Desambiguación del nombre de los autores en revistas científicas

ARTICLE · JULY 2014

READS
38

3 AUTHORS:

Luis Enrique Alonso Sierra
DATYS
5 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Yusniel Hidalgo Delgado
University of Information Sciences
16 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Amed Leiva Mederos
Central University "Marta Abreu" of Las Villas
18 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 3, Julio-Septiembre, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 131-150
http://rcci.uci.cu

Tipo de artículo: Artículo de revisión
Temática: Inteligencia artificial
Recibido: 14/11/2013 | Aceptado: 24/06/2014

Desambiguación del nombre de los autores en revistas científicas

Disambiguation of Names of Authors in Scientific Journals

Luis Enrique Alonso Sierra[1*], Yusniel Hidalgo Delgado[2], Amed Abel Leiva Mederos[3]

[1*] Grupo de Web Semántica. Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. CP.: 19370.
[2] Grupo de Web Semántica. Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. CP.: 19370. Correo-e: yhdelgado@uci.cu
[3] Departamento de Ciencias de la Información, Facultad de Ingeniería Industrial, Universidad Central "Marta Abreu" de las Villas, Villa Clara, Cuba. Correo-e: amed@uclv.edu.cu

[*] Autor para correspondencia: lealonso24@gmail.com

Resumen

La ambigüedad en el nombre de los autores en las revistas científicas es un problema que afecta a las publicaciones de este tipo. Dicho problema se refiere a la posibilidad de representar el nombre de los autores de diferentes formas en los metadatos bibliográficos presentes en los repositorios digitales. Este se puede manifestar de dos formas diferentes, (1) pueden aparecer nombres de autores iguales, pero que no se refieren al mismo autor y (2) aparecen nombres diferentes, pero que se refieren al mismo autor. En este artículo se presenta un análisis crítico de las principales aproximaciones existentes en la literatura para solucionar el problema antes mencionado. Se realizó una revisión bibliográfica en las principales Bases de Datos referenciadas a nivel mundial, con el objetivo de identificar los elementos más actuales y rigurosos posibles. Se pudo constatar que la variedad de técnicas utilizadas para resolver el problema de la ambigüedad abarcan desde la utilización de técnicas de minería de datos hasta la utilización de la web como fuente de información. Cada una de las soluciones planteadas posee limitaciones y ventajas que dependen de las características de los datos utilizados. Finalmente, se concluye que no existe una solución definitiva para resolver el

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 3, Julio-Septiembre, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 131-150
http://rcci.uci.cu

problema tratado debido a que los resultados de las aproximaciones no son cien por ciento completos y dependen estrechamente de los datos utilizados.
Palabras clave: bibliometría, desambiguación, minería de datos, nombre de autor, revistas científicas.

Abstract

The ambiguity in the names of authors in scientific journals is a problem that affects such publications. This problem concerns the possibility of representing the name of the authors of different ways in bibliographic metadata inside of digital repositories. This problem can manifest itself in two different ways. At first place, it is possible to find names of authors syntactically identical, but that do not refer to the same author. The second case refers to the appearance of different names that refer to the same author. This paper shows a study of the main approaches found in the literature to solve the above problem, in addition to a critical analysis of these solutions. To carry out research, a literature review was conducted in major databases referenced globally, with the goal of possibly exposing the latest and thorough elements. After having conducted the study, it was found that the variety of techniques used to solve the problem of ambiguity, range from the use of data mining techniques to the use of the web as an information source. Each of the proposed solutions has advantages and drawbacks, depending on the characteristics of the data used. Also it can be concluded that there is an ultimate solution for solving the problem addressed, due to its close dependence on the used data.
Keywords: *author name, bibliometrics, data mining, disambiguation, journals.*

Introducción

Con el creciente desarrollo de las nuevas tecnologías de la información, internet se ha convertido en una fuente de información importante para los investigadores. Con la aparición de la web 2.0 los usuarios de internet han tenido la posibilidad de publicar información en la red que muchas veces no cuenta con la calidad requerida, esto ha determinado la necesidad de encontrar formas de almacenar y publicar el conocimiento científico siguiendo normas rigurosas para la publicación de dicho conocimiento.

