Universidad de las Ciencias Informáticas

Facultad 3

Grupo de Investigación de Web Semántica

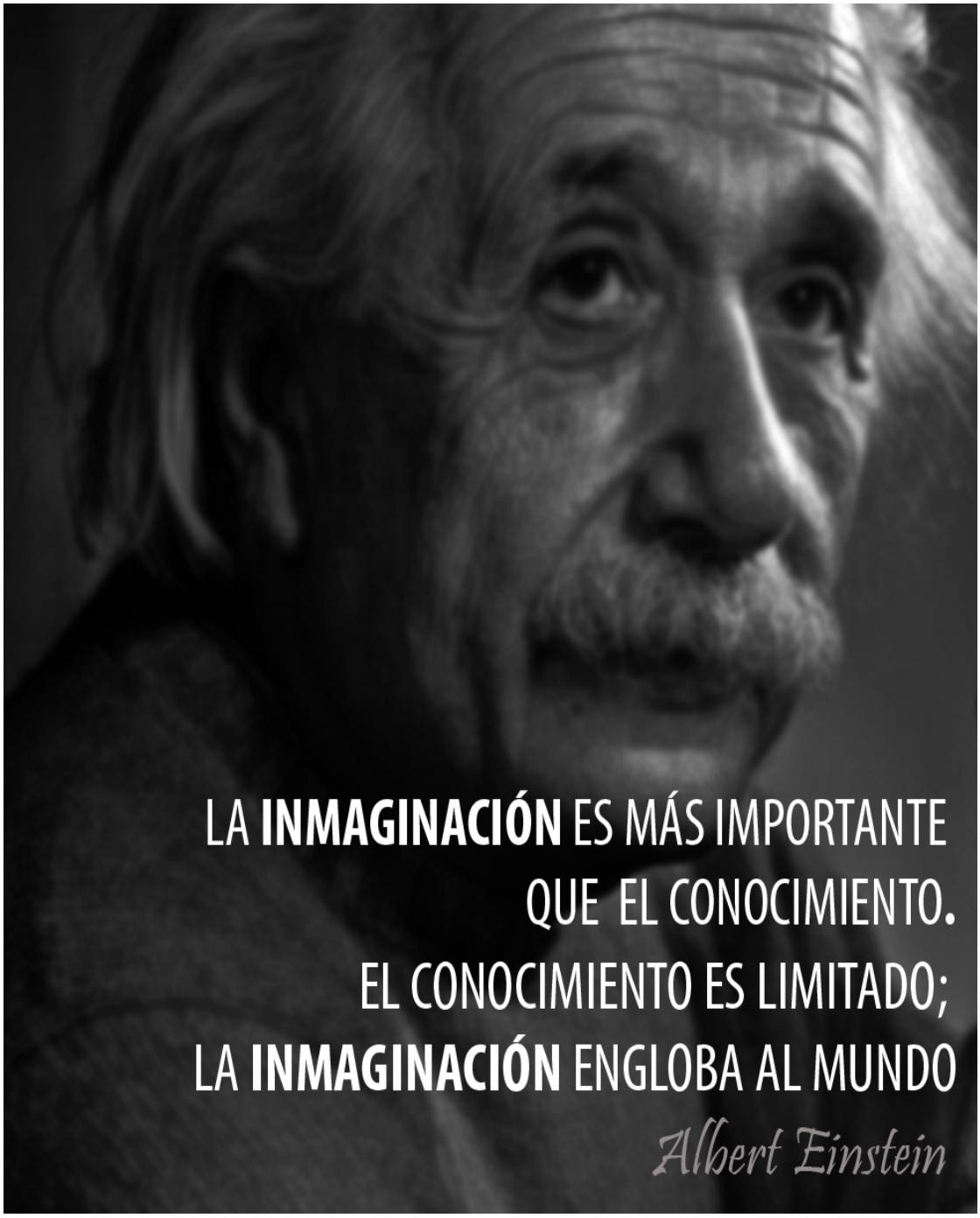
Desambiguación del nombre de los autores en metadatos bibliográficos usando técnicas de agrupamiento

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Autor: Luis Enrique Alonso Sierra

Tutor: Ing. Yusniel Hidalgo Delgado

Ciudad de la Habana, 6 de Junio de 2014


LA **INMAGINACIÓN** ES MÁS IMPORTANTE
QUE EL CONOCIMIENTO.
EL CONOCIMIENTO ES LIMITADO;
LA **INMAGINACIÓN** ENGLOBA AL MUNDO
Albert Einstein

Declaración de autoría

Declaro ser autor de la presente tesis y reconozco a la UCI los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____

Luis Enrique Alonso Sierra

_____

Profesor. Departamento de Programación, Facultad 3

Ing. Yusniel Hidalgo Delgado

Dedicatoria

A mi tía que se que donde quiera que esté está orgullosa de la persona en que me he convertido. Por enseñarme la diferencia entre el bien y el mal. Gracias por estar ahí para mí.

A mi novia por ser un punto de confianza donde apoyarme en los momentos difíciles. Gracias por ser como eres y por convertirme en la persona que soy.

Luis Enrique Alonso Sierra

Agradecimientos

Especialmente a mis padres por confiar y creer siempre en mí. Por cuidarme todos estos años y enseñarme las realidades de la vida.

A mi novia por compartir conmigo los momentos difíciles y alegres en estos 5 años.

A Abi por ser una de las personas que más admiro en la vida, por su tenacidad y empeño. Gracias por ser como eres.

A mi hermano por enseñarme cosas de la vida que yo debería enseñarle a él. Por ser la alegría de la familia.

A toda mi familia, a abuelo Ventura, abuela Niña, abuelo Sierra, a mi hermano Michel y Li Daniel, a Ivis, a tía Migdalia, a Adrián por ser parte de mi vida y darme cada uno un pedasito de su corazón y cariño, gracias por ser mi familia.

A Ñeca y Raúl por apoyarme y aconsejarme durante todos estos años.

A Yusniel por guiarme por los pasos del mundo académico, sin su apoyo no estaría en esta posición.

Al grupo de investigación de web semántica por las críticas.

A todos los que de una forma u otra han hecho posible mi tránsito por la universidad.

A la revolución por permitirme estudiar la carrera que me gusta.

Luis Enrique Alonso Sierra

Resumen

La ambigüedad en el nombre de los autores en los repositorios digitales, es un problema que afecta la calidad de los metadatos bibliográficos que estos almacenan. Se refiere a la posibilidad real de representar el nombre de los autores de dos formas: (1) nombres de autores sintácticamente diferentes pero que se refieren a la misma persona y (2) nombres de autores sintácticamente iguales pero que no se refieren a la misma persona. En esta investigación se propone un método para desambiguar el nombre de los autores en metadatos bibliográficos basado en la combinación de agrupamientos. El método propuesto utiliza algunos elementos de los metadatos bibliográficos presentes en los repositorios digitales para establecer las relaciones entre los autores. Estos elementos son: autores, co-autores, afiliación, títulos de las publicaciones y lugares de publicación. Los resultados obtenidos demuestran la viabilidad del método propuesto así como el aumento en la unicidad del nombre de los autores.

Índice General

Índice de figuras

Índice de tablas

Introducción

El desarrollo de las nuevas Tecnologías de la Información y las Comunicaciones (TIC) han convertido a internet en una fuente de consulta importante para los investigadores. Ello ha determinado la necesidad de encontrar formas de almacenar y publicar el conocimiento científico en espacios que cumplan con normas rigurosas para su difusión.

Existen numerosas formas de almacenar y divulgar el conocimiento científico. Entre las más conocidas se encuentran las bases de datos referenciadas, los repositorios digitales y las revistas científicas. Estas últimas se refieren a publicaciones periódicas en las que se intenta recoger el progreso de la ciencia, entre otras cosas, incluyendo informes sobre las nuevas investigaciones.

El desarrollo de la ciencia ha tomado un paso acelerado. Las tecnologías hoy, son usadas como herramientas para el desarrollo de nuevas tecnologías. Propiciado que el volumen de información científica crezca de forma vertiginosa, provocando que los investigadores se vean forzados a emplear gran cantidad de tiempo y esfuerzo en encontrar la información necesaria para sus investigaciones.

En este sentido, la bibliometría ayuda a los investigadores a reducir el tiempo y esfuerzo asociado a la búsqueda de información. Utiliza los metadatos de la producción científica de una determinada área del conocimiento y es capaz de determinar información importante para los investigadores. El uso de la bibliometría requiere que los metadatos empleados posean la mayor calidad posible para que los resultados arrojados en los análisis sean los esperados.

Un elemento importante dentro de los metadatos bibliográficos es el nombre de los autores. En ocasiones los investigadores utilizan este elemento como criterio para las búsquedas de información. Si los nombres de los autores no cuentan con la calidad requerida, es decir, no están representados correctamente en los

metadatos bibliográficos, entonces los resultados de las búsquedas no son correctos.

La identificación de un investigador dentro de los metadatos bibliográficos es una tarea compleja debido a las múltiples formas de representar un nombre. Generalmente se usan la inicial del nombre y el apellido, esto no traería grandes problemas debido a que los nombres en idiomas como el inglés, francés, alemán e italiano suelen estar formados por un nombre y un solo apellido. Las particularidades del idioma español hacen de este trabajo un proceso complejo. Los nombres en el idioma español generalmente están formados por un nombre (en ocasiones dos) y dos apellidos, propiciando que la cantidad de variantes de representación en una base de datos sea alta. Por ejemplo, en el caso de dos nombres, con dos apellidos, la cantidad de combinaciones posibles es 4^4 y en el caso de un nombre con dos apellidos es 3^3. Si se utilizaran otras formas como Y. H. Delgado o Y. Hidalgo-Delgado, entonces el número de posibles combinaciones es mayor. Un nombre estándar en el idioma español puede ser *"Rafael Bello Pérez"*, alguna de las variantes en que puede aparecer en una base de datos son:

- Bello Pérez, Rafael
- Bello Perez, Rafael
- Bello, Rafael
- Bellopérez, Rafael
- Pérez, R.B.

Como se puede apreciar no solo dificulta el trabajo las posibles variantes en que puede aparecer el nombre, sino que también pueden ocurrir errores en la escritura. Los resultados en las búsquedas de información poseerían un mayor grado de error. Esta situación es conocida como ambigüedad en el nombre de los autores.

El problema planteado está reflejado en los sistemas que manejan metadatos bibliográficos. No es hasta principios del este siglo donde se evidencian las primeras aproximaciones desarrolladas para dar solución a la ambigüedad en el nombre de los autores. Algunas de las revistas que han desarrollado proyectos con el fin de resolver la ambigüedad en el nombre de los autores son: CiteSeer[1] y MedLine[2]. Durante el período de tiempo que comprende desde el año 2000 hasta la actualidad (2014) las soluciones planteadas por los investigadores han tomado diferentes cursos y han arrojado resultados satisfactorios en muchas ocasiones.

En el Grupo de Investigación de Web Semántica[3] de la Universidad de las Ciencias Informáticas se desarrolla el proyecto **DBJournal**. Tiene como objetivo la recolección, publicación y consumo de metadatos bibliográficos como datos enlazados. Durante el proceso de recolección de metadatos se identificó el problema mencionado. Es necesario afirmar que el protocolo de recolección de metadatos soportado por el proyecto DBJournal es el OAI-PMH[1].

Los datos enlazados se refieren a un conjunto de buenas prácticas para la publicación y enlazados de datos estructurados en la web[2], poseen un conjunto de principios por los cuales es necesario regirse para la correcta publicación de datos estructurados en la web:

- **Identificar**: Utilizar URIs[4] para identificar cada recurso en la web.
- **Publicar**: Publicar cada recurso en una URI basada en HTTP[5] de modo que puedan ser fácilmente localizados y consultados.
- **Describir**: Proporcionar información detallada, útil o extra acerca de cada recurso publicado en la web.
- **Enlazar**: Enlazar los recursos publicados con otras URIs relacionadas, de forma que se potencie el descubrimiento de la información sobre la web.

[1] *http://citeseerx.ist.psu.edu*
[2] *http://www.medline.com*
[3] *http://gws-uci.blogspot.com/*
[4] *Identificador Universal de Recursos*
[5] *Lenguaje de Marcado de Hipertexto*

La ambigüedad en el nombre de los autores introduce ruido en el proceso de publicación de los metadatos bibliográficos como datos enlazados. Publicar un autor (recurso) con ambigüedad en su representación implica publicar un recurso varias veces, lo que trae consigo pérdida en la calidad de los recursos publicados en la web de los datos. Esto provoca que se afecte el rendimiento del entrelazado de los recursos en el proceso de publicación, así como su descubrimiento por otros sistemas informáticos. También la localización de los recursos en la web se afecta debido a que un recurso está publicado en diferentes URLs[6] y las relaciones que se establecen entre los autores y otros recursos (Artículos, Webs, entre otros) pueden ser excluyentes.

La ambigüedad en el nombre de los autores afecta diferentes factores relacionados con los metadatos bibliográficos, entre los que se mencionan en este documento se encuentran:

- Problemas en la recuperación de información sobre los autores.
- Dificultad en la realización de estudios bibliográficos.
- Dificultad en la publicación de metadatos bibliográficos como datos enlazados.

Los problemas antes mencionados poseen un origen común: **unicidad** en el nombre de los autores. Según la Real Academia de la Lengua Española[7], unicidad se refiere a la cualidad de ser único, irrepetible, solo, singular. A partir de lo anterior se puede formalizar unicidad en el nombre de los autores:

Definición. *Representación única, irrepetible y singular del nombre de los autores en una colección de metadatos bibliográficos.*

De acuerdo a la definición de unicidad en el nombre de los autores se puede expresar un relación de proporcionalidad inversa entre la **unicidad y la ambigüedad en el nombre de los autores**. Mientras mayor

[6] *Localizador Universal de Recurso*

[7] *http://www.rae.es*

es la ambigüedad en el nombre de los autores, menor es la unicidad de los registros de autores en los metadatos bibliográficos.

Luego de la definición de unicidad en el nombre de los autores es posible enunciar el **problema a resolver** de la investigación: la ambigüedad en el almacenamiento de los nombres de los autores dentro de los metadatos bibliográficos afecta su unicidad en el contexto del proyecto DBJournal.

A partir de lo anterior se puede definir como **Objeto de estudio** de la investigación: desambiguación del nombre de los autores en metadatos bibliográficos. Dentro del objeto de estudio se encuentra el **Campo de acción** de la misma: algoritmos para la desambiguación del nombre de los autores en metadatos bibliográficos.

El **Objetivo general** que se persigue es el siguiente: desambiguar el nombre de los autores en metadatos bibliográficos a través de un algoritmo utilizando técnicas de minería de datos para garantizar la unicidad de los mismos en el contexto del proyecto DBJournal.

