Algoritmos para solventar la falta de normalización de nombres de autor en los estudios bibliométricos

Rodrigo Costas
María Bordons *

Artículo recibido:
4 de noviembre de 2005.
Artículo aceptado:
25 de septiembre de 2006.

Resumen:

Se presentan dos algoritmos para detectar y solventar problemas de normalización de nombres de autores en datos procedentes de la base de datos *Science Citation Index* de Thomson ISI. El primer algoritmo permite detectar firmas diferentes que, por su parecido, podrían pertenecer a una misma persona. El segundo ayuda a determinar si dos firmas parecidas se corresponden o no con una misma persona en función del grado de similaridad existente entre los documentos de una y otra variante de firma. Para determinar la eficacia de los algoritmos se han utilizado como control los datos de autores normalizados de un estudio anterior. El algoritmo detecta un 67% de las variantes

* Ambos autores pertenecen al Centro de Información y Documentación Científica (CINDOC), CSIC., Madrid, España.
(rodrigo.costas@cindoc.csic.es); (mbordons@cindoc.csic.es).

de firma existentes en la población objeto de estudio y tiene un 74% de acierto en la determinación de si esas firmas corresponden a una misma persona.

Palabras clave: Normalización de nombres de autores; Bases de datos; *Science Citation Index*; Thomson ISI; Algoritmos; Variantes de firma.

Abstract

Algorithms to solve the lack of normalization in author names in bibliometric studies

Rodrigo Costas and María Bordons

Two algorithms to detect and solve normalization problems of author names in data originated in Thomson's ISI *Science Citation Index* are presented. The first algorithm allows detection of different names which could belong to the same person. The second one, based on the degree of similarity between two variants of the same name on a document, helps to determine whether two similar names correspond or not to the same person. In order to determine the efficacy of the algorithms, a control of normalized author data from a previous study has been used. The First algorithm detects 67% of name variants existing in the population under study, and the second one was successful in 74% of the cases.

Keywords: Author name normalization; *Science Citation Index*; Thomson ISI; Algorithms; Name variations.

1. Introducción

Tradicionalmente las bases de datos bibliográficas se han utilizado para almacenar y recuperar información y de este modo contribuir al proceso de transmisión del conocimiento científico. No obstante, estas bases también representan una gran fuente de datos para los estudios bibliométricos, ya que en ellas se encuentra representada la producción científica de países, regiones y áreas científicas. Estas fuentes permiten generalmente descargar los datos que se han consultado para que puedan ser tratados y analizados posteriormente con otras herramientas informáticas.

La importancia que las bases de datos tienen para los estudios bibliométricos se pone de manifiesto en la definición de bibliometría propuesta por Katz y Hicks (1997), quienes la consideran: el arte de explorar las bases de datos en la búsqueda de indicadores que reflejen la actividad investigadora, así como las interacciones entre individuos, grupos, instituciones, sectores, etcétera.

Entre los datos procedentes de las bases de datos bibliográficas existen tres que tienen una relevancia capital para los estudios bibliométricos: los autores, que permiten estudiar la productividad de los investigadores; la afiliación institucional, importante para aprender sobre la actividad científica de las instituciones; y los datos de carácter temático, que permiten analizar la actividad científica por disciplinas. Sin embargo, estos campos no siempre presentan una correcta normalización, lo que dificulta la realización de cálculos automáticos (Spinak, 1995) y constituye un importante inconveniente para su explotación bibliométrica (Lardy y Herzhaft, 1992; Frías y Romero Gómez, 1998).

El campo autor es uno de los campos más sensibles a la falta de normalización ya que las variaciones sobre la forma en que figura un autor en sus diferentes publicaciones pueden estorbar el cálculo de su producción real, al dificultar el desarrollo de estudios bibliométricos a nivel micro.