Una de las formas de almacenamiento y publicación del conocimiento científico es a través de las revistas científicas. Estas no son más que publicaciones periódicas en la que se intenta recoger el progreso de la ciencia, entre otras cosas, incluyendo informes sobre las nuevas investigaciones (Lawson, 2000).

En la actualidad el desarrollo de las tecnologías y la evolución del conocimiento han tomado un paso acelerado, las nuevas tecnologías, son usadas como herramientas para el desarrollo de otras. Debido a este fenómeno el volumen de

información almacenada en las revistas científicas tiende a ser elevado, lo que dificulta la facilidad y rapidez en la obtención de la información que requieren los investigadores para realizar sus estudios. Por este motivo se han desarrollado técnicas y procedimientos para facilitar el trabajo a los investigadores. La bibliometría es uno de estos elementos, ésta se define como una parte de la cienciometría que aplica métodos matemáticos y estadísticos a toda la literatura de carácter científico y a los autores que la producen, con el objetivo de estudiar y analizar la actividad científica (Pérez, 2002).

La bibliometría por su parte usa los metadatos bibliográficos de un determinado conjunto de información al cual se le pretenden realizar estudios bibliométricos. Por su parte los metadatos son un conjunto de datos estructurados y codificados que describen características de instancias, conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas (M.d, 2011).

Uno de los elementos importantes en los estudios bibliométricos son los nombres de los autores, si estos no cuentan con la calidad requerida los estudios bibliométricos no tendrán los resultados esperados. Entiéndase como calidad de los nombres de los autores a la correcta representación de los mismos en las bases de datos científicas. Este problema está relacionado con la ambigüedad del nombre de los autores el cual se refiere a la posibilidad de representar de diferentes formas los nombres de los mismos y su respectiva identificación en las revistas científicas.

La ambigüedad en el nombre de los autores puede evidenciarse de dos formas diferentes. La primera es que pueden encontrarse nombres sintácticamente iguales que no se refieren al mismo autor. Esta situación es la más compleja de resolver debido a que la representación del nombre es uno de los elementos más importantes en el proceso de desambiguación. Muchas de las soluciones estudiadas asumen o parten del hecho de que si dos nombres son iguales entonces estos se refieren al mismo autor.

La segunda y más común es que pueden aparecer nombres sintácticamente diferentes para referirse a un mismo autor. Un elemento importante en esta situación es el idioma de procedencia en que esté representado el nombre del autor. En idiomas como el inglés y el francés, los autores suelen representarse con un nombre y un apellido, lo cual disminuye la complejidad del proceso de desambiguación. En idiomas como el español los nombres son representados con un nombre (en ocasiones hasta nombres compuestos) y dos apellidos, trayendo consigo la posibilidad de

representar dicho nombre de diversas maneras. Por ejemplo, un nombre común en el idioma español puede ser: *Luis Enrique Alonso Sierra*, algunas formas de representar el mismo pueden ser las siguientes:

- *Luis E. Alonso*
- *Luis Enrique Alonso*
- *Luis E. Alonso Sierra*
- *Luis Enrique Alonso Sierra*
- *L. E. Alonso Sierra.*

Como se puede apreciar la variedad puede ser amplia, esto sin tener en cuenta los errores de escritura que pueden aparecer en los mismos.

De acuerdo a la bibliografía consultada y por las características del problema en cuestión, la principal herramienta utilizada para solucionar el problema de la ambigüedad del nombre de los autores es mediante la aplicación de técnicas de minería de datos. La minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Hernández, 2002). Por su parte la minería de datos es un área del conocimiento amplia de la cual se derivan un conjunto de técnicas que hacen posible la aplicación de la misma en diversas áreas de la sociedad. A continuación se mencionan algunas de las más importantes para la presente investigación.

Una de las técnicas que se desarrollaron como parte de la minería de datos es el Aprendizaje Supervisado. El mismo es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores) donde una componente del par son los datos de entrada y el otro los resultados deseados. Por otro lado existe otra técnica llamada Aprendizaje no Supervisado. Este es 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.