Para lograr del objetivo general se definen los siguientes **Objetivos específicos**:

- Elaborar el marco teórico y el estado del arte del objeto de estudio de la investigación mediante el análisis bibliográfico documental para identificar tendencias y adoptar posiciones al respecto.
- Diseñar un algoritmo para la desambiguación del nombre de los autores mediante el uso de combinación de algoritmos de agrupamiento.
- Implementar el algoritmo para la desambiguación del nombre de los autores en metadatos bibliográficos.
- Validar los resultados del algoritmo de desambiguación del nombre de los autores en metadatos bibliográficos a través de un diseño experimental.

Lo anterior conduce a la siguiente **Idea a defender**: Si se desarrolla un algoritmo para la desambiguación del nombre de los autores en los metadatos bibliográficos entonces se aumentará la unicidad en los mismos.

Para la realización de la investigación se emplearon los **métodos científicos** siguientes:

Teóricos:

- Histórico-lógico
- Hipotético-deductivo
- Analítico-sintético

Se enfocan las problemáticas asociadas a la desambiguación del nombre de los autores en metadatos bibliográficos desde un enfoque histórico-lógico. El inicio de la investigación estuvo caracterizado por la realización de un estudio del estado del arte donde se exponen las principales deficiencias y fortalezas de las soluciones existentes en la literatura.

Durante la investigación se siguió un método hipotético-deductivo debido a que partiendo del planteamiento del problema concreto se dedujeron objetivos específicos e idea a defender que permitieron solucionar el problema en cuestión usando los métodos adecuados.

El método analítico-sintético se siguió para identificar las partes que componen el problema de la ambigüedad en el nombre de los autores. Definiéndose las causas y los efectos que este provoca.

Empíricos:

- Experimentación
- Medición

El método de experimentación se utiliza para la realización de los experimentos diseñados con el objetivo de validar la propuesta de solución.

Por su parte el método de medición se utiliza en el cálculo de la efectividad de la propuesta de solución y de la calidad de las respuestas finales.

Para garantizar la validez de los resultados se precisa disponer de un conjunto de datos confiables y representativos de metadatos bibliográficos que permitan afirmar que los resultados obtenidos son aplicables en los escenarios previstos. Por lo anterior se seleccionó como **Población** los metadatos bibliográficos de las revistas científicas latinoamericanas que diseminan sus metadatos a través del protocolo **OAI-PMH**. Por su parte la **Muestra** seleccionada son las revistas científicas cubanas que diseminan sus metadatos bibliográficos a través del protocolo **OAI-PMH**.

La investigación está estructurada en 3 capítulos. En el capítulo 1 se realiza un estudio del estado del arte que permite identificar las principales tendencias en las soluciones existentes en la literatura así como las principales deficiencias y fortalezas. El capítulo 2 describe en detalle el método propuesto para solucionar el problema de la ambigüedad en el nombre de los autores. Finalmente en el capítulo 3 se arriban a las conclusiones finales teniendo en cuenta los resultados obtenidos en los experimentos realizados. Además se describen las pruebas realizadas a la propuesta de solución.

Capítulo 1

Fundamentación teórica

1.1. Introducción

La ambigüedad en el nombre de los autores es un problema que afecta a los sistemas que gestionan metadatos bibliográficos. En la actualidad se han desarrollado aproximaciones con el objetivo de solventar la situación mencionada. En el capítulo se exponen los fundamentos básicos de la desambiguación del nombre de los autores en metadatos bibliográficos, destacándose conceptos y definiciones que ayudan a comprender el resto de la investigación. Se presentan los resultados obtenidos luego de realizado un estudio de las aproximaciones más representativas existentes en la literatura, demostrando que son insuficientes, lo cual justificó el desarrollo de la investigación.

1.2. Análisis bibliométrico

Para la realización de la investigación se llevó a cabo un estudio documental que abarca principalmente la literatura publicada en los últimos 5 años. Se consultaron numerosas fuentes bibliográficas, entre las que se encuentran bases de datos referenciadas como SCOPUS[1], Springer[2] e IEEE[3]. También se visitaron los sitios web oficiales de algunas bases de datos y revistas que han desarrollado aproximaciones con el objetivo de resolver el problema de la ambigüedad en el nombre de los autores, por ejemplo CiteeSer[4] y MedLine[5]. A continuación se muestra una tabla resumen de la bibliografía consultada.

[1] *http://www.scopus.com/home.url*
[2] *http://www.springer.com*
[3] *http://ieeexplore.ieee.org/Xplore/home.jsp*
[4] *http://citeseerx.ist.psu.edu*
[5] *http://www.medline.com*

Tipo de bibliografía	**Total**	**Últimos 5 años**
Artículos en revistas científicas	15	10
Artículos en congresos	14	12
Libros	5	1
Secciones de libros	1	1
Tesis de pre-grado	1	1
Reportes de investigación	1	1
Páginas webs	1	-

Tabla 1.1: Resumen de la revisión bibliográfica realizada

La tabla anterior muestra que se consultaron un total de 38 trabajos científicos de los cuales 26 de ellos fueron publicados en los últimos 5 años lo cual representa un 68 % de la bibliografía consultada.

1.3. Marco teórico

En este acápite se describen los principales conceptos asociados al dominio del problema con el objetivo de avalar la investigación, permitiendo de esta forma crear la base de conocimientos necesaria para su desarrollo.

La ambigüedad en el nombre de los autores ha sido definida de varias formas, por ejemplo: en [3] se define como la asignación de autores diferentes a un nombre de autor en metadatos bibliográficos. En [4] se define el problema como la imposibilidad de referirse a un autor correctamente dentro de un conjunto de datos bibliográficos que contiene nombres iguales al que se hace referencia. También [5] define el problema en cuestión como la asignación de diferentes publicaciones científicas a un autor común pero que pertenecen

a autores diferentes o la existencia de publicaciones científicas asignadas a diferentes autores con nombres distintos pero que se refieren a un solo autor.

A continuación se presenta la definición de ambigüedad en el nombre de los autores utilizada en la investigación.

Definición. *Posibilidad de representación de los nombres de los autores de formas diferentes. Dicha situación comprende dos casos fundamentales:*

- *Aparición de nombres sintácticamente diferentes pero que se refieren a la una persona (autor).*
- *Aparición de nombres sintácticamente iguales pero que se refieren a personas (autores) diferentes.*

Otro de los conceptos asociados al problema en cuestión es el de **Metadatos**. Estos se refieren a datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas[6].

De acuerdo a las herramientas y técnicas utilizadas en la resolución del problema de la ambigüedad en el nombre de los autores, la **Minería de datos** es la principal herramienta utilizada, según [7] se define como: el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes volúmenes de datos almacenados en distintos formatos.

Por su parte la minería de datos es un área del conocimiento amplia, dentro de ella se distinguen temáticas de importancia para la investigación, por ejemplo, el **Aprendizaje supervisado** el cual se define como: técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten en pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados[7]. Dentro del aprendizaje supervisado se distinguen los algoritmos de clasificación, estos se definen como: algoritmos utilizados para asignar un elemento entrante no etiquetado

en una categoría concreta conocida. Estos algoritmos, permiten ordenar o disponer por clases elementos entrantes, a partir de cierta información característica de éstos[7].

El aprendizaje no supervisado pertenece también a la minería de datos, este se define como: un método donde un modelo es ajustado a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori[7]. Dentro del aprendizaje no supervisado se distinguen los algoritmos de agrupamiento, estos se refieren a: procedimientos de agrupación de una serie de vectores de acuerdo con un criterio de cercanía. La cercanía se define en términos de una determinada función de distancia, como la euclidiana, aunque existen otras más robustas o que permiten extenderla a variables discretas[7].

De acuerdo a la evolución histórica-lógica han surgido nuevos métodos y herramientas que mejoran los conocimientos anteriores y a su vez sus resultados. Este es el caso de la **Combinación de agrupamientos**, la que se define como: utilización de los resultados de varios algoritmos de agrupamiento para la obtención de un resultado superior a la utilización de un algoritmo de agrupamiento simple[8].

La **distancia de edición** es un elemento referente al procesamiento del lenguaje natural[9]. Existen diferentes definiciones de distancia de edición, en la investigación se utiliza la definida por **Damerau-Levenshtein**. Esta se refiere a: la cantidad de operaciones básicas que es necesario realizarle a una cadena de caracteres para convertirla en otra, siendo estas operaciones la adición, modificación, eliminación y permutación de un carácter[10].

Los **modelos probabilísticos** pueden ser utilizados para la resolución del problema de la ambigüedad en el nombre de los autores debido a que son herramientas que predicen cuan parecidos son dos instancias de un problema determinado. Por tanto se definen como: la forma que pueden tomar un conjunto de datos obtenidos de muestreos de datos con comportamiento que se supone aleatorio. Pueden ser modelos probabilísticos discretos o continuos[11].

1.4. Estado del arte

Las aproximaciones estudiadas pueden dividirse en dos grandes grupos, de acuerdo a (1) la técnica de minería de datos utilizada y (2) la fuente de datos empleada. La clasificación de las soluciones de acuerdo a la técnica de minería de datos se puede dividir en 4 grupos de soluciones, (1) soluciones que usan técnicas de agrupamiento, (2) soluciones que usan técnicas de clasificación, (3) soluciones que utilizan modelos probabilísticos y (4) soluciones que usan una combinación de los métodos anteriores.

Por su parte la clasificación de las soluciones de acuerdo a fuente de datos utilizados se puede dividir en dos grupos, (1) soluciones que usan los metadatos bibliográficos de los repositorios digitales, (2) soluciones que usan la web como fuente de información. A continuación se muestra una taxonomía con lo antes expuesto.

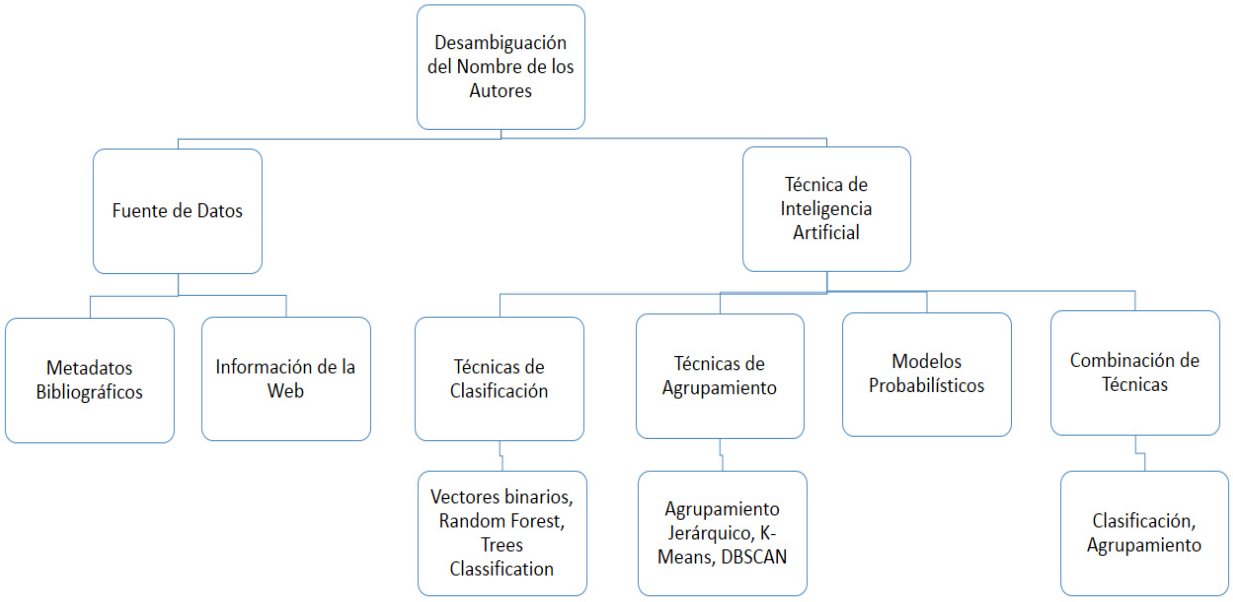


Figura 1.1: Taxonomía de la clasificación de las soluciones consultadas

A continuación se describen cada una de las clasificaciones mencionadas anteriormente además de un análisis crítico de las soluciones que pertenecen a cada uno de los grupos.

1.4.1. Principales aproximaciones

1.4.1.1. Soluciones usando técnicas de clasificación

Las soluciones desarrolladas utilizando técnicas de clasificación, tratan de establecer una correspondencia entre las entradas y las salidas deseadas del sistema.

En [12] se toma como caso de estudio la red social de investigadores ArnetMiner[6]. Se manejan todos los elementos existentes en la red social: co-autores, afiliación, citas de los artículos, similitud entre los títulos de las publicaciones, las páginas web de los autores y la retro-alimentación de los usuarios en la red social. Durante el análisis de los datos utilizados se construye un vector de valores binarios donde se coloca 1 si los nombres que se están comparando cumplen con una determinada característica. El valor asignado al vector depende del elemento que se esté comparando. Por ejemplo, partiendo de que los autores que se comparan comparten el primer nombre (exactamente igual), cuando se comparan los segundos nombres de los autores, se coloca 1 si estos son iguales, 0 en cualquier otro caso.

En [13] se presenta una aproximación orientada a la explotación de toda la información referente a los autores en los datos utilizados: co-autor, afiliación, lugar de publicación, entre otros elementos. También permite calcular el número exacto de autores presentes en un conjunto de datos determinado. En la aproximación se define el problema como se muestra a continuación: Dada una lista de publicaciones $P = p_1, p_2, p_3, \ldots, p_n$ suponga que existen m diferentes personas $t_1, t_2, t_3, \ldots, t_n$ compartiendo un nombre, entonces la tarea consiste en asignar a cada persona la publicación que en realidad escribió.

En [14] se proponen cuatro pasos fundamentales para resolver el problema de la ambigüedad en el nombre de los autores. Primero, se realiza el filtrado de los datos por los nombres y la afiliación, luego se construye un vector de similitudes, después son agrupados los autores y finalmente se realiza su

[6] *http://arnetminer.org/*

clasificación. En la aproximación también son calculadas las tasas de error permitidas por el algoritmo utilizado. Con los experimentos realizados sobre el método propuesto por los autores se pudo demostrar la efectividad de la solución desarrollada.

En [5] se reconoce la necesidad de contar con conjunto de datos amplio con el objetivo de obtener resultados satisfactorios en la solución del problema de ambigüedad en el nombre de los autores. La aproximación es una actualización de un trabajo previo encaminado a generar y clasificar datos de entrenamiento. La propuesta actual genera un gran volumen de datos que pueden ser utilizados como conjunto de datos de entrenamiento. El proceso consiste en la selección de un pequeño grupo de datos, este conjunto es enviado a especialistas para que estos los clasifiquen. En los experimentos realizados se muestra que con un pequeño conjunto de datos (cerca del 5 % de los datos) el rendimiento del proceso de desambiguación mejora en aproximadamente en el 10 %.