Junto a las recomendaciones a los autores sobre la importancia de firmar las publicaciones de una forma normalizada y estable a lo largo del tiempo, surgen también hoy indicaciones dirigidas a las revistas y a las bases de datos (ver por ej. Ruíz-Pérez *et al* 2002; Fernández y García, 2003), principalmente a las internacionales, las cuales tienen que enfrentar el problema de las diferencias que existen entre los diferentes países al estructurar los nombres personales. Así, la estructura de nombre personal predominante en las bases de datos internacionales es la formada por una o dos iniciales de nombres, seguida de un solo apellido (por ej. J.H.Smith), pero con frecuencia son mal recogidos los nombres hispanos si los autores incluyen dos apellidos (por ejemplo, J. García Sánchez, puede ser recogido como J.G.Sánchez). Es indudable el interés de propuestas a priori, orientadas a dar recomendaciones a autores, revistas y bases de datos con la intención de lograr una mayor normalización de los nombres de autores en las publicaciones y bases de datos. Sin embargo, también se han planteado soluciones a posteriori, una vez introducidos los datos en la base de datos (Costas-Comesaña y García-Zorita, 2003; Torvik *et al*, 2005). En este ámbito destacan los algoritmos informáticos para comparar nombres personales *Personal Name Matching* o *Name Matching* (Camps Paré, 2003; Patman y Thompson, 2003; Thompson y Dozier, 2003; Patman y Thompson, 2005), los cuales permiten comparar dos

cadenas de nombres y determinar la probabilidad de que ambas designen a la misma persona. Normalmente estos algoritmos utilizan información adicional aparte del nombre, al emplear otros elementos tales como variantes de deletreos, información fonética, la distribución de las teclas del ordenador, etcétera. Sin embargo, debido a la complejidad propia de esta tarea, muchas veces es necesaria la intervención humana para determinar si las cadenas de nombres similares detectadas corresponden o no a la misma persona. El software *Synoname* (Gross, 1991; Borgman y Siegfried, 1992) desarrollado por el Consorcio Getty, detecta nombres parecidos, candidatos a pertenecer al mismo investigador, pero que no se aceptan como correctos hasta que hay una autorización humana.

Actualmente no existen aplicaciones informáticas que empleen estos algoritmos y que sean flexibles para ser utilizados como una herramienta más en la investigación bibliométrica. En este trabajo se pretende profundizar en esta problemática y proponer soluciones que permitan facilitar la normalización de los datos procedentes de algunas de las bases de datos que tienen mayor interés bibliométrico, como son las de Thomson ISI (SCI, SSCI y A&HCI).

2. Objetivos

1. Desarrollar y presentar algoritmos metodológicos para detectar variantes de firma de los investigadores en los registros bibliográficos procedentes de las bases de datos de Thomson ISI.
2. Cuantificar el grado de similaridad entre los documentos asignados a cada variante de firma, con el fin de determinar si dichas firmas pertenecen efectivamente a una sola persona o a más de una.
3. Analizar la efectividad del funcionamiento tanto de la metodología de detección de variantes de firmas como de la cuantificación de la similaridad entre ellas.

3. Metodología

3.1. Algoritmo de detección de variantes de firma similares, procedentes de las Bases de Datos de Thomson ISI

El algoritmo propuesto busca detectar firmas de autores "parecidas", partiendo de la estructura general de los nombres hispánicos que constan de dos apellidos y uno o dos nombres propios:

APE1 APE2, NOM1 [NOM2]→GARCÍA RUIZ, JOSE MANUEL

Teniendo en cuenta las prácticas de indización de nombres seguidas por Thomson ISI (Ruíz-Pérez *et al*, 2002), la parte final del nombre presente en el documento es tomada como apellido y las restantes cadenas son tomadas como nombres, recogiéndose como iniciales, estableciéndose así que del ejemplo anterior se derivan 9 variantes potenciales de firma "lógicas", que son las siguientes:

1. GARCÍA J
2. GARCÍARUIZ J
3. GARCÍA JM
4. GARCÍARUIZ JM
5. RUIZ JG
6. RUIZ JMG
7. GARCÍA M
8. GARCÍARUIZ M
9. RUIZ MG

El algoritmo que se presenta compara por parejas las variantes del ejemplo anterior y establece que tienen alguna posibilidad de corresponder a la misma persona. Hay que tener en cuenta que existen algunas combinaciones de variantes de firma que por sí mismas no se pueden asociar (p. ej. "GARCÍA J // RUIZ JG" o "GARCÍA JM // RUIZ JMG") dado que no tienen suficientes vínculos textuales en común.

