ayuda*. http://images.webofknowledge.com/WOKRS522_2R1/help/es_LA/WOS/contents.html

Cota, R. G., Gonçalves, M. A., & Laender, A. H. F. (2007). *A Heuristic-based Hierarchical Clustering Method for Author Name Disambiguation in Digital Libraries*. XXII Simposio Brasileiro de Banco de Dados, Brasil.

DOMO. (2019). *Domo Resource—Data Never Sleeps 7.0*. DOMO. https://www.domo.com/learn/data-never-sleeps-7

EC3 - Grupo de Evaluación de la Ciencia y la Comunicación Científica. (2016). *Co-author Index*. Co-author Index. http://www.coauthorindex.info/

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, *41*(2), 15-26. https://doi.org/10.1145/2350036.2350040

Ferreiro, L. (1993). *Bibliometría: (Análisis bivariante)* (Vol. 16). Eypasa.

Ferrer, N. F., & Pérez-Montoro, M. (2011). *Búsqueda y recuperación de la información*. Editorial UOC.

Gasparyan, A. Y., Ayvazyan, L., & Kitas, G. D. (2013). Multidisciplinary Bibliographic Databases. *Journal of Korean Medical Science*, *28*(9), 1270. https://doi.org/10.3346/jkms.2013.28.9.1270

Giles, C. L., Zha, H., & Han, H. (2005). Name disambiguation in author citations using a K-way spectral clustering method. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*, 334-343. https://doi.org/10.1145/1065385.1065462

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259-264. https://doi.org/10.1087/20120404

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review*, *32*, e22. https://doi.org/10.1017/S0269888917000182

In-Su, K., Seung-Hoon, N., Seungwoo, L., Pyung, K., Hanmin, J., Won-Kyung, S., & Jong-Hyeok, L. (2009). On co-authorship for author disambiguation | Elsevier Enhanced Reader. *Information Processing and Management*, *45*, 84-97. https://doi.org/10.1016/j.ipm.2008.06.006

La Razón. (2021, febrero 4). *Apellidos más comunes en cada país del mundo*. La Razón. https://www.larazon.es/sociedad/20210204/xqui2rlkizamhj2vthjgkvxryy.html

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

Marqués, M. (2011). *Bases de datos*. Universitat Jaume I, Servei de Comunicació i Publicacions.

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, *7*(4), 767-773. https://doi.org/10.1016/j.joi.2013.06.006

NetCredit. (2019, noviembre 24). The Most Common Last Name in Every Country. *NetCredit Blog*. https://www.netcredit.com/blog/most-common-name-country/

ORCID. (s. f.). *ORCID Registry*. ORCID. Recuperado 26 de abril de 2021, de https://info.orcid.org/documentation/features/orcid-registry/

phpMyAdmin. (s. f.). *PhpMyAdmin*. PhpMyAdmin: Bringing MySQL to the Web. Recuperado 8 de junio de 2021, de https://www.phpmyadmin.net/

Pinto, M. (2018). *Búsqueda y Recuperación de Información*. Electronic Content Management Skills. http://www.mariapinto.es/e-coms/busqueda-y-recuperacion-de-informacion/

Schulz, C., Mazloumian, A., Petersen, A. M., Penner, O., & Helbing, D. (2014). Exploiting citation networks for large-scale author name disambiguation. *EPJ Data Science*, *3*(1), 11. https://doi.org/10.1140/epjds/s13688-014-0011-3

Shoaib, M., Daud, A., & Amjad, T. (2020). *Author Name Disambiguation in Bibliographic Databases: A Survey*. 24.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*(1), 1-43. https://doi.org/10.1002/aris.2009.1440430113

Sobrido, M., Gutiérrez, U., & González, C. (2016). De la normalización de la firma científica a la identificación digital del autor. *Index de Enfermería*, *25*(1-2), 56-59.

Somoza, M. (2015). *Búsqueda y recuperación de información en bases de datos de bibliografía científica*. Ediciones Trea. https://bv.unir.net:2769/es/ereader/unir/117492

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

Tekles, A., & Bornmann, L. (2019). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *arXiv:1904.12746 [cs]*. http://arxiv.org/abs/1904.12746