1.4.1.2. Soluciones usando técnicas de agrupamiento

Las soluciones basadas en el uso de técnicas de agrupamiento son definidas por una función de similitud para establecer los criterios de agrupamientos entre los nombres de los autores.

En [15] se usa un mecanismo para identificar páginas web con información referente a los autores. Para su identificación se utiliza un modelo de clasificación mediante redes neuronales. Después de que son identificadas, no es posible la extracción de la información de forma directa ya que estas no la contienen de forma estructurada. Por tanto, en la propuesta se creó un mecanismo para la extracción de información referente a la afiliación, los co-autores y los títulos de trabajos. Por último, se procede a realizar el proceso de agrupamiento. Según los autores el proceso de selección y extracción realizado sobre las páginas web mejoró el proceso de agrupamiento.

En [16] se establecen dos tipos de relaciones entre las publicaciones: (1) correlación de tema y (2) correlación web, con el objetivo de explorar las relaciones entre las publicaciones que compartan el nombre de autor. La correlación de tema se refiere a la relación que puede existir entre las temáticas de las publicaciones. La correlación web se refiere a la relación que puede existir entre las publicaciones en las páginas web. Luego de determinadas cada una de estas correlaciones se procede a realizar el proceso de agrupamiento teniendo en cuenta estos dos elementos.

En [17] se desarrolla un método para la selección de un modelo que estima el número correcto de autores en un conjunto de metadatos bibliográficos. Este modelo está basado en la relación existente entre los co-autores. Se muestra además que dada las características del problema, el método desarrollado para la selección del modelo ofrece los resultados exactos. Con la aproximación desarrollada se resuelve el problema de determinar cuántos autores o agrupamientos existen en el conjunto de datos utilizados para la desambiguación.

En [18] se reconoce que los datos almacenados en las bibliotecas digitales institucionales poseen mayor calidad en la información y mayor grado de organización lo que facilita el proceso de desambiguación. Los repositorios de internet no poseen estas características por lo que es necesario enriquecerlos. En la aproximación, el método utilizado para este proceso fueron los modelos de tópicos, para los cuales es necesario contar con una fuente de información con determinadas características, usando para esto Wikipedia[7] . Luego se usa agrupamiento aglomerativo para realizar el proceso de desambiguación del nombre de los autores.

En [4] se analizan diversos factores que intervienen en el bajo rendimiento de las propuestas de solución existentes en la literatura: el enorme espacio de búsqueda de la solución y la diferencia entre la cantidad de citas de los autores (algunos aparecen solo unas pocas veces, mientras que otros son muy productivos

[7] *http://www.wikipedia.org/*

científicamente). En la aproximación se proponen tres resultados principales:(1) un método que se encarga de explorar reglas de asociación para realizar el proceso de desambiguación, (2) un método que se encarga de extraer reglas de asociación por demanda, lo que reduce de forma significativa el espacio de búsqueda de la solución propuesta y (3) una extensión del segundo método con la capacidad de auto-entrenarse, reduciendo la cantidad de datos de entrenamientos necesitados por la solución propuesta.

La propuesta [3] está basada en la creación de un conjunto de datos de entrenamiento aplicable a las soluciones existentes en la bibliografía. Los pasos seguidos para su creación son: (1) determinación de la fuente de datos a utilizar, (2) determinación del conjunto de nombres de autores en los datos seleccionados, (3) generación de las citas de los autores en el conjunto de datos, (4) recolección de información referente a los autores, (5) asignación de identificadores a los nombres de los autores y (6) verificación y repetición del paso anterior. Luego de generado los datos de entrenamiento se probó el método propuesto en una solución basada en técnicas de agrupamiento, arrojando resultados satisfactorios.

La solución descrita en [19] se basa en los elementos relacionados con el contexto de los autores (co-autores, logares de publicación, afiliaciones) para realizar la desambiguación. La propuesta está basada en la utilización del algoritmo K-Means para generar agrupamientos de los posibles autores ambiguos. En la aproximación se realiza un proceso de validación cruzada donde los agrupamientos formados usando como criterio de comparación, por ejemplo, los co-autores son contrastados con los agrupamientos formados usando otro criterio de comparación. Los resultados obtenidos por los autores demuestran la validez de la propuesta, la cual sobrepasa los resultados obtenidos por otras aproximaciones existentes en la literatura.

1.4.1.3. Soluciones usando modelos probabilísticos

Las aproximaciones basadas en modelos probabilísticos establecen relaciones entre las características de los metadatos para determinar la probabilidad de que dos artículos sean escritos por un autor.

En [20] se presenta un modelo probabilísticos para la generación automática del conjunto de datos de entrenamiento. Además, permite estimar la probabilidad de que un par de artículos de la base de datos MedLine[8], que comparten el nombre de autor sean escritos por la misma persona. Dicha probabilidad está basada en que estos comparten el título, la revista de publicación, co-autores en común, entre otros. Esta aproximación marcó un punto de partida para la creación de nuevos métodos de solución relacionados con el tema.

En [21] se presenta un marco de trabajo que permite la incorporación de atributos y sus relaciones dentro de un modelo probabilístico. Se experimenta en una aproximación dinámica para la estimación del número de nombres de autores únicos en el conjunto de datos utilizado, además se desarrolló una medición de distancia adaptativa para estimar la distancia entre los objetos del modelo.

En [22] se utiliza un modelo probabilístico para resolver el problema de la ambigüedad en el nombre de los autores. Teniendo como principal característica para realizar el proceso de desambiguación el título de las publicaciones realizadas por cada autor. Los títulos son analizados, luego colocados en una agrupación aquellas publicaciones que posean mayor probabilidad de referirse a una sola y cuyo dominio tenga una fuerte relación con las publicaciones de dicha agrupación. Este proceso continúa hasta que no es posible agrupar más los nombres de los autores, terminado así el proceso de desambiguación.

En [23] se hace un análisis sobre el bajo rendimiento de los métodos usados en la mayoría de las soluciones que abordan el problema de la ambigüedad en el nombre de los autores. Debido a que, en la mayoría de los casos las agrupaciones que se conforman generalmente son pequeñas. A partir de este análisis los autores proponen un nuevo método para la solución del problema. Basado en la selección de un nuevo conjunto de atributos a partir del cual se lleva a cabo el proceso de agrupamiento. Luego de la selección de dicho conjunto, este es utilizado para determinar un ratio de probabilidad, para el cual, mayores valores

[8] *http://www.medline.com*

significa que hay mayor probabilidad de que dos conjuntos de autores se refieran a una persona. Además se propone un método para determinar la cantidad exacta de autores, dado un nombre, a partir de las estadísticas extraídas de un repositorio digital. Con los resultados de los experimentos los autores demuestran que el rendimiento del método propuesto es mejor que los métodos tradicionales.

1.4.1.4. Soluciones usando una combinación de métodos

Las aproximaciones estudiadas en la investigación comprenden la utilización de combinaciones de métodos para su desarrollo.

En [24] se afirma que los métodos de aprendizaje supervisado arrojan mejores resultados para este tipo de solución pero es necesaria la intervención de los humanos en el proceso de generación de los datos de entrenamiento. Los autores proponen un método para la solución de la desambiguación basado en dos pasos. El primero de ellos es utilizado para la generación de los datos de entrenamiento a través de un algoritmo de agrupamiento, basado en la similitud entre el nombre de los co-autores. El segundo paso utiliza un algoritmo de aprendizaje supervisado para realizar el proceso de desambiguación. El objetivo es detectar los autores no incluidos en ninguno de los datos de entrenamiento generados en el paso anterior.

En [25] se reconoce la importancia que tiene poseer un equilibrio entre precisión y rapidez en los métodos de desambiguación. A partir de esto, se propone un algoritmo para resolver el problema de la ambigüedad usando todos los campos disponibles en los metadatos bibliográficos. Además el proceso de comparación entre dos autores es dinámico. Los elementos que se toman en cuenta para comparar un par de autores no son necesariamente iguales que los que se usan para comparar otros, varían en dependencia de la disponibilidad de información. Se tiene en cuenta la diferencia en las temáticas de publicación de los autores y las fechas de publicación de los trabajos. La propuesta tiene la particularidad de no preseleccionar elementos previamente para realizar el proceso de desambiguación. Esto ocasiona que el conjunto de datos a

comparar sea mayor aumentando la exhaustividad de la propuesta. Los resultados expuestos en el informe muestran que la solución mejora en rapidez y precisión con respecto a las propuestas existentes en la bibliografía.

En [26] se propone una herramienta para la evaluación de los métodos propuestos en la literatura sobre la desambiguación del nombre de los autores. Además las aproximaciones no tienen en cuenta la adición de nuevos registros en las revistas digitales ni los cambios que puedan aparecer en los intereses de los investigadores. Después de realizar las pruebas pertinentes sobre tres soluciones desarrolladas se demuestra la efectividad de la herramienta desarrollada.

En [27] se utiliza un modelo basado en grafos para resolver el problema de la desambiguación del nombre de los autores. Luego de utilizar un método basado en la partición de grafos se realiza el proceso de desambiguación, teniendo como base un conjunto de datos de entrenamiento. Los datos de entrenamiento utilizados son determinados por una solución propuesta en la aproximación, con el objetivo de que estos sean la menor cantidad posible.

1.4.1.5. Clasificaciones de las soluciones de acuerdo a la naturaleza de los datos

Las soluciones estudiadas se pueden clasificar en dos grandes grupos de acuerdo a la naturaleza de los datos que utilizan: **soluciones que usan los metadatos de los repositorios digitales y soluciones que usan la web como fuente de información.**

Las primeras utilizan diversos mecanismos (protocolos, librerías, etc.) para obtener los registros bibliográficos (metadatos) de los repositorios digitales. En este caso los metadatos cuentan con un mayor nivel de detalle y organización lo que permite que el trabajo con estos sea menos complicado, ocurriendo todo lo contrario con la información obtenida de la web.

Las segundas utilizan la información que se encuentra pública en la web referente a los autores en los repositorios digitales. En este caso los datos son obtenidos a través de consultas a motores de búsqueda. Las formas de componer las consultas pueden ser variadas, por ejemplo: *nombre del autor + título de la publicación*. Otro caso puede ser: *nombre del autor + título de la publicación + afiliación del autor (en caso de estar disponible)*. Con el objetivo de facilitar el trabajo con los datos obtenidos estos deben ser procesados previamente.

1.4.2. Sistemas de metadatos para la identificación y desambiguación del nombre de los autores

Tener un registro único de cada uno de los autores en la web sería un gran paso de avance para solventar el problema de la ambigüedad en el nombre de los autores. La idea mencionada consiste en contar con un mecanismo de identificación que permita a los autores registrar sus datos solo en una ocasión[28], luego para identificarse en una revista solo usa el mecanismo de identificación proporcionado por la base de datos donde registró sus datos. Este proceso minimizaría en gran medida la aparición de errores de escritura en los nombres de los autores, además que permitiría a los sistemas que usan este tipo de información acceder a ella de una forma mucho más eficiente y rápida.

Muchas iniciativas han surgido teniendo como base esta idea, entre ellas se pueden mencionar:

Library of Congress Authorities[9]: Esta biblioteca combina nombre, temática y títulos de los autores registrados en ella, está formada por registros generados por bibliotecas de los Estados Unidos aunque existen contribuciones de otras instituciones de este tipo, como por ejemplo la biblioteca británica. Los registros de los autores están almacenados en el formato de autoridad MARC.

[9] *http://authorities.loc.gov*

Virtual International Authority File (VIAF)[10]: Proyecto conjunto de varias bibliotecas internacionales que tiene como objetivo disminuir los costos y aumentar la utilidad de los archivos de autoridad. Comparando y relacionando estos haciéndolos hacen accesibles desde la web. Teniendo en cuenta que es un proyecto conjunto internacional es necesario contar con varias formas de introducir los datos de un autor. Almacena los archivos de autoridad en formato MARC y UNIMARC.

ResearcherID[11]: De acuerdo a la información existente, esta iniciativa es una comunidad multidisciplinaria que provee un identificador único a cada uno de los autores que participen en el proyecto. Fue creada y es soportada por Thomson Reuters. Cada uno de los autores debe crear una página en la cual se registran sus datos, por ejemplo los artículos científicos, los libros, citas que hayan recibido los trabajos, entre otras. Cada una de las páginas creadas por los autores es de libre acceso.

International Standard Name Identifier (ISNI)[12]: El propósito de esta iniciativa es asignar un número único a los autores que aparezcan en publicaciones tanto en línea como impresas. Este número es similar al ISBN que aparece en los libros pero se diferencia en que, por ejemplo, un libro con dos ediciones, cada una de las ediciones tienen ISBN diferentes, mientras que con la iniciativa el número asignado será siempre igual.

Digital Author Identification System (DAI)[13]: Este es un ejemplo de un sistema de identificación de los nombres de los autores internacional. Consiste en la asignación de un número a cada uno de los profesores e investigadores que se encuentran registrados en el sistema. El número asignado por el sistema sigue el patrón y es compatible con ISNI.

[10] *http://viaf.org/*
[11] *http://www.researcherid.com/*
[12] *http://www.isni.org/*
[13] *http://www.rug.nl/bibliotheek/informatie/digitaleBibliotheek/daikort?lang=en*

Open Researcher and Contributor ID (ORCID)[14]: Iniciativa creada con el objetivo de solucionar el problema de la ambigüedad en el nombre de los autores. Esta iniciativa crea un registro único de cada uno de los autores y un mecanismo de enlazado con otras iniciativas de este tipo. ORCID permite mejorar el rendimiento del proceso de descubrimiento de información relacionada con los autores. El proceso comienza con el registro de los datos, luego le es asignado un identificador único el cual es usado como mecanismo de identificación cuando dicho autor firme un artículo o contribución.

Anteriormente se mencionaron algunas de las principales soluciones desarrolladas con el objetivo de resolver el problema de la ambigüedad en el nombre de los autores. Aunque no son pocas las soluciones y los métodos utilizados para mitigar dicho problema, estas no están expensas de problemas que dificultan la obtención de los mejores resultados. A continuación se realiza un análisis crítico de estas.

En las soluciones revisadas no se tienen en cuenta la calidad de los datos utilizados. Es necesario tener en cuenta las características, inconsistencias y ruidos que pueden aparecer en estos para determinar los métodos que mejor se adapten a sus particularidades. De igual forma, no se realiza un previo procesamiento de los datos. En muchas ocasiones las características de los datos permiten la realización de pre-procesamiento de forma tal que los resultados de las soluciones mejoren considerablemente. También se asume que si aparecen dos nombres iguales entonces se refieren a una persona. Teniendo en cuenta el estudio previo, podemos afirmar que esta es una suposición incorrecta. Se toman como punto de partida para realizar el proceso de desambiguación que los nombres de los autores deben coincidir completamente para comenzar el proceso antes mencionado, esta condición obtendría resultados erróneos si existiesen errores de escritura en los nombres que se están analizando. También se puede afirmar que la mayoría de las soluciones están orientadas al idioma inglés, muy pocas se centran en otros idiomas y sus particularidades, como por ejemplo el español.