El algoritmo funciona comparando una firma (A1) con otra firma (A2). Incluye 13 sentencias que se ejecutan sucesivamente una detrás de otra; si el algoritmo pasa por las 13 sin encontrar ninguna coincidencia se considerará que las firmas comparadas no son "parecidas", mientras que si en algún caso se cumplen las condiciones señaladas, las dos firmas se considerarán "sospechosas" de pertenecer a una misma persona.

Sentencia 1:
Resuelve las siguientes combinaciones:

A1: GARCÍA J
A2: GARCÍARUIZ J

A1. GARCÍA JM
A2. GARCÍARUIZ JM

A1. GARCÍA M
A2. GARCÍARUIZ M

Se identifican casos en que las iniciales de A1 y A2 son iguales y coinciden las cuatro primeras letras de los dos apellidos (la selección de los cuatro caracteres iniciales de los apellidos es decidida por el usuario, y puede ser aumentada o reducida).

Sentencia 2:

A1. GARCÍA J
A2. GARCÍA JM

A1. GARCÍARUIZ J
A2. GARCÍARUIZ JM

Identifica aquellos casos en que los apellidos coinciden, A1 tiene una inicial, A2 tiene dos iniciales, y los dos coinciden en la primera inicial.

Sentencia 3:

A1. GARCÍA J
A2. GARCÍARUIZ JM

A1. GARCÍA JM
A2. GARCÍARUIZ J

Identifica aquellos casos donde el número de iniciales de las firmas es de uno y dos respectivamente, y que coinciden en la primera inicial del nombre y en los cuatro primeros caracteres del apellido.

Sentencia 4:

A1. RUIZ JG
A2. GARCÍARUIZ J

A1. RUIZ JMG
A2. GARCÍARUIZ J

A1. RUIZ MG
A2. GARCÍARUIZ M

Identifica aquellos casos en los que el apellido de A1 está contenido en A2, A1 tiene dos o tres iniciales, A2 tiene una inicial, A1 y A2 coinciden en la primera inicial del nombre, y la inicial final del A1 es igual que la primera letra del apellido de A2.

Sentencia 5:

A1. GARCÍA M
A2. GARCÍA JM

A1. GARCÍARUIZ M
A2. GARCÍARUIZ JM

Identifica como "pareja parecida" aquellos casos en los que los apellidos coinciden, A1 tiene una inicial, A2 tiene dos iniciales y la primera inicial de A1 coincide con la inicial final de A2.

Sentencia 6:

A1. GARCÍA JM
A2. GARCÍARUIZ M

Esta sentencia identifica casos en los que coinciden los cuatro primeros caracteres de los apellidos, el número de iniciales de A1 son dos y el de A2 es uno, y la inicial final de A1 es igual a la inicial de A2.

Sentencia 7:

A1. RUIZ JG
A2. GARCÍARUIZ JM

Identifican aquellos casos donde el apellido de A1 está contenido en A2, las dos firmas tienen dos iniciales, coinciden en la primera inicial, y la inicial final de A1 es igual a la primera letra del apellido de A2.

Sentencia 8:

A1. RUIZ JMG
A2. GARCÍARUIZ JM

Identifica los casos en que el apellido de A1 está contenido en A2, A1 tiene tres iniciales y A2 dos iniciales, coinciden en la primera inicial, y la inicial final de A1 es igual a la primera letra del apellido de A2.

Sentencia 9:

A1. GARCÍA M
A2. GARCÍARUIZ JM

Detecta los casos donde los cuatro primeros caracteres de los apellidos coinciden, A1 tiene una inicial y A2 tiene dos iniciales, y la primera inicial de A1 es igual que la inicial final de A2.