Por otra parte, los algoritmos de Agrupamiento son procedimientos de agrupación de una serie de vectores de acuerdo con un criterio de cercanía. Esta cercanía se define en términos de una determinada función de distancia, como la euclídea, aunque existen otras más robustas o que permiten extenderla a variables discretas.

Existe otro tipo de técnica que se encarga de Clasificar las relaciones establecidas en correctas e incorrectas, a esta se le conoce como Clasificador. Un clasificador no es más que un algoritmo utilizado 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.

Por otro lado, los Modelos Probabilísticos son formas 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 (Vilares, 2009).

En la presente investigación se realiza una valoración crítica de las principales aproximaciones existentes para desambiguar el nombre de los autores en publicaciones científicas. Se pretende además, identificar las principales herramientas, métodos y técnicas utilizadas en la solución del problema objeto de estudio. También se pretende analizar de forma crítica los principales problemas detectados por los investigadores en las soluciones estudiadas. Esta investigación pretende servir como un punto de partida para el desarrollo de soluciones a la ambigüedad del nombre de los autores.

En las secciones siguientes se describen los materiales y métodos usados para la realización de la investigación. Luego se describen cada una de las soluciones estudiadas, clasificándolas de acuerdo a las técnicas usadas para su realización. Además se describen algunas iniciativas que han surgido con el objetivo de solventar el problema de la ambigüedad a través de metadatos públicos en la web. También se realiza un análisis de las principales deficiencias encontradas en dichas soluciones. Por último se presentan las conclusiones de la investigación.

Desarrollo

Para la realización del estudio documental se consultaron numerosas fuentes bibliográficas, como revistas científicas indexadas en SciELO y SCOPUS. Se identificaron las principales revistas científicas a las cuales se les han aplicado soluciones para la desambiguación del nombre de los autores y además se visitaron sus sitios oficiales para lograr una mayor veracidad en la información mostrada. Además, se consultaron libros de autores que están relacionados con la Minería de Datos en los últimos 5 años.

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 3, Julio-Septiembre, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 131-150
http://rcci.uci.cu

Las aproximaciones estudiadas pueden dividirse en dos grandes grupos de acuerdo a (1) la técnica de minería de datos utilizada, (2) la fuente de datos utilizada. 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, (4) soluciones que usan una combinación de los métodos anteriores. Luego, la clasificación de las soluciones de acuerdo a fuente de datos utilizados se pueden 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.