[14] *http://www.orcid.org*

1.4.3. Ventajas y desventajas de las soluciones estudiadas

Las soluciones basadas en técnicas de agrupamiento permiten la obtención de resultados sin necesidad de tener un conocimiento previo de la información que será tratada. No es necesario poseer un conjunto de datos de entrenamiento, como es el caso de los algoritmos de clasificación. También posibilitan que el proceso de desambiguación sea automatizado, eliminando la intervención de la actividad humana. Por otro lado este tipo de soluciones poseen limitantes que dificultan su utilización. Los resultados obtenidos tienen una menor calidad que los obtenidos por otros tipos de soluciones, por ejemplo, las soluciones basadas en clasificadores. También se puede plantear que, cuando se utilizan técnicas de agrupamiento no se conoce el número de agrupaciones que se deben crear en el proceso de desambiguación, introduciendo errores en los resultados obtenidos.

Las soluciones basadas en clasificadores permiten que los resultados obtenidos tengan un grado de fidelidad alto en comparación con otras técnicas, como por ejemplo, el agrupamiento. En muchas ocasiones este tipo de solución es la más eficaz para su utilización debido a la forma de modelar el problema de la ambigüedad. Entre sus limitaciones se encuentra que, es necesario conocer información previa de los datos tratados, tener un conjunto de datos de entrenamiento para utilizarlos en la construcción del modelo creado por el clasificador. Esto hace que sea necesaria la intervención de la actividad humana para determinar las principales características de los datos utilizados, además de ser una situación poco común en los escenarios reales.

Las soluciones basadas en la utilización de modelos probabilísticos son unas de las menos complejas para su utilización. La determinación de los elementos que compondrán el modelo probabilístico y sus respectivas ponderaciones es un trabajo relativamente sencillo. Las ponderaciones pueden ser determinadas utilizando métodos heurísticos. Por otro lado, la utilización de este tipo de soluciones deben ser aplicadas en entornos muy controlados, es decir, donde las características de los datos utilizados sean conocidas. Esto hace que este tipo de solución solo se pueda aplicar, con resultados satisfactorios, cuando se conocen las

características de los datos tratados.

Las soluciones basadas en la utilización de una combinación de los métodos estudiados tratan de resolver los problemas de los restantes enfoques mencionados en la investigación. Entre las dificultades que se tratan de resolver se encuentran: la generación de datos de entrenamiento con algoritmos de agrupamiento, para la posterior utilización de estos en los algoritmos de clasificación. Es decir, están centradas en solucionar el problema de la ambigüedad del nombre de los autores teniendo como base las dificultades encontradas en otros enfoques de solución. Esto hace que los resultados obtenidos con este tipo de solución muchas veces tengan mayor calidad que los resultados obtenidos con otros tipos de soluciones. También, debido a la utilización de diferentes métodos de solución, aumenta la aparición de errores y dificultades propias de cada uno de los tipos de soluciones estudiadas.

1.5. Combinación de agrupamientos

La selección de un algoritmo de agrupamiento para la solución de un problema determinado, de forma tal que los resultados obtenidos sean los mejores, es una tarea compleja. De lo anterior surge la idea de combinar los resultados provenientes de varios algoritmos de agrupamiento (*cluster ensemble* por su traducción al inglés) para obtener una solución final superior a la utilización de un algoritmo de agrupamiento simple. El origen de esta idea proviene del éxito obtenido en la combinación de algoritmos de clasificación.

El objetivo de la combinación de agrupamientos es obtener resultados con mayor calidad que los obtenidos por algoritmos individuales, ya que esta combinación puede equilibrar los errores cometidos por estos. Se puede afirmar que el resultado obtenido por la combinación de varios algoritmos de agrupamientos debe ser más robusto y ajustado al problema que se pretende resolver[29].

Una buena estrategia de combinación, debe permitir encontrar nuevas estructuraciones más consistentes que las existentes, entendiendo por consistente a una estructuración que comparte gran cantidad de información o es muy similar al conjunto de particiones. Esta estructuración de consenso debe ser, lo más invariante posible a pequeñas variaciones en los datos, debe ser suficientemente robusta ante información ruidosa[29].

Uno de los procesos asociados a la combinación de agrupamientos es la **generación** de los resultados de los algoritmos de agrupamiento. En el proceso de generación pueden utilizarse diferentes modelos de algoritmos, diferentes algoritmos, distintas representaciones de los objetos o utilizando diferentes subconjuntos de objetos.

El proceso fundamental dentro de la combinación de agrupamientos es el **consenso** de los resultados de los algoritmos, debido a que se obtiene el resultado final de la combinación. Existen dos tendencias principales en el proceso de consenso: (1) basada en la idea de votación entre objetos, donde es analizada la cantidad de veces que un elemento pertenece a un agrupamiento, o la cantidad de veces que pares de elementos se encuentran juntos en un agrupamiento y (2) basada en la idea de la búsqueda de la partición media.

Las funciones de consenso basadas en la co-asociación parten de la idea de que diferentes agrupamientos de los datos pueden tener cada uno sus fortalezas y debilidades. Se espera que la contribución conjunta de estos tenga un efecto compensatorio en los resultados y permitiendo su mejora con respecto a agrupaciones obtenidas al utilizar un solo algoritmo de agrupamiento.

Considerando las particiones obtenidas a partir de diferentes algoritmos, éstas son mapeadas en un espacio intermedio: la matriz de co-asociación, donde la posición (i, j) representa el número de veces que el par de objetos i y j co-ocurren en un agrupamiento. Cada co-ocurrencia es vista como un voto para que ambos objetos se encuentren en el mismo conjunto en la partición de consenso. Dividiendo los valores de la

matriz por el número de particiones en la combinación de agrupamientos se obtienen los votos normalizados.

Teniendo en cuenta las limitaciones detectadas en las aproximaciones existentes en la literatura y sus características positivas se pretende desarrollar un algoritmo para la desambiguación del nombre de los autores. Estará caracterizado por la utilización de la combinación de agrupamientos como herramienta clave de minería de datos.

1.6. Conclusiones parciales

Luego de analizadas las aproximaciones estudiadas se puede afirmar que no existe una solución exacta para la desambiguación del nombre de los autores. Las aproximaciones son abordadas utilizando cuatro enfoques, (1) técnicas de agrupamiento, (2) técnicas de clasificación, (3) modelos probabilísticos y (4) usando una combinación de los enfoques vistos anteriormente. Las soluciones estudiadas también pueden clasificarse de acuerdo a la naturaleza de los datos, (1) soluciones que utilizan los metadatos de los repositorios digitales y (2) soluciones que usan la web como fuente de información. Se puede afirmar que las aproximaciones existentes no realizan un previo procesamiento de la información. No se tiene en cuenta la calidad de los datos utilizados. Además, ninguna de las aproximaciones estudiadas está orientada a las características del idioma español. También se puede afirmar que la combinación de agrupamientos constituye una herramienta idónea para solucionar el problema de la ambigüedad en el nombre de los autores.

Capítulo 2

Propuesta de solución

2.1. Introducción

La ambigüedad en el nombre de los autores es un problema ocasionado por la falta de unicidad en su registro en los metadatos bibliográficos. En el capítulo se describe la propuesta de solución desarrollada para la resolución del problema en cuestión. Se detallan los algoritmos desarrollados, además de explicar los procedimientos seguidos para obtener la solución propuesta.

2.2. Descripción general de la propuesta

De acuerdo a las propuestas existentes en la literatura con respecto a la resolución del problema en cuestión, la mayoría[22][24][25] explotan las relaciones que se pueden establecer entre los contextos de los autores. El contexto de los autores se refiere a elementos existentes en los metadatos bibliográficos que permiten crear un perfil de los mismos. Entre los elementos disponibles en los metadatos bibliográficos se encuentran los co-autores, los títulos de las publicaciones, los lugares de publicación y las afiliaciones. Los elementos mencionados permiten establecer relaciones de similitud entre los autores y así realizar el proceso de desambiguación.

Como se define en el título de la investigación y en su diseño metodológico, el principal objetivo del trabajo es desarrollar un algoritmo para desambiguar el nombre de los autores en metadatos bibliográficos. De acuerdo a [30] se define como **Algoritmo** a un conjunto de instrucciones finitas y ordenadas que permiten la resolución de un problema, transformando los datos de entrada en los datos de salida. La definición anterior muestra un componente importante dentro del desarrollo de algoritmos: los datos de entrada. Es necesario que posean una estructura bien definida con el objetivo de garantizar la portabilidad y estandarización de los algoritmos desarrollados. La estructura de estos puede estar definida de diversas

formas y en diversos formatos.

Para el desarrollo de la investigación se definió una estructura de datos en formato XML[1]. Utilizando las facilidades que el formato XML ofrece se estructuró de la forma que se muestra a continuación:

```
<Root>
    <author>
        <id>1</id>
        <name>grisel infante costa</name>
        <affiliation>falta</affiliation>
        <titles>
            <title>Akademos, un sistema automatizado para la gestión académica.</title>
        </titles>
        <sources>
            <source>Serie Científica; Vol 1, No 1 (2007)</source>
        </sources>
        <co_authors>
            <co_author>yordanis camejo valdivia</co_author>
            <co_author>darien cepero rojas</co_author>
            <co_author>dianly santiler alvarez</co_author>
        </co_authors>
    </author>
<Root>
```

Figura 2.1: Estructura de datos de entrada

Se hace necesario definir un formato de salida estandarizado y bien estructurado que permita su re-utilización por terceras aplicaciones y sistemas. Dicho formato se definió en XML como se muestra a continuación:

[1] *http://www.w3.org/XML/*

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<Root>
    <author>
        <id>1</id>
        <name id="27">pedro y. yobanis pinero perez</name>
        <sameAs>
            <name id="121">pedro yobanis pinero perez</name>
        </sameAs>
    </author>
</Root>
```

Figura 2.2: Estructura de datos de salida

La propuesta de solución comprende tres procesos fundamentales, (1) identificación de las relaciones existentes entre los autores usando la distancia de edición definida por Damerau-Levenshtein, (2) generación de un vector de similitudes utilizando las relaciones identificadas. Compuesto por las similitudes entre los elementos del contexto de los autores (co-autores, afiliación, lugares de publicación y títulos de las publicaciones) y (3) utilizar la combinación de agrupamientos para determinar si las relaciones establecidas en el proceso anterior son correctas.

Con el objetivo de lograr una mayor comprensión de la propuesta de solución se muestra el proceso general de desambiguación:

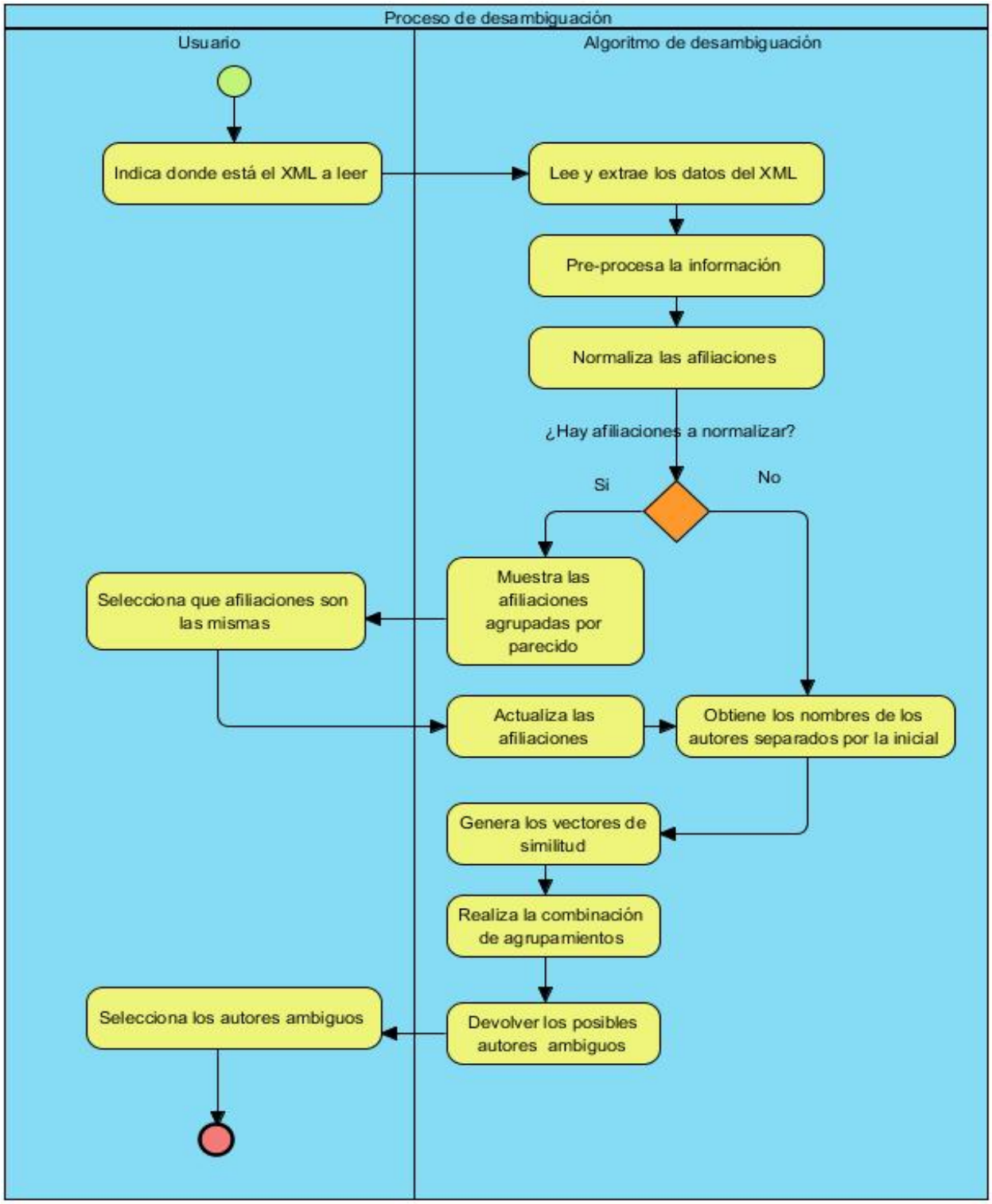


Figura 2.3: Proceso general de desambiguación

2.2.1. Representación de los vectores de similitud

La calidad de los metadatos utilizados en las soluciones es un elemento importante en el problema tratado. Realizar tareas de pre-procesamiento de la información (metadatos) elevaría la calidad de los resultados finales debido a que se eliminarían datos ruidosos. Las tareas llevadas a cabo en este sentido fueron: (1) conversión de todos los elementos que representaban cadenas de caracteres a minúscula, (2) eliminación de las tildes, (3) eliminación de los caracteres extraños y (4) eliminación de las partículas de los nombres.

2.2.1.1. Homogenización de las afiliaciones

Otro proceso realizado en el pre-procesamiento de la información fue la homogenización de las afiliaciones de los autores debido a que es uno de los elementos de mayor peso en el proceso de desambiguación[20]. Por ejemplo, las afiliaciones *"Universidad de las Ciencias Informáticas"* y *"Facultad 3, Universidad de las Ciencias Informáticas"* sintácticamente representan dos instituciones diferentes, refiriéndose en realidad a una sola. Este proceso se realizó utilizando la distancia de edición. Tomando como referencia dicho valor se agruparon todas las afiliaciones cuyo valor de distancia de edición fuera superior a un umbral que permitiera considerar que las afiliaciones comparadas representan variantes de una institución. El umbral puede ser calculado de dos formas: (1) a través de un umbral absoluto y (2) a través de un umbral relativo. Debido a que la utilización de un umbral absoluto no es eficaz en la formación de grupos, se propone un umbral relativo.

$$\beta = \alpha * \text{mín}(|A|, |B|) \tag{2.1}$$

Donde A y B representan la cantidad de palabras de las afiliaciones y α representa una constante determinada por los resultados de los experimentos.

Para comparar las afiliaciones se tomaron como una lista de palabras y se eliminaron los puntos. En ocasiones las afiliaciones no tienen las palabras en un orden único y la similitud que resultaría de aplicar la distancia de edición sería mayor que la que resultaría intuitivamente. Seguidamente se determinan qué

palabras son lo suficientemente parecidas. Esto se logra limitando el valor de la distancia de edición a que sea menor o igual a 1, lo cual condiciona que los errores que puedan aparecer sean solamente errores de escritura. Luego se calcula la razón que existe entre las palabras que coinciden en las dos afiliaciones y todas las palabras. Finalmente se compara el valor calculado con el umbral, si es mayor, entonces las afiliaciones comparadas son agrupadas, luego los autores determinan si las afiliaciones agrupadas en un conjunto se refieren a una institución.

2.2.2. Relaciones entre los nombres de los autores

Uno de los elementos de importancia en la desambiguación es la similitud que existe entre los nombres de los autores[20]. En la propuesta de solución se desarrolló una función de similitud entre los nombres de los autores, tiene como principal componente la distancia de edición y su objetivo es establecer relaciones entre los nombres de los autores y cuantificarlas.

2.2.2.1. Función de similitud

Similar al trabajo realizado en [31], la función de similitud se define como un conjunto de comparaciones entre las sub-cadenas que componen los nombres de los autores. Permiten devolver la distancia de edición entre las cadenas de caracteres, tolerando también errores de escritura. Para sub-cadenas de longitud siete se tolera un error de escritura y para sub-cadenas de longitud catorce se toleran dos errores de escritura. A continuación se muestra una tabla con el sistema de puntuación empleado para realizar la comparación entre las sub-cadenas:

Valor devuelto	**Interpretación**
0	La coincidencia es total
1	Hay un cambio de edición entre las dos sub-cadenas
2	Hay dos cambios de edición entre las sub-cadenas y la menor de ellas tiene un longitud menor que catorce
∞	No se cumple ninguna de las situaciones anteriores

Tabla 2.1: Sistema de puntuación para la comparación de sub-cadenas

Otro elemento a tratar en la función de similitud entre los nombres fueron las partículas presentes en ellos. Estas se refieren a palabras de longitud corta que no representan nombres: *del, la, el, etc*.... Luego de detectarlas fueron eliminadas en la comparación de los nombres debido a que no aportan información relevante en la comparación.

También se detectaron y procesaron los apellidos compuestos. En el contexto de la investigación los apellidos compuestos se refieren a errores en la escritura, por ejemplo: **Pedro GarcíaSierra**. Para solucionar este problema se llevó a cabo el siguiente proceso:

Teniendo, por un lado, una palabra **a** que forma parte de las palabras del nombre y, por otro lado, un conjunto de palabras **C** cuya unión puede haber formado la primera. De ser cierto lo anterior, podemos afirmar que $\mathbf{a} \in \mathbf{C}$, por tanto se concluye que probablemente el conjunto de palabras **C** ha sido compuesta por la unión de dos elementos del nombre (por ejemplo, la unión de los dos apellidos). A continuación se muestra una tabla con el sistema de puntuación empleado para cuantificar la identificación de apellidos compuestos:

Valor devuelto	**Interpretación**
0	Tanto el comienzo como el final del apellido del autor se corresponden con palabras del nombre del otro autor
1-4	Es un posible apellido compuesto pero se habrían producido errores en su escritura
∞	No se cumple ninguna de las situaciones anteriores

Tabla 2.2: Sistema de puntuación para la identificación de los apellidos compuestos

La función de similitud entre los nombres de los autores tiene un objetivo doble. Primero, determinar si dos nombres de autores son parecidos o no y segundo cuantificar la similitud. A los resultados de la **comparación de sub-cadenas** y **comprobación de apellidos compuestos** se le llamará **disimilitud**.

Para cada palabra del nombre del autor **A** se puntuará las coincidencias con el nombre del autor **B**. Los siguientes casos se evalúan en orden. Una vez que se cumplen las pre-condiciones de uno se aplican los consecuentes y se obvian los demás.

Interpretación	**Operación a realizar**
El apellido coincide con alguno de los apellidos del otro autor	$similitud = similitud + 30 - disimilitud * 10$
El apellido coincide con alguna de las palabras del nombres del otro autor	$similitud = similitud + 10 - disimilitud * 5$
Comprobar si el apellido puede ser compuesto	$similitud = similitud + 5 - disimilitud * 10$
No era ninguno de los casos anteriores	Se acaba la comparación y el valor devuelto es 0

Tabla 2.3: Sistema de puntuación para la comparación entre los apellidos

Los valores establecidos en cada una de las ecuaciones responden a los niveles de importancia referente a las partes que conforman los nombres de los autores. En el caso de las coincidencias de los apellidos los niveles de similitud deben ser mayores. La coincidencia que aparece entre un apellido y un nombre atribuye una similitud media. Por último, la coincidencia entre los nombres provee menores niveles de similitud. Esto se aplica debido a las particularidades del idioma español. En muchas ocasiones existen autores que comparten las palabras del nombre y no se refieren a la misma persona. Lo contrario ocurre con los apellidos, es poco probable que si dos autores comparten los apellidos no se refieran a una sola persona, a menos que posean lazos familiares.

Luego de encontrar una correspondencia para cada uno de los apellidos del autor es necesario establecer una correspondencia entre los elementos restantes del nombre. Estos elementos pueden ser o palabras del nombre o iniciales. En caso de que sean palabras del nombre la puntuación es similar a la anterior (2.2.2.1), la diferencia principal es que no se suelen juntar los nombres, por lo tanto no se evalúa esta posibilidad. En caso de que sean iniciales se realizan las operaciones mostradas a continuación:

Interpretación	**Operación a realizar**
La inicial corresponde con la inicial de alguno de los nombres a los que todavía no se le ha encontrado una correspondencia	$similitud = similitud + 20$
La inicial corresponde con la inicial de algún apellido del otro autor a los que no se le encontró correspondencia	$similitud = similitud + 5$

Tabla 2.4: Sistema de puntuación para la comparación entre las iniciales

Luego de determinar el valor de la función de similitud este se compara con un umbral, determinado de forma similar al utilizado en la homogenización de las afiliaciones. Si este valor es mayor que el umbral utilizando entonces los nombres posiblemente sean representaciones diferentes de un autor.

2.2.3. Relaciones de similitud

Utilizando los elementos disponibles en los repositorios digitales (co-autores, afiliaciones, lugares de publicación, títulos de las publicaciones) se establecen relaciones entre los nombres de los autores.

La utilización de los co-autores constituye un elemento importante en el proceso de desambiguación[20]. Para el establecimiento de la similitud entre los co-autores se siguió el proceso descrito a continuación.

Algoritmo 1 Similitud entre los co-autores

Entrada: Lista de co-autores del autor A con longitud $L1$, Lista de co-autores del autor B con longitud $L2$.

Salida: Similitud entre los co-autores.

Inicializar: Hacer similitud mayor sim_may y suma similitud sum_sim variable con valor 0
para $i = 1$ hasta $L1$ **hacer**
    para $j = 1$ hasta $L2$ **hacer**
        Calcular $similitud$ entre el co-autor i-ésimo y el co-autor j-ésimo
        si $similitud > sim\_may$ **entonces**
            $sim\_may = similitud$
        fin si
    fin para
    $sum\_sim+ = sim\_may$
fin para
devolver $sum\_sim/L1$

Un elemento importante en la validación de la solución propuesta es el análisis de la eficiencia de los algoritmos presentados. Esta se define como el uso óptimo de los recursos que el algoritmo utilice[30], en este caso, el tiempo de ejecución, teniendo en cuenta que dicho algoritmo llega a los objetivos propuestos. Existen diferentes formas de evaluar la eficiencia de los algoritmos, siendo una de ellas el cálculo de la

complejidad temporal haciendo uso de la notación asintótica.

La complejidad temporal de un algoritmo es la función $O(n)$ que mide el número de instrucciones realizadas por el algoritmo para procesar los n elementos de entrada. Cada instrucción tiene asociado un costo temporal[30]. Podría considerarse que los valores de los n casos que se presentan como entrada son los correspondientes: a un caso típico, a un caso promedio, o al peor caso.

Para el algoritmo **2.2.3**, la mayor complejidad temporal radica en la aparición de dos ciclos anidados (líneas 2 y 3), por tanto la complejidad temporal del primer ciclo es de $O(n^2)$, el resto de las sentencias del algoritmo son de complejidad constante $O(1)$. Lo que produce una complejidad temporal de $O(n^2)$. Para una explicación detallada del análisis de la complejidad temporal ver **A.1**.

Otra relación de similitud que se establece es entre las afiliaciones de los autores, siendo estos elementos de importancia en el proceso de desambiguación[20]. Debido a la homogenización realizada con anterioridad la similitud que se establece entre las afiliaciones de los autores, es determinar si estas coinciden o no.

También se estableció una relación entre los lugares de publicación. Para el establecimiento de esta relación se siguió un procedimiento similar al utilizado en la relación entre los co-autores 2.2.3. Primeramente se verifica cuantos lugares de publicación coinciden entre los dos autores comparados y finalmente se divide el número de coincidencias entre la cantidad de lugares de publicación menor de los dos autores.

Finalmente se estableció una relación de similitud entre los títulos de las publicaciones de los autores. Partiendo de que los títulos de las publicaciones son un conjunto de palabras separadas por espacios, se sigue un procedimiento similar al establecido para la homogenización de las afiliaciones. Se comparan cada una de las palabras del título del autor **A** con las del autor **B**, utilizando la distancia de edición. Luego, se seleccionan las palabras que coinciden en los dos títulos y finalmente se encuentra la razón entre la cantidad

de palabras coincidentes en los títulos y la cantidad de palabras del menor de ellos.

De acuerdo a las ventajas detectadas en las soluciones que emplean algoritmos de clasificación se puede afirmar que su utilización en la solución propuesta sería la más efectiva. También se mencionaba que dichos algoritmos necesitan datos de entrenamiento para realizar el proceso de desambiguación. Lo anterior constituye la principal deficiencia que poseen los algoritmos de clasificación para su empleo. La literatura consultada evidenció el uso de conjuntos de datos de entrenamiento en las aproximaciones similares pero estos no están compuestos por nombres en el idioma español.

El párrafo anterior evidenció la imposibilidad de utilizar técnicas de clasificación, lo que propicia la utilización de técnicas de agrupamiento en la solución propuesta. Como se menciona en el capítulo 1 la combinación de agrupamientos es una técnica que reduce las limitaciones de las técnicas de agrupamiento, haciendo de esta una herramienta ideal para solucionar el problema de la ambigüedad en el nombre de los autores.

2.2.4. Combinación de agrupamientos

La combinación de agrupamientos es el principal elemento dentro de la propuesta de solución. En la literatura se han propuesto diversos algoritmos para realizar este proceso[32], [33]. En la propuesta de solución se utiliza un método basado en la co-ocurrencia[32]. Partiendo de una matriz de co-ocurrencia donde se representan la cantidad de veces que un par de objetos han sido colocados en el mismo agrupamiento luego de ejecutar N métodos algoritmos, dicho número se divide entre N para obtener las cantidades normalizadas y luego se compara con el umbral fijo $0{,}5$. Todos aquellos objetos que en la matriz de co-ocurrencia sean mayores que dicho umbral entonces son colocados en un agrupamiento. El procedimiento es descrito a continuación.