Sentencia 10:

A1. RUIZ MG
A2. GARCÍARUIZ JM

Identifica aquellas combinaciones en las que el apellido de A1 está contenido en A2, las dos firmas tienen dos iniciales, la inicial final de A1 es igual a la primera letra del apellido del A2, y la primera inicial de A1 es igual a la inicial final de A2.

Sentencia 11:

A1. RUIZ JG
A2. RUIZ JMG

Identifica los casos en los que coinciden los apellidos, el número de iniciales de A1 es dos y el de A2 es tres, coinciden en la primera inicial, y la inicial final de A1 es igual a la inicial final de A2.

Sentencia 12:

A1. RUIZ JG
A2. RUIZ MG

Identifica los casos en los que coinciden los apellidos de las firmas y el número de iniciales es dos en ambos casos, y en los que coinciden las iniciales finales.

Sentencia 13:

A1. RUIZ MG
A2. RUIZ JMG

Detecta los casos en los que coinciden los apellidos de las firmas y el número de iniciales de A1 es dos y el de A2 es tres, coinciden las iniciales finales, y la primera inicial de A1 es igual a la segunda inicial de A2.

Si una vez ejecutadas las 13 sentencias no se han detectado "parejas parecidas", se considerará que las firmas que se están comparando no son textualmente susceptibles de pertenecer a una misma persona.

El algoritmo contempla la mayor parte de los casos de variantes de firma que un autor puede presentar. Alguno de los casos más interesantes que detecta son aquellos en los que el segundo apellido del autor es el que indiza, mientras que el primero se incluye como inicial (GARCÍARUIZ J – RUIZ JG o GARCÍARUIZ JM – RUIZ JMG), dado que estas combinaciones son difíciles de detectar incluso en revisiones manuales.

Debe tenerse en cuenta que pueden darse casos en los que dos autores tengan firmas similares, y por tanto sean detectados como susceptibles de ser una misma persona, pero que en realidad no lo sean. Es aquí donde se hace patente la necesidad de contar con algún mecanismo de análisis de los documentos firmados bajo cada variante para determinar si éstas pertenecen o no a la misma persona.

3.2. Algoritmo para cuantificar la similaridad entre variantes de firmas

No basta con detectar firmas susceptibles de corresponder a una misma persona, también es necesario cuantificar este parecido y, en función de su mayor o menor similaridad, aceptar o rechazar si una pareja de firmas pertenece a una misma persona. Para determinar el grado de similitud entre firmas, se ha partido, al igual que Torvik *et al* (2005), de la hipótesis de que los documentos firmados por un determinado autor, con frecuencia presentan características comunes (coautores, revistas, palabras clave, lugares de trabajo, referencias, etcétera).

De este modo, dado un par de firmas que pueden corresponder a una misma persona, se analizan los coautores, los lugares de trabajo y las revistas de publicación de sus documentos, y se calcula el grado de coincidencia que hay entre los documentos de las dos firmas.

Para el cálculo de la similaridad o parecido entre los documentos de cada variante de firma se ha realizado una adaptación de la medida del coseno,

utilizada en recuperación de la información (Harman, 1992; Lee *et al*, 1997). La adaptación consiste en considerar a cada autor como un vector de elementos (de coautores, de revistas o de centros de trabajo), donde cada elemento está o es ponderado por el número de documentos en los que aparece.

Así tenemos que la adaptación de la medida del coseno quedaría del siguiente modo:

$$VS = \frac{\sum (FA1_i) * (FA2_i)}{\sqrt{\left(\sum (FA1_i^2) * \sum (FA2_i^2)\right)}}$$

Dónde:

FA1= es el número de veces que el elemento "i" aparece en los documentos de A1.

FA2= es el número de veces que el elemento "i" aparece en los documentos de A2.