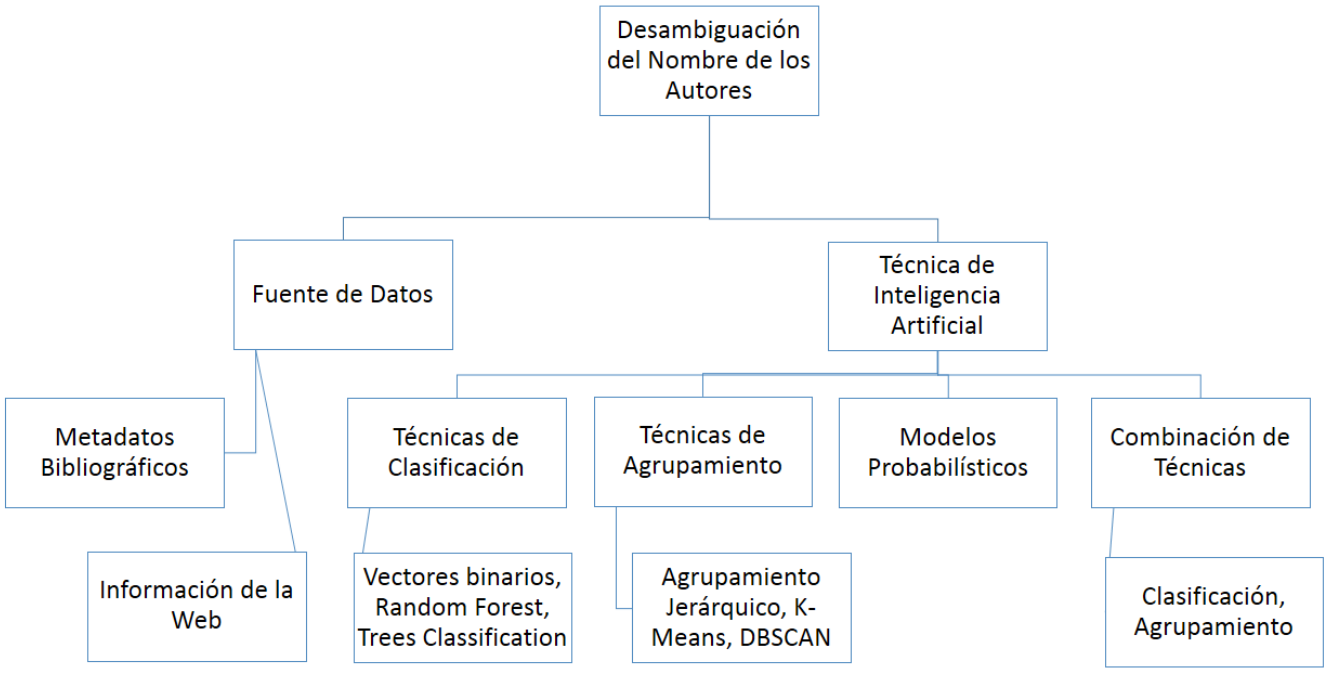

Figura 1. Taxonomía de las técnicas identificadas en las soluciones y la fuente de datos utilizada para las mismas

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 (Lin, 2010) se toma como caso de estudio la red social de investigadores ArnetMiner[1]. Dicha solución está basada en un algoritmo de clasificación, en ella se manejan todos los elementos presentes en dicha red social: coautores, afiliación, citas de los artículos, similitud entre los títulos de los artículos, las páginas web de los autores y la retroalimentación 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 (Treeratpituk, 2009) 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 = {p1, p2, p3,…, pn} suponga que existen m diferentes personas {t1, t2, t3,…, tn} compartiendo un mismo nombre, entonces la tarea consiste en asignar a cada persona la publicación que en realidad escribió. Tomando esta definición como base estamos en presencia de un problema típico de clasificación.

En (Wang, 2012) 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 la clasificación de los mismos. En la aproximación también son calculadas las tasas de error permitidas por el algoritmo utilizado. Con los experimentos realizados sobre el método propuesto se puede demostrar la efectividad de la solución desarrollada.

En (Ferreira, 2012) 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, tanto con la utilización de aprendizaje supervisado como no supervisado. 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 clasifiquen dicho conjunto de datos.

[1] http://arnetminer.org/

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 el 10%.

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 (Zhu, 2011) se usa un mecanismo para identificar páginas web con información referente a los autores. Para la identificación de las páginas web se utiliza un modelo de identificación mediante redes neuronales. Después de identificadas las páginas web, no era posible la extracción de la información de forma directa ya que estas no contienen la información necesaria de forma estructurada. Por tanto, se creó un mecanismo para la extracción de información referente a la afiliación, los coautores y los títulos de trabajos. Por último, se procede a realizar el proceso de agrupamiento. Con el proceso de extracción de información realizado previamente se mejoró dicho proceso de agrupamiento.

En (Yang, 2008) 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 mismo 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 (Kern, 2011) se hace un análisis de las estrategias de desambiguación basada en el método de agrupamiento. Se desarrolla un método para la selección de un modelo que estima el número correcto de autores presentes en un conjunto de datos bibliográficos. Este modelo está basado en la correlación existente entre los coautores. 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 *clusters* están presentes en el conjunto de datos utilizados para la desambiguación.

En (Bernardi, 2011) se reconoce que los datos presentes en las bibliotecas digitales institucionales poseen mayor calidad en la información y mayor grado de organización lo que facilita el proceso de desambiguación. Por otro lado los repositorios de internet no poseen estas características por lo que es necesario enriquecer los mismos. 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[2]. Luego se usa agrupamiento aglomerativo para realizar el proceso de desambiguación del nombre de los autores.

En (Veloso, 2012) 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 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, (3) una extensión del segundo método con la capacidad de auto-entrenarse, reduciendo esto la cantidad de datos de entrenamientos necesitados por la solución propuesta.