Algoritmo 2 Combinación de agrupamientos

Entrada: n objetos, α combinación de agrupamientos conformada por m particiones del conjunto de objetos.

Salida: Partición de consenso.

Inicializar: Hacer una matriz de co-asociación co_assoc nula de dimensiones $n\ x\ n$.
Inicializar co_assoc
para $i = 0$ hasta m **hacer**
    Correr el i-ésimo método de agrupamiento y producir la partición α_i.
    para $i = 0$ hasta m **hacer**
        para $j = 0$ hasta m **hacer**
            si objeto i-ésimo y j-ésimo pertenecen al mismo agrupamiento en α_i **entonces**
                $co\_assoc(i, j) = co\_assoc(i, j) + 1/m$
            fin si
        fin para
    fin para
fin para
para $i = 0$ hasta m **hacer**
    para $j = 0$ hasta m **hacer**
        si $co\_assoc(i, j) > 0{,}5$ **entonces**
            Unir los objetos i-ésimo y j-ésimo en el mismo agrupamiento. Si los objetos pertenecen a dos agrupamientos diferentes se unen en uno solo.
        fin si
    fin para
fin para
Cada objeto no incluido en ningún agrupamiento forma un agrupamiento que solo lo contenga a él.

De acuerdo con [32] la complejidad temporal del algoritmo mostrado **2.2.4** es de $O(n^2)$.

Como se menciona en el capítulo anterior, en la combinación de agrupamientos, la generación de los resultados de los algoritmos empleados puede ser llevada a cabo de varias formas. La empleada en la investigación se refiere a la utilización de diferentes modelos de algoritmos. De acuerdo a las características de la solución, donde los valores empleados son numéricos. También, el proceso de pre procesamiento realizado disminuye considerablemente los datos ruidosos. Además, los grupos de autores ambiguos suelen ser pequeños[23]. Todo lo anterior propició la utilización del algoritmo de agrupamiento K-Means en la fase de generación de los resultados. Aunque dicho algoritmo posee limitaciones que dificultan su empleo, es fácil de comprender e implementar. Entre las limitaciones que posee el algoritmo K-Means se encuentran las siguientes:

- Falla cuando los puntos de un agrupamiento están muy cercanos al centroide de otro grupo.
- No obtiene buenos resultados cuando los agrupamientos tienen diferentes formas y tamaños.
- Son muy susceptibles al problema de la inicialización.
- Son lentos ante conjuntos de datos grandes.

La propuesta de solución descrita en la investigación tiene como principal resultado el diseño de un algoritmo para la desambiguación del nombre de los autores. Entre las herramientas usadas en el desarrollo del algoritmo se encuentran la distancia de edición y la combinación de agrupamientos. A continuación se describe el algoritmo propuesto.

Algoritmo 3 Algoritmo de desambiguación

Entrada: Lista de objetos a desambiguar con longitud N, $umbral$ de comparación.

Salida: Lista de autores desambiguados.

1: **Inicializar:** Hacer una lista de autores parecido aut_par vacía, lista resultado $result\_list$ de autores desambiguados.
2: **para** $i = 0$ hasta N **hacer**
3: $aut\_par.add(objeto\ i - esimo)$
4: **para** $i = i + 1$ hasta N **hacer**
5: Calcular $similitud$ entre el objeto i-ésimo y j-ésimo.
6: **si** $similitud > umbral$ **entonces**
7: $aut\_par.add(objeto\ j - esimo)$
8: Calcular sim_vec entre los elementos restantes de los datos de entrada presentes en aut_par.
9: **fin si**
10: **fin para**
11: Combinación de agrupamientos().
12: Adicionar un objeto de cada agrupamiento formado a la lista de autores desambiguados
13: **fin para**
14: **devolver** $result\_list$

Finalmente podemos determinar la complejidad temporal del algoritmo mostrado. Como se puede apreciar posee dos ciclos anidados $O(n^2)$ (líneas 2 y 4), mientras que en el primer ciclo existe una sentencia de complejidad $O(n^2)$ (línea 12), el resto de las sentencias son de complejidad contante $O(1)$. Por tal motivo la complejidad temporal global del algoritmo es de $O(n^4)$. Para una explicación detallada de la complejidad temporal del algoritmo ver **A.2**.

Los algoritmos presentados en la investigación poseen una cota superior de complejidad temporal menor o igual a $O(n^3)$. Dichos algoritmos poseen una complejidad temporal aceptable con respecto a los estándares

de diseño, donde la cota superior $O(n^3)$ así se considera.

La propuesta descrita en este capítulo trata de solucionar algunos problemas encontrados en los trabajos existentes en la literatura. Por ejemplo, se realiza un pre-procesamiento de la información. Se utiliza como herramienta de minería de datos la combinación de agrupamientos, lo cual constituye un aporte a la literatura debido a que no se ha reportado su utilización en este tipo de problemas.

2.3. Conclusiones parciales

En este capítulo se explicó un algoritmo para la desambiguación del nombre de los autores en metadatos bibliográficos. La propuesta está basada principalmente en la distancia de edición y la combinación de agrupamientos. La utilización de estas técnicas constituye un aporte a la literatura ya que trata de resolver algunos de los problemas comunes en las soluciones propuestas. La solución expuesta difiere de otras aproximaciones debido a que se realiza un tratamiento previo a los datos utilizados. También se puede afirmar que la combinación de agrupamientos constituye una técnica avanzada en la minería de datos propiciando que los resultados obtenidos sean superiores a los existentes, puesto que estos se obtienen utilizando técnicas de minería de datos clásicas.

Capítulo 3

Validación de la solución propuesta

3.1. Introducción

En el capítulo anterior se explicaron las técnicas utilizadas para diseñar la propuesta de solución. Este capítulo tiene como objetivo la validación de dicha propuesta. Se describen los elementos desarrollados y empleados para realizar la validación. Se explica una aplicación informática para la desambiguación del nombre de los autores la cual utiliza los algoritmos desarrollados. También se definen las métricas para la evaluación de la solución. Finalmente se realizan los estudios experimentales diseñados.

Para una mejor comprensión del contenido de este capítulo se hace necesario definir el término experimento. Se refiere a un estudio de investigación en el que se manipulan deliberadamente una o más variables independientes (supuestas causas) para analizar las consecuencias de esa manipulación sobre una o más variables dependientes (supuestos efectos) dentro de una situación de control para el investigador[34]. Una definición más simple pero igualmente acertada es: un estudio que involucra la manipulación intencional de una acción para analizar sus posibles efectos[34]. Los experimentos se dividen en 3 grupos, pre-experimentos, cuasi-experimentos y experimentos puros.

Los experimentos puros difieren de los pre-experimentos y los cuasi-experimentos en el control sobre la situación experimental. Para realizar un experimento puro se debe:

1. Manipular una o más variables independientes.
2. Medir el efecto de la variable independiente sobre la variable dependiente.
3. Controlar la validez interna de la situación experimental.

Los pre-experimentos se caracterizan por no poseer un grupo de control o patrón para realizar la comparación. Esto significa que las posibilidades de manipular las variables independientes son reducidas. La principal característica de los cuasi-experimentos es que la asignación de los participantes a los grupos no se hace de forma aleatoria ni por emparejamiento.

3.2. Aplicación informática para la desambiguación del nombre de los autores

En la investigación se desarrolló una aplicación informática para realizar la desambiguación del nombre de los autores. La herramienta permite importar un conjunto de datos en formato XML[1] como se definió en el capítulo anterior y realizar la desambiguación. Como salida, se genera un fichero en formato XML con los resultados obtenidos, estructurado de la forma descrita anteriormente, además de actualizar el conjunto de datos de entrada. Para el desarrollo de la aplicación se empleó el lenguaje de programación Java[2] en su versión 7 y la API[3] de la herramienta Weka[4] para la implementación del algoritmo desarrollado.

La aplicación permite realizar las acciones necesarias para desambiguar el nombre de los autores existentes en los datos de entrada. Primeramente permite la importación de un fichero XML donde están contenidos los datos a analizar, brindando información referente al fichero importado. Se realiza el proceso de desambiguación y muestra una tabla, agrupando los nombres clasificados como ambiguos. Seguidamente permite seleccionar, de los nombres agrupados, cuales se refieren al mismo autor, además de seleccionar cuál se tomará como guía para actualizar los restantes seleccionados. También brinda información detallada de cada autor seleccionado como ambiguo. Finalmente genera un fichero XML con los resultados obtenidos y

[1] *http://www.w3.org/XML/*

[2] *http://java.com*

[3] *Interfaz de programación de aplicaciones (del inglés Application Programming Interface)*

[4] *http://www.cs.waikato.ac.nz/-ml/weka/*

actualiza el conjunto de datos de entrada.

La Figura 3.1 muestra la desambiguación realizada por la aplicación sobre un conjunto de autores dado mostrando los grupos obtenidos. El panel de la izquierda muestra información sobre el fichero analizado. El panel de la derecha muestra los grupos formados por la desambiguación además de las acciones que es posible realizar sobre los nombres ambiguos, por ejemplo, mostrar información detallada sobre un autor seleccionado.

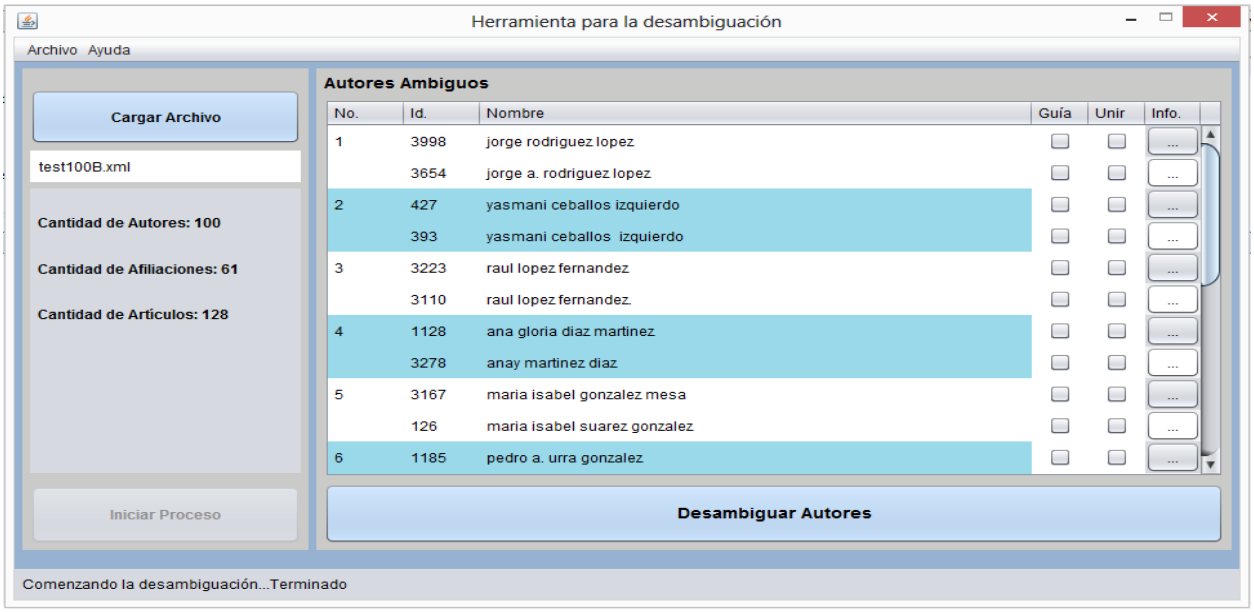


Figura 3.1: Aplicación para desambiguación del nombre de los autores

3.3. Métricas de evaluación

Para la validación de la solución desarrollada se emplearon métricas que evalúen los resultados obtenidos y demuestren su viabilidad. De acuerdo a la revisión de la literatura realizada[5][19][22][23] las métricas más empleadas son **Precisión**, **Exactitud**, **Recall** y **F- Measure**.

Estas métricas son empleadas en problemas relacionados con la recuperación de información[35]. Se usan para evaluar los resultados de las búsquedas realizadas sobre un conjunto de datos determinado. Por otra parte, pueden aplicarse a problemas de clasificación supervisada [36] para medir el rendimiento de un algoritmo de clasificación. Como se menciona en [36], tienen en cuenta la dispersión de los datos analizados y el balance de los grupos formados por tanto son ideales para la validación de la solución propuesta . El balance de los grupos formados es un elemento de importancia debido a que la solución agrupa todos los autores no ambiguos en un conjunto, afectando el equilibrio de los agrupamientos formados.

El proceso de desambiguación del nombre de los autores se puede modelar como un problema de clasificación. Se trata de encontrar los autores que hacen referencia a la misma persona o a personas diferentes. Como se menciona en el párrafo anterior las métricas son empleadas en la clasificación supervisada, esto significa que para cada uno de los autores clasificados como ambiguos es necesario verificar si en realidad se refieren a la misma persona. Para realizar esta verificación se desarrolló un trabajo manual donde: se buscó información referente a cada uno de los autores clasificados como ambiguos en un grupo para comprobar la clasificación realizada. Lo anterior permite seleccionar las métricas mencionadas como elementos de validación para la investigación realizada.

A continuación se explican cada una de las métricas de validación:

3.3.1. Precisión

El término Precisión (*P*, del inglés Precision) es la proporción de los casos predichos positivos que fueron correctos. En el caso particular de la desambiguación del nombre de los autores, se refiere a la proporción de los autores ambiguos agrupados correctamente. A continuación se muestra la ecuación que define la métrica.

$$P = \frac{VP}{FP + VP} \tag{3.1}$$

3.3.2. Exactitud

El término Exactitud (*A*, del inglés Accuracy) es la proporción de la cantidad total de predicciones que fueron correctas. En el caso del problema en cuestión es la proporción de la cantidad de autores que fueron clasificados correctamente. A continuación se muestra la ecuación que define la métrica.

$$A = \frac{VP + VN}{P + N} \tag{3.2}$$

3.3.3. Recall

El término *Recall* (*R*) es una medida de completitud y representa el porciento de predicciones correctas que fueron etiquetadas como tal. En el caso del problema en cuestión es el porciento de autores ambiguos clasificados como ambiguos. A continuación se muestra la ecuación que define la métrica.

$$R = \frac{VP}{VP + FN} \tag{3.3}$$

3.3.4. F-Measure

El término *F-Measure* (*F*) se refiere a una combinación de las métricas **Precisión** y **Recall**, asignándoles igual peso a ambas. A continuación se muestra la ecuación que define la métrica.

$$F = \frac{2 * P * R}{P + R} \tag{3.4}$$

Cada una de las métricas mostradas hacen referencia a términos en las ecuaciones que las definen. A continuación se describen cada uno de estos:

VP: Cantidad de autores ambiguos clasificados correctamente (Verdaderos Positivos).

FP: Cantidad de autores ambiguos mal clasificados (Falsos Positivos).

FN: Cantidad de autores clasificados como no ambiguos incorrectamente (Falsos Negativos).

P: Cantidad de autores clasificados como ambiguos (Positivos).

N: Cantidad de autores clasificados como no ambiguos (Negativos).

3.4. Resultados experimentales

Se definió un experimento para probar la validez del algoritmo desarrollado, para ello se emplea la aplicación informática presentada anteriormente y las métricas de validación definidas en la sección anterior.

3.4.1. Diseño experimental

De acuerdo a la clasificación de los experimentos mostrada al inicio del capítulo, se emplea en la investigación un cuasi-experimento. Los participantes de los grupos experimentales confeccionados no fueron asignados aleatoriamente ni por emparejamientos lo cual permite clasificar el experimento diseñado como tal.

Se definieron 3 grupos de autores diferenciados por el volumen de datos (100, 200 y 300 autores respectivamente). Cada grupo está compuesto por 3 conjuntos de datos diferenciados entre sí por el nivel de ambigüedad. Estos niveles son alto, medio y bajo respectivamente refiriéndose al 75 %, 50 % y 25 % de autores ambiguos. A continuación se muestra una tabla con el diseño experimental propuesto.

		Aplicación de la solución
R $G_{100Alta}$	X_1	O_1
R $G_{100Media}$	X_1	O_2
R $G_{100Baja}$	X_1	O_3
R $G_{200Alta}$	X_1	O_4
R $G_{200Media}$	X_1	O_5
R $G_{200Baja}$	X_1	O_6
R $G_{300Alta}$	X_1	O_7
R $G_{300Media}$	X_1	O_8
R $G_{300Baja}$	X_1	O_9

Tabla 3.1: Diseño experimental propuesto