Por ejemplo, si se comparan los coautores de los documentos de las firmas "Casas V" y "Casas VJ" (firmas parecidas que podrían pertenecer a una misma persona), se obtienen los vectores de la Tabla 1:

Tabla 1. Ejemplo de análisis de Coautores de "Casas V" y "Casas VJ"

	Coautores				
Autores	Pérez J	García P	Vívez O	Lamino T	Milchord SA
A1: "Casas V"	2 docs.	3 docs.	6 docs.	2 docs.	0 docs.
A2: "Casas VJ"	0 docs.	1 docs.	2 docs.	4 docs.	4 docs.

De este modo se calcularía la similaridad por coautores de la siguiente manera:

$$\frac{((2*0)+(3*1)+(6*2)+(2*4)+(0*4))}{\sqrt{((2^2+3^2+6^2+2^2+0^2)*(0^2+1^2+2^2+4^2+4^2))}} = 23/44{,}28 = 0{,}52$$

En el algoritmo final propuesto, se obtienen 3 valores de similaridad para cada pareja de firmas comparadas: uno por los coautores, otro por los centros de trabajo y otro por las revistas de publicación, y se obtiene un valor de similaridad final (VS) consistente en la media de estos tres valores que oscilará entre 0 y 1:

$$VS = \frac{\text{Sim (Coautores)} + \text{Sim (Centros trabajo)} + \text{Sim (Revistas)}}{3}$$

El proceso presenta la posibilidad de trabajar iterativamente. Esto supone que cuando una pareja de firmas presenta un VS muy alto, la información de la nueva firma se le asigna automáticamente a su autor, y es utilizada a su vez para compararse con el resto de firmas pendientes de revisión, lo cual le da mayor fiabilidad a la comparación. Sin embargo, esta característica debe utilizarse con precaución dado que una mala asignación automática podría provocar que firmas de personas diferentes se asimilaran como propias de una sola persona.

3.3. Metodología para la evaluación de los algoritmos propuestos

Para comprobar la efectividad de los algoritmos propuestos, éstos se han aplicado a datos previamente analizados en un estudio anterior (Costas Comesaña, 2003), en el cual se estudió la producción científica ISI (versión CD-ROM) durante el periodo 1994-2001 de 333 investigadores del Área de Recursos Naturales del Consejo Superior de Investigaciones Científicas (CSIC), principal organismo dedicado a la investigación en España. Se cuenta, pues, con la relación de investigadores del área y su lugar de trabajo. La producción final de dichos investigadores ascendió a 3.302 documentos.

En este trabajo se desea comprobar que los algoritmos propuestos detectan las variantes de firma identificadas en el estudio anterior, y que los datos de la normalización de dicho estudio sirven como control de la eficacia de los algoritmos. Debe tenerse en cuenta también que la validez de los datos del estudio anterior está refrendada por los expertos del área de recursos naturales que lo supervisaron.

3.3.1. Descripción de los datos de control

Para el presente análisis se ha contado con la información correspondiente a las variantes de firma de los investigadores, identificadas en el estudio anterior, que se obtuvieron por un procedimiento semi-automático complementado con búsquedas manuales y revisión por expertos. La Figura 4 incluye una muestra de la tabla de autores con sus variantes de firma. Así por ejemplo, se observa que el autor "ÁLVAREZ COBELAS, MIGUEL" aparece firmando sus documentos como "Cobelas MA" y como "Álvarezcobelas M".

Autor	Firma
ALDASORO MARTÍN,JUAN JOSÉ	Aldasoro JJ
ALONSO LÓPEZ, JUAN CARLOS	Alonso JC
ALONSO MARTÍNEZ, MARÍA BELEN	Alonso B
ÁLVAREZ COBELAS, MIGUEL	Álvarezcobelas M
ÁLVAREZ COBELAS, MIGUEL	Cobelas MA
ÁLVAREZ SALGADO, XOSÉ ANTONIO	Álvarezsalgado XA
ÁLVAREZ SALGADO, XOSÉ ANTONIO	Álvarezsalgado X

Fig. 1. Muestra de la tabla de autores con sus variantes de firma reales, procedentes del estudio anterior

Hay que señalar que el 82% de los autores firmaban siempre de la misma manera, frente a un 18% de autores que firmaban con dos o más variantes (tabla 2).