La propuesta (Kang, 2011) está basada en la creación de un conjunto de datos de entrenamiento aplicable a las soluciones existentes en la bibliografía. Los pasos seguidos para la creación de dicho conjunto de datos son: (1) determinación de la fuente de datos a utilizar, (2) determinación del conjunto de nombres de autores presentes en los datos seleccionados, (3) generación de las citas de los autores en el conjunto de dato 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 el mismo fue probado en una solución basada en técnicas de agrupamiento, arrojando la misma resultados satisfactorios.

Soluciones usando modelos probabilísticos

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

[2] http://www.wikipedia.org/

En (Torvik, 2005) se presenta un modelo probabilístico 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, que compartan el mismo nombre de autor sean escritos por la misma persona, basado en que los mismos comparten el título, la revista de publicación, coautores en común, entre otros. Esta aproximación marcó un punto de partida para la creación de nuevos métodos de solución relacionados con el tema.

En (Tang, 2008) 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 (J.Pricilla, 2013) 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. Estos títulos son analizados, luego colocados en la misma agrupación aquellas publicaciones que posean mayor probabilidad de referirse al mismo título 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 (Li, 2012) 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 dicho análisis los autores proponen un nuevo método para la solución del problema. Dicho método está 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 significa que hay mayor probabilidad de que dos conjuntos de autores se refieran a la misma 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.

Soluciones usando una combinación de métodos

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

En (Ferreira, 2010) se afirma que los métodos de aprendizaje supervisados 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 coautores. 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 (Gurney, 2012) 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. Además el proceso de comparación entre dos autores es dinámico, es decir, los elementos que se toman en cuenta para comparar un par de autores no son necesariamente los mismos que para comparar otros, varían en dependencia de la disponibilidad de la información. Tienen 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 que no preselecciona elementos previamente para realizar el proceso de desambiguación. Esto ocasiona que el conjunto de datos a comparar sea mayor, pero aumenta 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 (Ferreira, 2012) se propone una herramienta para la evaluación de los métodos propuestos en la literatura sobre la desambiguación del nombre de los autores. Además dichas 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 (Cheng, 2013) 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

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 las revistas científicas** 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 las revistas científicas. 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 las revistas científicas. 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). En este caso los datos obtenidos deben ser tratados una vez que se recuperen con el objetivo de facilitar el trabajo con los mismos.

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 determinado *token* o mecanismo de identificación que permita a los autores registrar sus datos solo en una ocasión en un determinado sitio o base de datos (Beall, 2010), 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: 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 existe 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.

Virtual International Authority File (VIAF): Es un 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 y luego haciéndolos 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 mismo autor. Este proyecto almacena los archivos de autoridad en formato MARC y UNIMARC.

ResearcherID: 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. Esta iniciativa fue creada y es soportada por Thomson Reuters. En la iniciativa cada uno de los autores deben crear una página en la cual se registran los elementos relacionados con los autores, por ejemplo los artículos científicos, los libros de los autores, citaciones 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): El propósito de esta iniciativa es asignar un número único a los autores que aparezcan en publicaciones tanto online 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 distintas, cada una de las ediciones tienen ISBN diferentes, mientras que con la iniciativa el número asignado será siempre el mismo.

Digital Author Identification System (DAI): 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.

Open Researcher & Contributor ID (ORCID): Iniciativa creada con el objetivo de solucionar el problema de la ambigüedad de los nombres 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 un autor determinado. El proceso comienza con el registro de los datos de un autor, luego le es asignado un identificador único el cual es usado como mecanismo de identificación cuando dicho autor firme un artículo o contribución.

Discusión

Las revistas científicas constituyen una de las principales fuentes de consulta por parte de la comunidad científica mundial, lo cual determina que la calidad de los registros bibliográficos que las mismas poseen deben tener la mayor calidad posible. En este artículo 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, dichas soluciones no están expensas a problemas que dificultan la obtención de los mejores resultados.