La simbología utilizada en la tabla anterior es la siguiente:

- G_{XY} : Grupo de participantes, el subíndice X representa el grupo de datos conformado, los posibles valores que puede tomar son $100, 200$ *y* 300. El subíndice Y representa el nivel de ambigüedad en los conjuntos de datos, los posibles valores que puede tomar son *Alta (*$75\,\%$*), Media (*$50\,\%$*) y Baja (*$25\,\%$*)*.
- R: Asignación al azar de los participantes en cada uno de los grupos de autores conformados.
- X_1 : Tratamiento o estímulo, en este caso la aplicación del algoritmo propuesto.
- O_x : Observación realizada luego de la aplicación del algoritmo propuesto.

Seguidamente se aplica la solución propuesta a cada uno de los grupos y se calculan las métricas definidas en las secciones anteriores, facilitando la utilización del diseño experimental propuesto para cada cálculo de las métricas.

Las observaciones O_1, O_2, O_3 representan el cálculo de las métricas definidas para el grupo correspondiente al tamaño de 100 autores. Las observaciones O_4, O_5, O_6 se refieren al cálculo de las métricas para el grupo de 200 autores. Finalmente las observaciones O_7, O_8, O_9 son los cálculos de las métricas para el grupo que posee 300 autores en su registro. Las mediciones se basan en la comparación de los resultados del algoritmo propuesto con su respectivo conjunto de control (patrón). El objetivo de las observaciones es detectar un aumento en la unicidad en los nombres de los autores.

3.4.2. Características de los datos

La revisión de la literatura permitió identificar una serie de conjuntos de datos utilizados en validaciones de aproximaciones similares. Ninguno de estos conjuntos de datos pueden ser adoptados por la investigación debido a que todos están compuestos por nombres en idiomas como inglés, alemán y francés. Muy pocos casos poseen nombres en idioma español lo cual dificulta su utilización en la validación de la solución. De acuerdo a lo anterior se hace necesario confeccionar un conjunto de datos con las características necesarias para garantizar que la validación diseñada sea correcta.

Los datos fueron obtenidos de revistas científicas cubanas que diseminan sus metadatos sobre el protocolo OAI-PMH[1] cuyas temáticas oscilan entre la medicina, las ciencias de la información y computación[37]. La selección de los registros bibliográficos se basó en el método de muestreo aleatorio simple[38] inicialmente, garantizando la aleatoriedad de los datos y la representatividad de las muestras seleccionadas. Seguidamente a cada uno de los conjuntos de datos conformados se le introdujeron los registros de autores ambiguos garantizando la proporcionalidad mencionada anteriormente.

La tabla que se muestra a continuación resume las características de los grupos confeccionados.

Cantidad de autores	Cantidad de autores ambiguos	Cantidad de artículos	Cantidad de afiliaciones
100	75	160	82
	50	146	79
	25	128	61
200	150	297	152
	100	281	191
	50	250	83
300	225	441	237
	150	404	159
	75	343	103

Tabla 3.2: Descripción de los conjuntos de datos usados en la experimentación

La tabla muestra la cantidad de autores ambiguos para cada uno de los conjuntos de datos confeccionados, la cantidad de artículos (títulos) y afiliaciones. Estos elementos son de importancia para la investigación debido a que son usados como componentes de los vectores de similitud generados en la propuesta de solución.

3.4.3. Análisis de los resultados

De acuerdo al diseño experimental propuesto se realizaron varias pruebas a cada uno de los grupos conformados encaminadas al cálculo de las métricas definidas. Se realizaron con el objetivo de detectar factores que influyen en los resultados obtenidos, en este caso el nivel de ambigüedad en los datos y el volumen de información.

La primera métrica calculada en cada uno de los grupos de autores formados y sus respectivos conjuntos de datos fue **precisión**.

Los resultados obtenidos muestran variaciones entre los grupos experimentales de 100 autores y el resto. También se puede apreciar que apenas existen variaciones entre los grupos de 200 y 300 autores. Los resultados obtenidos son analizados por el nivel de ambigüedad. El grupo con ambigüedad baja posee una precisión media de 90 %, a diferencia de los grupos de media y alta ambigüedad que poseen 96 % y 93 % respectivamente. El análisis anterior permite concluir que la solución se desempeña con mayor **precisión** cuando el volumen de datos es elevado y la ambigüedad también (más del 50 % de los autores ambiguos).

La tabla siguiente muestra los resultados obtenidos en los cálculos de la precisión con los grupos de pruebas conformados.

Cantidad de autores	**Nivel de ambigüedad**	**Precisión**
100	Alta	0.91
	Media	0.92
	Baja	0.83
200	Alta	0.97
	Media	0.98
	Baja	0.92
300	Alta	0.91
	Media	0.99
	Baja	0.95

Tabla 3.3: Resultados de la métrica precisión

A partir de los datos obtenidos en las pruebas realizadas para la métrica precisión se aplicó la medida de tendencia central (media) y se calculó su desviación estándar.

La desviación estándar de una serie de mediciones es el promedio de desviación de cada una de las mediciones con respecto a la media de las mismas. Cuanto mayor es la dispersión de los datos con respecto a

la media mayor es la desviación estándar. A continuación se muestra la ecuación que define el promedio mencionado.

$$S = \sqrt{\frac{\sum(X - X_m)^2}{N}} \tag{3.5}$$

En la ecuación anterior la variable X representa los valores de las mediciones realizadas, la variable X_m representa la media de las mediciones obtenidas. La variable N representa la cantidad de mediciones realizadas.

Para el cálculo de la desviación estándar se sigue el procedimiento descrito a continuación:

1. Se ordenan las mediciones.
2. Se calcula la media de las mediciones realizadas.
3. Se determina la desviación de cada medición con respecto a la media.
4. Se eleva al cuadrado cada desviación y se obtiene la sumatoria de las desviaciones elevadas al cuadrado o $\sum(X - X_m)^2$.
5. Se aplica la formula con los valores obtenidos.

A partir de lo anterior se obtiene un valor medio de la precisión de **0.93** o **93 %** con una desviación estándar de **0.09**.

Luego de analizada la precisión de la solución se calculó su **exactitud**. El valor medio obtenido para el grupo de 100 autores fue de 93 % mientras que la de los grupos de 200 y 300 fue de 95 % y 96 % respectivamente. La exactitud media obtenida fue de 93 % para los grupos de autores con nivel alto de ambigüedad y de 95 % y 96 % para los grupos con ambigüedad baja y media respectivamente. Lo anterior permite concluir que la exactitud de la solución propuesta está condicionada por el grado de ambigüedad y la cantidad de datos analizados, siendo el grado de ambigüedad alto y el bajo volumen de datos los elementos

que más afectan su exactitud.

La tabla que se muestra a continuación resume los resultados obtenidos en las pruebas realizadas.

Cantidad de autores	**Nivel de ambigüedad**	**Exactitud**
100	Alta	0.88
	Media	0.96
	Baja	0.96
200	Alta	0.98
	Media	0.90
	Baja	0.98
300	Alta	0.93
	Media	0.99
	Baja	0.96

Tabla 3.4: Resultados de la métrica exactitud

Al igual que en el caso de la métrica precisión se calcula la media de los resultados y su respectiva desviación estándar para la métrica **Exactitud**. A partir de los valores obtenidos en las mediciones realizadas se obtuvo una media de **0.95 o 95 %** y una desviación estándar de **0.03**.

La métrica *Recall* está condicionada por la clasificación de los autores ambiguos, tanto los que se refieren a la misma persona como los que no. La solución desarrollada en ninguna de las pruebas realizadas clasificó incorrectamente autores no ambiguos propiciando la obtención de un 100 % para esta métrica.

Finalmente se calculó la métrica *F-Measure*. El grupo de autores con 100 registros es el más afectado con un valor medio de 94 % mientras que para los grupos de 200 y 300 autores se obtuvo un 96 % y 97 % respectivamente. En el caso de los conjuntos de datos con un nivel de ambigüedad bajo (25 % de los autores)

el valor medio obtenido fue de 93 % mientras que para los conjuntos de datos de medio y alto nivel de ambigüedad se obtuvo un valor de 98 % y 96 % respectivamente. A partir de lo anterior se puede concluir que la solución propuesta obtiene mejores resultados para la métrica *F-Measure* cuando el volumen de datos es elevado y el nivel de ambigüedad también.

A continuación se muestra una tabla donde se resumen los resultados obtenidos en el cálculo de la métrica *F-Measure*.

Cantidad de autores	**Nivel de ambigüedad**	**F-Measure**
100	Alta	0.95
	Media	0.96
	Baja	0.90
200	Alta	0.98
	Media	0.99
	Baja	0.95
300	Alta	0.95
	Media	0.99
	Baja	0.93

Tabla 3.5: Resultados de la métrica F-Measure

Con los resultados mostrados se obtiene un valor medio para la métrica *F-Measure* de **0.96 o 96 %** mientras que su desviación estándar es de **0.02**.

Los resultados obtenidos reflejan las condiciones en que la solución propuesta se comporta de mejor forma, así como los casos peores. En todas las métricas calculadas uno de los factores que propició buenos resultados fue el alto volumen de datos.

El análisis realizado permite concluir que el nivel de ambigüedad influye en los resultados obtenidos. En el caso de las métricas precisión, *Recall* y *F-Measure* los niveles de ambigüedad alto y medio arrojaron mejores resultados, siendo lo contrario en el caso de la exactitud. Esto puede tomarse en consideración para posibles escenarios donde los niveles de exactitud de la solución no requieran ser los mejores pero si otras métricas, como por ejemplo, la precisión.

Como se menciona en el capítulo 1 existe una relación de proporcionalidad inversa entre la ambigüedad y la unicidad del nombre de los autores en metadatos bibliográficos. Los valores obtenidos con las métricas calculadas demuestran que se disminuye el nivel de ambigüedad en los datos analizados por tanto la unicidad aumenta.

La métrica **precisión** establece la proporción de los autores clasificados como ambiguos que en realidad se refieren a la misma persona. Teniendo un 93 % de precisión en los resultados obtenidos por el algoritmo, se puede afirmar que: de cada 100 autores ambiguos que se refieren a la misma persona, 93 son clasificados correctamente, disminuyendo la ambigüedad de los autores analizados, aumentando su unicidad.

La métrica **exactitud** establece la proporción de los autores clasificados correctamente, tanto ambiguos como no ambiguos. Teniendo un 95 % de exactitud en los resultados obtenidos se puede afirmar que: de cada 100 autores analizados, 95 de ellos se clasifican correctamente. Esto disminuye el nivel de ambigüedad en los nombres de los autores lo que aumenta su unicidad.

3.5. Conclusiones parciales

La validación de la solución propuesta estuvo conducida por la realización de experimentos, demostrando la validez de los resultados obtenidos. Los experimentos se llevaron a cabo con la utilización de la aplicación informática desarrollada. El análisis realizado sobre los resultados obtenidos permiten afirmar que la solución

desarrollada obtiene mejores resultados cuando el volumen de datos es elevado y en la mayoría de los casos el nivel de ambigüedad también. Para la métrica **exactitud** los mejores resultados son obtenidos cuando el grado de ambigüedad es menor, esto puede tomarse en consideración en los escenarios donde la exactitud de la respuesta no sea un elemento crítico y si lo sea, por ejemplo, la precisión. Finalmente se puede afirmar que se logra disminuir la ambigüedad en el nombre de los autores, aumentando así su unicidad.

Conclusiones

En la presente investigación se plantearon una serie de objetivos los cuales se fueron cumpliendo progresivamente, permitiendo arribar a las siguientes conclusiones:

- El problema de la ambigüedad en el nombre de los autores es abordado siguiendo 4 enfoques: (1) técnicas de clasificación, (2) técnicas de agrupamiento, (3) modelos probabilísticos y (4) usando una combinación de los dos primeros. Con respecto a los datos usados en las soluciones se identificaron dos aristas: (1) soluciones que usan los metadatos bibliográficos existentes en los repositorios digitales y (2) soluciones que usan la web como fuente de información a través de motores de búsqueda.

- Las soluciones existentes en la literatura poseen una serie de limitaciones que imposibilitan su aplicación en el contexto de investigación, entre las que se encuentran: (1) no se realiza pre-procesamiento de la información y (2) no están orientadas a las particularidades del idioma español.

- La solución propuesta está basada en las particularidades del idioma español. Realiza pre-procesamiento de la información con el objetivo de elevar la calidad de los datos utilizados a través de su normalización. Permite la detección de los nombres de los autores ambiguos haciendo uso de la distancia de edición y la combinación de agrupamientos, esta última constituye un aporte a la literatura debido a que no se ha reportado su utilización en este tipo de problemas.

- Los experimentos desarrollados haciendo uso de la aplicación informática propuesta demostraron la viabilidad de la solución garantizando un aumento en la unicidad del nombre de los autores en metadatos bibliográficos.

Recomendaciones

A partir del estudio realizado en la presente investigación se proponen las siguientes recomendaciones a tener en cuenta en la solución desarrollada:

- Utilizar otro algoritmo de agrupamiento en la fase de combinación de agrupamientos de acuerdo a las limitaciones que posee el algoritmo empleado (K-Means).
- Realizar experimentos sobre conjuntos de datos cuyos nombres no estén representados solamente en el idioma español.
- Mejorar la complejidad temporal del algoritmo propuesto.

Referencias bibliográficas

[1] José Manuel Barrueco and Imma Subirats Coll. Open archives initiative. protocol for metadata harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo. *El Profesional de la Informacion*, 12(2):99 – 106, 2003. ISSN 1386-6710. URL http://elprofesionaldelainformacion.metapress.com/app/home/contribution.asp?referrer=parent&backto=issue,3,21;journal,66,93;linkingpublicationresults,1:105302,1.

[2] Tim Berners-Lee. Linked data, 2006. URL http://www.w3.org/DesignIssues/LinkedData.html.

[3] In-Su Kang, Pyung Kim, Seungwoo Lee, Hanmin Jung, and Beom-Jong You. Construction of a large-scale test set for author disambiguation. *Information Processing & Management*, 47(3):452–465, 2011-05. ISSN 0306-4573. URL http://www.sciencedirect.com/science/article/pii/S0306457310000865.

[4] Adriano Veloso, Anderson A. Ferreira, Marcos André Gon\ccalves, Alberto H. F. Laender, and Wagner Meira,Jr. Cost-effective on-demand associative author name disambiguation. *Information Processing & Management*, 48(4):680–697, 2012-07. ISSN 0306-4573. URL http://dx.doi.org/10.1016/j.ipm.2011.08.005.