Tabla 2. Distribución de autores según el número de firmas con las que aparecen en los documentos (Tabla de control)

Nº de firmas	Nº de autores en ISI	%
1	251	81,8
2	42	13,7
3	14	4,6
Total	307	

Nota: 307 investigadores con producción ISI.

A partir de la tabla que se muestra en la figura 4 se generó una "Tabla control", cuya estructura se muestra en la figura 5, que incluía todas las combinaciones de parejas de firmas reales de cada investigador; es decir, que sólo recoge aquellos autores para los que se identificaron dos o más variantes de firma, y que incluyeron un total de 86 entradas distintas. Esta tabla de control se utilizará para analizar la efectividad del algoritmo que detecta variantes de firmas similares.

En la Figura 2 se observan las variantes de firmas con las que los investigadores han firmado sus documentos. En este ejemplo todos los autores tienen dos variantes, salvo "VALERO GARCÉS, BLAS LORENZO" que tiene tres ("Valerogarcés BL", "Valerogarcés B" y "Garcés BLV"). Para evitar duplicados innecesarios las parejas se crean siempre de modo que la firma con más caracteres está en "FIRMA1" y la más corta en "FIRMA2".

Nombre	Firma1	Firma2
BRANDLE MATESANZ, JOSÉ LUIS	Brandle JL	Brandle J
MARTÍNEZ FRÍAS, JESÚS	Martínezfrías J	Martínez J
GARCÍA DEL CURA, MARÍA DE LOS ANGELES	Delcura MAG	Delcura G
ÁLVAREZ COBELAS, MIGUEL	Álvarezcobelas M	Cobelas MA
MALDONADO BARAHONA, MANUEL	Maldonaldo M	Maldonado M
RODRÍGUEZ BADIOLA, EDUARDO	Rodríguezbadiola E	Badiola ER
VALERO GARCÉS, BLAS LORENZO	Valerogarcés B	Garcés BLV
VALERO GARCÉS, BLAS LORENZO	Valerogarcés BL	Garcés BLV
VALERO GARCÉS, BLAS LORENZO	Valerogarcés BL	Valerogarcés B

Fig. 2. Ejemplo Tabla de control de combinaciones de firmas reales de los investigadores

3.3.2. Fiabilidad del algoritmo de similaridad

Para comprobar la fiabilidad del algoritmo de similaridad entre variantes de firma se ha seleccionado para cada autor una variante de referencia, que es aquélla ligada al lugar de trabajo correcto del autor.

Siguiendo las metodologías propuestas por Fernández *et al* (1993), Bordons *et al* (1995) y Zulueta et al (1999), se generó una tabla denominada *Autor-Centro* (véase ejemplo Figura 3), en la cual cada firma de autor le es asignada a un centro de trabajo normalizado. Este proceso se basa en asignarle a todos los firmantes de un documento con un solo lugar de trabajo dicha dirección, y a continuación, identificar aquellos documentos en los que haya quedado un solo autor y una sola dirección sin asignar, que se añaden a la Tabla Autor-Centro.

Auth	Provincia	Instit	Organismo	Centro normalizado
Abad A	08	H	HGTP	H.Germans Trias.Pujol,Badalona
Abad E	08	2	020304	I.Inv.Quim.Amb.CSIC,Barcelona
Abad JL	08	2	020304	I.Inv.Quim.Amb.CSIC,Barcelona
Abad JP	28	21	050105	C.Biol.Mol.CSIC-UAM,Madrid
Abad M	46	1P	2AG	ETSI.Agron.UPV
Abadía A	50	2	090101	E.Exptl.Aula Dei CSIC,Zaragoza
Abadía J	50	2	090101	E.Exptl.Aula Dei CSIC,Zaragoza
Abadortega MD	18	21	030261	I.A.Cienc.Tierr.CSIC-U.Granada
Garcíaabadgarcía MT	28	2	------	CSIC (varios),Madrid
Martínezabad M	46	H	HINSA2	H.Dr.Peset,Valencia
Sánchezabadía S	28	2	------	CSIC (varios),Madrid