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 las particularidades de los datos utilizados. De igual forma, no se realiza un previo procesamiento de los mismos. En muchas ocasiones las características de los datos permiten la realización de dicho procesamiento de forma tal que los resultados de las soluciones mejoren considerablemente. Por otro lado se asume que si aparecen dos nombres iguales entonces esos nombres se refieren a la misma persona. Teniendo en cuenta el estudio previo, podemos afirmar que esta es una suposición incorrecta. También se toman como punto de partida para realizar el proceso de desambiguación que los nombres de los autores deben coincidir completamente para entonces comenzar a realizar el proceso antes mencionado, esta condición obtendría resultados erróneos si existiesen errores de escritura en los nombres que se están analizando. También podemos 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.

Ventajas y desventajas de las soluciones estudiadas

Las soluciones basadas en técnicas de agrupamiento permiten la obtención de resultados sin la necesidad de tener un conocimiento previo de la información que será tratada, es decir, no es necesario poseer un conjunto de datos de entrenamiento, como es el caso de los algoritmos de clasificación. También posibilita que el proceso de desambiguación sea automatizado, permite la eliminación de la intervención de la actividad humana en el proceso de desambiguación. Por otro lado este tipo de soluciones tienen limitantes que dificultan su utilización. Los resultados

obtenidos poseen una menor calidad que los resultados 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 conocemos con exactitud el número de agrupaciones o *clusters* que se deben crear en el proceso de desambiguación, introduciendo esto, 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, las técnicas de 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 hacer que sea necesaria la intervención de la actividad humana para determinar las principales características de los datos utilizados.

Las soluciones basadas en la utilización de modelos probabilísticos son unas de las menos complicadas para su utilización. La determinación de los elementos que compondrán el modelo probabilístico y sus respectivas ponderaciones es un trabajo relativamente sencillo. Dichas 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. Conocer las características de los datos utilizados es un elemento importante para la determinación de los principales elementos que contendrá el modelo probabilístico. 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 presente investigación. Con este enfoque se tratan de resolver algunos problemas, entre los que 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, se propicia la aparición de errores y dificultades propias de cada uno de los tipos de soluciones estudiadas anteriormente.

La principal importancia que tiene la desambiguación del nombre de los autores es que, una vez que cada registro en la revista científica se corresponda con uno y solo uno de los investigadores asociados a los artículos, los resultados de las búsquedas de información referentes a dichos autores arrojaran los resultados correctos. Otro aspecto importante radica en la posibilidad de que los estudios bibliométricos realizados sobre los metadatos de las revistas (una vez realizado el proceso de desambiguación), determinarán con exactitud resultados como: qué autor es el más representativo del tema de estudio en cuestión, qué artículos se deben consultar primero en la investigación, entre otros resultado de importancia.

Conclusiones

En el presente trabajo se realizó un análisis crítico de las aproximaciones existentes en la literatura sobre el problema de la ambigüedad del nombre de los autores en las revistas científicas. Las aproximaciones sobre la temática son abordadas utilizando cuatro enfoques, (1) usando técnicas de agrupamiento, (2) usando técnicas de clasificación, (3) usando métodos probabilísticos, (4) usando una combinación de los enfoques vistos anteriormente. Por otro lado, las soluciones estudiadas también pueden dividirse de acuerdo a la naturaleza de los datos, (1) soluciones que utilizan los metadatos de las revistas digitales, (2) soluciones que usan la web como fuente de información. También se puede concluir que las aproximaciones existentes no realizan un previo procesamiento de la información, tampoco tienen en cuenta la calidad de los datos utilizado en las soluciones. Además, ninguna de las aproximaciones estudiadas está orientada a las características del idioma español.

Luego de la revisión realizada los investigadores pretenden desarrollar un algoritmo para solucionar el problema de la ambigüedad del nombre de los autores en las revistas científicas cubanas. Para la realización de dicho algoritmo serán utilizadas técnicas de la minería de datos y elementos asociados al procesamiento del lenguaje natural.