[5] Anderson A. Ferreira, Tales Mota Machado, and Marcos André Gonçalves. Improving author name disambiguation with user relevance feedback. *Journal of Information and Data Management*, 3(3): 332, 2012-09-27. ISSN 21787107. URL http://seer.lcc.ufmg.br/index.php/jidm/article/view/200.

[6] PrakashM. Nadkarni. What is metadata? In *Metadata-driven Software Systems in Biomedicine*, Health Informatics, pages 1–16. Springer London, 2011. ISBN 978-0-85729-509-5. URL http://dx.doi.org/10.1007/978-0-85729-510-1_1.

[7] José Hernández Orallo, Ma José Ramírez Quintana, and César Ferri Ramírez. *Introducción a la Minería de Datos*, volume 1. Pearson Prentice Hall, 2004. ISBN 9788483225585. URL http://www.publidisa.com/preview-libro-9788483225585.pdf.

[8] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003-03. ISSN 1532-4435. URL http://dx.doi.org/10.1162/153244303321897735.

[9] Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. ISBN 9780262133609.

[10] Andrés Marzal and Enrique Vidal. Computation of normalized edit distance and applications. *IEEE*, 5 (19):926–932, 1993.

[11] Jesús Vilares. El modelo probabilístico: características y modelos derivados. *Revista General de Información y Documentación*, 18:345 – 363, 2009-02-24. ISSN 1988-2858. URL http://revistas.ucm.es/index.php/RGID/article/view/RGID0808110345A.

[12] Quan Lin, Bo Wang, Yuan Du, Xuezhi Wang, Yuhua Li, and Songcan Chen. Disambiguating authors by pairwise classification. *Tsinghua Science & Technology*, 15(6):668–677, 2010-12. ISSN 1007-0214. URL http://www.sciencedirect.com/science/article/pii/S1007021410701140.

[13] Treeratpituk Pucktada and Giles C Lee. Disambiguating authors in academic publications using random forest. In *9th ACM/IEEE-CS join conferens*, pages 39–48. ACM, 2009. URL http://doi.acm.org/10.1145/1555400.1555408.

[14] Jian Wang, Kaspars Berzins, Diana Hicks, Julia Melkers, Fang Xiao, and Diogo Pinheiro. A boosted-trees method for name disambiguation. *Scientometrics*, 93(2):391–411, 2012-11-01. ISSN 0138-9130, 1588-2861. URL http://link.springer.com/article/10.1007/s11192-012-0681-1.

[15] Jia Zhu, Gabriel Fung, and Liwei Wang. Efficient name disambiguation in digital libraries. In Haixun Wang, Shijun Li, Satoshi Oyama, Xiaohua Hu, and Tieyun Qian, editors, *Web-Age Information Management*, number 6897 in Lecture Notes in Computer Science, pages 430–441. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23534-4, 978-3-642-23535-1. URL http://link.springer.com/chapter/10.1007/978-3-642-23535-1_37.

[16] YANG, Kai-Hsiang, PENG, Hsin-Tsung, JIANG, Jian-Yi, LEE, Hahn-Ming, and HO, Jan-Ming. Author name disambiguation for citations using topic and web correlation. In *Proceedings of the 12th European conference*, page 185–196. Heidelberg: Springer-Verlag, 2008. URL http://dx.doi.org/10.1007/978-3-540-87599-4_19.

[17] KERN, Roman, ZECHNER, Mario, and GRANITZER, Michael. Model selection strategies for author disambiguation. In *Proceedings of the 2011 22nd International Workshop*, page 155–159. IEEE Computer Society, 2011. URL http://dx.doi.org/10.1109/DEXA.2011.54.

[18] Raffaella Bernardi and Dieu-Thu Le. Metadata enrichment via topic models for author name disambiguation. In *Proceedings of the 2009 international conference on Advanced language technologies for digital libraries*, NLP4DL'09/AT4DL'09, page 92–113. Springer-Verlag, 2011. ISBN 978-3-642-23159-9. URL http://dl.acm.org/citation.cfm?id=2039901.2039908.

[19] Muhammad Imran, Syed Zeeshan Haider Gillani, and Maurizio Marchese. A real-time heuristic-based unsupervised method for name disambiguation in digital libraries. *D-Lib Magazine*, 19(9/10), September 2013. ISSN 1082-9873. URL http://www.dlib.org/dlib/september13/imran/09imran.html.

[20] Vetle I. Torvik, Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. A probabilistic similarity metric for medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, 56(2):140–158, 2005. ISSN 1532-2890. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20105/abstract.

[21] TANG, Jie, ZHANG, Jing, ZHANG Duo, and LI, Juanzi. A unified framework for name disambiguation. In *Proceedings of the 17th international conference*, page 1205–1206. ACM, 2008. URL http://doi.acm.org/10.1145/1367497.1367728.

[22] J. Pricilla. An efficient framework for name disambiguition in digital library. *International Journal Of Engineering And Computer Science*, 2(4):1097–1105, 2013. ISSN 2319-7242. URL http://ijecs.in/ijecsissue/wp-content/uploads/2013/04/1097-1105ijecs.pdf.

[23] Shaohua Li, Gao Cong, and Chunyan Miao. Author name disambiguation using a new categorical distribution similarity. In *Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECML PKDD'12, page 569–584. Springer-Verlag, 2012. ISBN 978-3-642-33459-7. URL http://dx.doi.org/10.1007/978-3-642-33460-3_42.

[24] FERREIRA, Anderson A, VELOSO, Adriano, GONCALVES, Marcos André, and LAENDER, Alberto H.F. Effective self-training author name disambiguation in scholarly digital libraries. In *Proceedings of the 10th annual joint conference*, page 39–48. ACM, 2010. URL http://doi.acm.org/10.1145/1816123.1816130.

[25] Thomas Gurney, Edwin Horlings, and Peter van den Besselaar. Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 91(2):435–449, 2012-05-01. ISSN 0138-9130, 1588-2861. URL http://link.springer.com/article/10.1007/s11192-011-0589-1.

[26] Anderson A. Ferreira, Marcos André Gon\ccalves, Jussara M. Almeida, Alberto H. F. Laender, and Adriano Veloso. A tool for generating synthetic authorship records for evaluating author name disambiguation methods. *Information Sciences*, 206:42–62, 2012-11. ISSN 0020-0255. URL http://dx.doi.org/10.1016/j.ins.2012.04.022.

[27] Yu Cheng, Zhengzhang Chen, Jiang Wang, Ankit Agrawal, and Alok Choudhary. Bootstrapping active name disambiguation with crowdsourcing. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, page 1213–1216. ACM, New York, NY, USA, 2013. ISBN 978-1-4503-2263-8. URL http://doi.acm.org/10.1145/2505515.2507858.

[28] Jeffrey Beall. Metadata for name disambiguation and collocation. *Future Internet*, 2(1):1–15, 2010. ISSN 1999-5903. URL http://www.mdpi.com/1999-5903/2/1/1.

[29] Sandro Vega-Pons and Jose Ruis-Chulcloper. Combinación de agrupamientos: un estado del arte, 2010.

[30] Alfred V. Aho. *Estructuras de datos y algoritmos*. Addison-Wesley iberoamericana, 1988. ISBN 0201640244.

[31] José Luis Pinar Gámez. Identificación de autores en bases de datos bibliográficas, 2007.

[32] Ana Fred. Finding consistent clusters in data partitions. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, number 2096 in Lecture Notes in Computer Science, pages 309–318. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-42284-6, 978-3-540-48219-2. URL http://link.springer.com/chapter/10.1007/3-540-48219-9_31.

[33] Evgenia Dimitriadou, Andreas Weingessel, and Kurt Hornik. Voting-merging: An ensemble method for clustering. In Georg Dorffner, Horst Bischof, and Kurt Hornik, editors, *Artificial Neural Networks — ICANN 2001*, number 2130 in Lecture Notes in Computer Science, pages 217–224. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-42486-4, 978-3-540-44668-2. URL http://link.springer.com/chapter/10.1007/3-540-44668-0_31.

[34] Correa C. Rojas M. Grau, R. *Metodología de la Investigación. 2da edición*. EL POIRA, Editores S.A., Ibagué, Colombia, 2004.

[35] John Tait (Eds.) Sharon McDonald. *Advances in Information Retrieval*. 2004.

[36] Morgan Kaufmann. *Data Mining. Concepts and Techniques*. 2012.

[37] Yusniel Hidalgo-Delgado, Liudmila Reyes-Álvarez, Amed Leiva-Mederos, María del Mar Roldán-García, and José F. Aldana-Montes. Bm2lod: Platform for publishing bibliographic data as linked open data. In *Proceeding of 7th IADIS International Conference on Information Systems*, pages 27–34. IADIS Press, 2014.

[38] Steven K. Thompson. *Simple Random Sampling*, pages 9–37. John Wiley and Sons, Inc., 2012. ISBN 9781118162934. URL `http://dx.doi.org/10.1002/9781118162934.ch2`.

Anexo A

Análisis de la complejidad temporal de los algoritmos propuestos

A.1. Complejidad temporal Algoritmo similitud entre los co-autores

Algoritmo 1 Similitud entre los co-autores

Entrada: Lista de co-autores del autor A con longitud $L1$, Lista de co-autores del autor B con longitud $L2$.

Salida: Similitud entre los co-autores.

1: **Inicializar:** Hacer similitud mayor sim_may y suma similitud sum_sim variable con valor 0
2: **para** $i = 1$ hasta $L1$ **hacer**
3: **para** $j = 1$ hasta $L2$ **hacer**
4: Calcular $similitud$ entre el co-autor i-ésimo y el co-autor j-ésimo
5: **si** $similitud > sim\_may$ **entonces**
6: $sim\_may = similitud$
7: **fin si**
8: **fin para**
9: $sum\_sim+ = sim\_may$
10: **fin para**
11: **devolver** $sum\_sim/L1$

Para el cálculo de la complejidad temporal del anterior algoritmo calcularemos la complejidad de cada una de las sentencias que el mismo posee. La lista que se muestra a continuación corresponde con la numeración de las sentencias presentes en el algoritmo.

Entrada: Los datos que se pasan por parámetro son pasados directamente no por referencia por tanto la complejidad de esta linea es de $O(n)$.

Línea 1 La inicialización de las variables se considera una operación básica por tanto la complejidad temporal de esta linea es de $O(1)$.

Línea 2 La complejidad temporal de un ciclo depende de evaluar la condición de parada, de la cantidad de iteraciones del ciclo y del coste de ejecutar el cuerpo del ciclo. En la condición del *para* aparece una asignación sobre i que se ejecuta sólo una vez y que debemos contabilizar independientemente del ciclo, esta asignación tiene coste $O(1)$. También hay otra asignación sobre i con coste $O(1)$ que se ejecuta tantas veces como iteraciones realiza el ciclo. Esta asignación la tendremos en cuenta cuando calculemos el coste del cuerpo del ciclo. Para calcular el coste del *para* necesitamos el número de iteraciones que efectúa el ciclo denominada n, y el coste del cuerpo del ciclo.

Línea 3 En el cuerpo del ciclo anterior existe otro ciclo *para*, como en el caso del *para* anterior debemos tener en cuenta la evaluación de la condición de parada, la cantidad de iteraciones del ciclo y el coste de ejecutar el cuerpo del ciclo.

Línea 4 El calculo de la complejidad temporal de calcular la similitud entre el co-autor i-ésimo y el j-ésimo es de $O(1)$ debido a que es una operación básica.

Línea 5 La complejidad de una sentencia condicional está dada por coste de evaluar la condición, en este caso es $O(1)$ y el coste de ejecutar el cuerpo de la condición, en este caso se trata de una operación básica, una asignación por tanto la complejidad temporal es de $O(1)$.

Línea 9 La complejidad temporal del incremento y luego de la asignación de la variable es $O(1)$ debido a que ambas operaciones se consideran operaciones básicas($O(1)$).

Línea 11 La complejidad temporal en este caso es de $O(1)$ debido a que esta operación se considera una operación de coste constante $O(1)$.

La complejidad temporal de un algoritmo está condicionado por el coste máximo de las sentencias que lo componen. En el caso de dicho algoritmo existen 2 ciclos anidados, la complejidad temporal de primero

depende de la complejidad del segundo. El segundo ciclo tiene una complejidad temporal de $O(n)$. Por tanto la complejidad temporal de dicho algoritmo es $O(n^2)$.

A.2. Complejidad temporal Algoritmo de desambiguación

Algoritmo 2 Algoritmo de desambiguación

Entrada: Lista de objetos a desambiguar con longitud N, $umbral$ de comparación.

Salida: Lista de autores desambiguados.

1: **Inicializar:** Hacer una lista de autores parecido aut_par vacía, Lista resultado $result\_list$ de autores desambiguados.
2: **para** $i = 0$ hasta N **hacer**
3: $aut\_par.add(objeto\ i - esimo)$
4: **para** $i = i + 1$ hasta N **hacer**
5: Calcular $similitud$ entre el objeto i-ésimo y j-ésimo.
6: **si** $similitud > umbral$ **entonces**
7: $aut\_par.add(objeto\ j - esimo)$
8: Calcular sim_vec entre los elementos restantes de los datos de entrada presentes en aut_par.
9: **fin si**
10: **fin para**
11: Combinación de agrupamientos().
12: Adicionar un objeto de cada agrupamiento formado a la lista de autores desambiguados
13: **fin para**
14: **devolver** $result\_list$

El análisis de la complejidad del algoritmo 3 está dado por el coste de cada una de las sentencias que lo componen. A continuación se analizan cada una.

Entrada Los datos que se pasan por parámetro son pasados directamente no por referencia por tanto la complejidad de esta linea es de $O(n)$.

Línea 1 La inicialización de las variables es una operación básica por tanto el coste de la línea es de $O(1)$.

Línea 2 La complejidad temporal de un ciclo depende de evaluar la condición de parada, de la cantidad de iteraciones del ciclo y del coste de ejecutar el cuerpo del ciclo. En la condición del *para* aparece una asignación sobre i que se ejecuta sólo una vez y que debemos contabilizar independientemente del ciclo, esta asignación tiene coste $O(1)$. También hay otra asignación sobre i con coste $O(1)$ que se ejecuta tantas veces como iteraciones realiza el ciclo. Esta asignación la tendremos en cuenta cuando calculemos el coste del cuerpo del ciclo. Para calcular el coste del *para* necesitamos el número de iteraciones que efectúa el ciclo denominada n, y el coste del cuerpo del ciclo.

Línea 3 La operación *add* posee una complejidad de $O(1)$ por tanto la complejidad temporal de la linea es de $O(1)$.

Línea 4 La complejidad temporal de un ciclo se explicó en la línea 2, por tanto se hace necesario calcular el coste de las sentencias del cuerpo del ciclo.

Línea 5 El cálculo de la similitud entre los objetos se considera una operación básica por tanto la complejidad de esta línea sea de $O(1)$.

Línea 6 La complejidad de una sentencia condicional está dada por coste de evaluar la condición, en este caso es $O(1)$ y el coste de ejecutar el cuerpo de la condición, en este caso se existen 2 sentencias, una contiene la operación *add* la cual es una operación básica, por tanto la complejidad es de $O(1)$, la otra sentencia es una asignación por tanto la complejidad temporal es de $O(1)$. Lo anterior produce una complejidad total de $O(1)$ para la sentencia condicional.

Línea 9 La llamada a un procedimiento está condicionado por la complejidad temporal del procedimiento, en este caso posee una complejidad temporal de $O(n^2)$, por tanto la complejidad temporal de la línea es de $O(n^2)$.

Línea 10 La acción de adicionar un objeto es una operación básica por tanto la complejidad de la línea es $O(1)$.

La complejidad temporal de un algoritmo está condicionado por el coste máximo de las sentencias que lo componen. En el caso de dicho algoritmo existen 2 ciclos anidados, la complejidad temporal de primero depende de la complejidad del segundo. El segundo ciclo tiene una complejidad temporal de $O(n^3)$ debido a que en su cuerpo posee una sentencia de complejidad $O(n^2)$. Por tanto la complejidad temporal de dicho algoritmo es $O(n^4)$.