Referencias

– BEALL, J. Metadata for Name Disambiguation and Collocation. *Future Internet* [online]. 2010. Vol. 2, no. 1, p. 1–15. DOI 10.3390/fi2010001. Available from: http://www.mdpi.com/1999-5903/2/1/1

– BERNARDI, R. and LE D. T. Metadata enrichment via topic models for author name disambiguation. En: *Proceedings of the 2009 international conference on Advanced language technologies for digital libraries* [online]. Berlin, Heidelberg: Springer-Verlag. 2011. p. 92–113. ISBN 978-3-642-23159-9. Available from: http://dl.acm.org/citation.cfm?id=2039901.2039908
– CHENG Y., CHEN Z., WANG J., AGRAWAL A. and CHOUDHARY A. Bootstrapping Active Name Disambiguation with Crowdsourcing. En: *CIKM'13*. 2013. DOI 10.1145/2505515.2507858.
– CULOTTA, A., KANANI, P., HALL, R., WICK, M., & MCCALLUM, A. Author disambiguation using error-driven machine learning with a ranking loss function. En: *Sixth International Workshop on Information Integration on the Web (IIWeb-07), Vancouver, Canada,* 2007.
– FERREIRA, A. A., VELOSO, A, GON\CCALVES, M. A. and LAENDER, A H.F. Effective self-training author name disambiguation in scholarly digital libraries. En: *Proceedings of the 10th annual joint conference on Digital libraries* [online]. New York, NY, USA: ACM. 2010. pp. 39–48. Disponible en: http://doi.acm.org/10.1145/1816123.1816130. Gold Coast, Queensland, Australia
– FERREIRA A. A., GON\CCALVES M. A., ALMEIDA J. M., LAENDER A. H. F. and VELOSO A. A tool for generating synthetic authorship records for evaluating author name disambiguation methods. *Inf. Sci.* [online]. 2012. Vol. 206, p. 42–62. DOI 10.1016/j.ins.2012.04.022. Available from: http://dx.doi.org/10.1016/j.ins.2012.04.022
– FERREIRA A. A., MACHADO T. M. and GONÇALVES M. A. Improving Author Name Disambiguation with User Relevance Feedback. *Journal of Information and Data Management* [online]. 2012. Vol. 3, no. 3, p. 332. Available from: http://seer.lcc.ufmg.br/index.php/jidm/article/view/200
– GURNEY T., HORLINGS E. and BESSELAAR P. Author disambiguation using multi-aspect similarity indicators. *Scientometrics* [online]. 2012. Vol. 91, no. 2, p. 435–449. DOI 10.1007/s11192-011-0589-1. Available from: http://link.springer.com/article/10.1007/s11192-011-0589-1
– HERNÁNDEZ J. O, Ramírez M.a J. Q. and Ramírez C. F. *Introducción a la Minería de Datos*. España. PEARSON PRENTICE HALL. Segunda edición. ISBN 84-205-4091-9. KERN R., ZECHNER M. and GRANITZER M. Model Selection Strategies for Author

Disambiguation. En: *Proceedings of the 2011 22nd International Workshop on Database and Expert Systems Applications* [online]. Washington, DC, USA: IEEE Computer Society. 2011. pp. 155–159. Disponible en: http://dx.doi.org/10.1109/DEXA.2011.54.

- J.PRICILLA. An Efficient Framework for Name Disambiguation In Digital Library. *International Journal Of Engineering And Computer Science*. 2013. Vol. 2, no. 4, p. 1097–1105.
- KANG I. S., KIM P., LEE S., JUNG H. and YOU B. J. Construction of a large-scale test set for author disambiguation. *Information Processing & Management* [online]. 2011. Vol. 47, no. 3, p. 452–465. DOI 10.1016/j.ipm.2010.10.001. Available from: http://www.sciencedirect.com/science/article/pii/S0306457310000865
- LAWSON A. E., ALKHOURY S., BENFORD R., CLARK B. R. and FALCONER K. A. What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. En: *Journal of Research in Science Teaching* [online]. 2000. Vol. 37, no. 9, pp. 996–1018. DOI 10.1002/1098-2736(200011)37:9<996::AID-TEA8>3.0.CO;2-J. Disponible en: http://onlinelibrary.wiley.com/doi/10.1002/1098-2736(200011)37:9<996::AID-TEA8>3.0.CO;2-J/abstract.
- LIN Q., WANG B., DU Y., WANG X., LI Y. and CHEN S. Disambiguating Authors by Pairwise Classification. *Tsinghua Science & Technology* [online]. 2010. Vol. 15, no. 6, p. 668–677. DOI 10.1016/S1007-0214(10)70114-0. Available from: http://www.sciencedirect.com/science/article/pii/S1007021410701140
- LI S., CONG G. and MIAO C. Author name disambiguation using a new categorical distribution similarity. En: *Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I* [online]. Berlin, Heidelberg : Springer-Verlag. 2012. p. 569–584. ECML PKDD'12. ISBN 978-3-642-33459-7. Available from: http://dx.doi.org/10.1007/978-3-642-33460-3_42
- M.D, Dr PRAKASH M. N. What Is Metadata? En: *Metadata-driven Software Systems in Biomedicine* [online]. S.l.: Springer London. Health Informatics. 2011. pp. 1–16. ISBN 978-0-

85729-509-5, 978-0-85729-510-1. Disponible en: http://link.springer.com/chapter/10.1007/978-0-85729-510-1_1.

- PÉREZ N. E. M. La bibliografía, bibliometría y las ciencias afines. En: *ACIMED* [online]. 2002. Vol. 10, no. 3, pp. 1–2. Disponible en: http://scielo.sld.cu/scielo.php?pid=S1024-94352002000300001&script=sci_arttext.
- TANG J., ZHANG J, ZHANG D and LI, J. A unified framework for name disambiguation. En: *Proceedings of the 17th international conference on World Wide Web* [online]. New York, NY, USA: ACM. 2008. pp. 1205–1206. Disponible en: http://doi.acm.org/10.1145/1367497.1367728. Beijing, China.
- TORVIK V. I., WEEBER M., SWANSON D. R. and SMALHEISER N. R. A probabilistic similarity metric for Medline records: A model for author name disambiguation: Research Articles. En: *J. Am. Soc. Inf. Sci. Technol.* [online]. January 2005. Vol. 56, no. 2, pp. 140–158. DOI 10.1002/asi.v56:2. Disponible en: http://dx.doi.org/10.1002/asi.v56:2.
- TREERATPITUK P. and GILES, C. L. Disambiguating authors in academic publications using random forests. En: *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* [online]. New York, NY, USA: ACM. 2009. pp. 39–48. Disponible en: http://doi.acm.org/10.1145/1555400.1555408.
- VILARES J. El modelo probabilístico: características y modelos derivados. En: *Revista General de Información y Documentación* [online]. 2009. Vol. 18, pp. 345 – 363. DOI -. Disponible en: http://revistas.ucm.es/index.php/RGID/article/view/RGID0808110345A.
- VELOSO A., FERREIRA A. A., GON\CCALVES M. A., LAENDER A. H. F. and MEIRA, JR., W. Cost-effective on-demand associative author name disambiguation. *Inf. Process. Manage.* [online]. 2012. Vol. 48, no. 4, p. 680–697. DOI 10.1016/j.ipm.2011.08.005. Available from: http://dx.doi.org/10.1016/j.ipm.2011.08.005
- WANG J., BERZINS K., HICKS D., MELKERS J., XIAO, F. and PINHEIRO D. A boosted-trees method for name disambiguation. *Scientometrics* [online]. 2012. Vol. 93, no. 2, p. 391–411. DOI 10.1007/s11192-012-0681-1. Available from: http://dx.doi.org/10.1007/s11192-012-0681-1

– YANG K. H., PENG H. T., JIANG J. Y., LEE H. M. and HO J. M.. Author Name Disambiguation for Citations Using Topic and Web Correlation. En: *Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries* [online]. Berlin, Heidelberg: Springer-Verlag. 2008. pp. 185–196. Disponible en: http://dx.doi.org/10.1007/978-3-540-87599-4_19.
– ZHU J., FUNG G. and WANG L. Efficient Name Disambiguation in Digital Libraries. En: *Web-Age Information Management* [online]. 2011. Springer Berlin Heidelberg. p. 430–441. Lecture Notes in Computer Science, 6897. ISBN 978-3-642-23534-4, 978-3-642-23535-1. Available from: http://link.springer.com/chapter/10.1007/978-3-642-23535-1_37