Fig. 3. Ejemplo Tabla Autor-Centro

La Figura 1 muestra la tabla con la combinación de firmas originales (AUTH) y los centros normalizados desglosados en 3 elementos diferentes: Provincia, Institución y Organismo. Cada entrada de la Tabla Autor-Centro está ligada con todos los documentos en los que aparece la firma de AUTH y el centro normalizado. Hay que señalar que la mayor parte de las firmas de autores quedan asociadas a uno o varios centros, aunque pueden quedar algunos que no están asignados a ningún centro (véase ejemplo Figura 4).

Auth	Provincia	Instit	Organismo	Centro normalizado
Abad I	--	----	------	Varios sin identificar
Abad JM	--	----	------	Varios sin identificar
Abad LR	--	----	------	Varios sin identificar
Sabadini R	--	----	------	Varios sin identificar

Fig. 4. Ejemplo de firmas que no se han podido asociar con ningún centro

En la Tabla Autor-Centro existen entradas que hemos denominado "Ciertas", en las que el centro normalizado coincide con el centro de trabajo real del investigador asignado, y que se revisaron cuidadosamente para garantizar que los documentos de esas entradas pertenecen a los investigadores. Asimismo, las entradas en las que esta correspondencia del centro de trabajo no existe fueron marcadas como "Dudosas", y son las que serán comparadas a través del algoritmo con las entradas "Ciertas" para determinar si pertenecen o no al investigador al que han sido asignadas (véase Figura 5).

Notas	Investigador asociado al autor	Autor	Provincia	Instit	Organismo	Centro Autor
CIERTO	AGUILAR-AMAT FERNÁNDEZ,JUAN	Amat JA	41	2	060401	060401
DUDOSO	AGUILAR-AMAT FERNÁNDEZ,JUAN	Amat JA	28	2	060501	060401
CIERTO	ALCARAZ MEDRANO,MIGUEL ÁNGEL	Alcaraz M	08	2	070103	070103
DUDOSO	ALCARAZ MEDRANO,MIGUEL ÁNGEL	Alcaraz M	--	----	------	070103
CIERTO	ALDASORO MARTÍN,JUAN JOSÉ	Aldasoro JJ	28	2	060102	060102
DUDOSO	ALDASORO MARTÍN,JUAN JOSÉ	Aldasoro JJ	--	----	------	060102
CIERTO	ALONSO LÓPEZ,JUAN CARLOS	Alonso JC	28	2	060501	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	Alonso J	28	2	050204	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	López JA	--	----	------	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	Alonso C	41	2	060401	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	Alonso JC	28	2	050402	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	Alonso C	28	3OPI	INIA	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	Alonso J	--	----	------	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	López JA	28	3OPI	CARL	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	Alonso C	18	2	060301	060501
DUDOSO	ALONSO LÓPEZ,JUAN CARLOS	Alonso C	28	21	050105	060501

CIERTO	ALONSO MARTÍNEZ,MARÍA BELEN	Alonso B	08	2	070103	070103
DUDOSO	ALONSO MARTÍNEZ,MARÍA BELEN	Alonso M	28	2	050204	070103
DUDOSO	ALONSO MARTÍNEZ,MARÍA BELEN	Martínez MA	28	21	050105	070103
DUDOSO	ALONSO MARTÍNEZ,MARÍA BELEN	Martínez MA	--	9	------	070103

Fig. 5. Muestra de la tabla Autor-Centro

Se puede observar en la Figura 5 que el autor "AGUILAR-AMAT FERNÁNDEZ, JUAN", tiene una entrada considerada "Cierta", dado que coincide con su centro de trabajo real (columna CENTRO AUTOR). Sin embargo, a continuación el mismo autor presenta una entrada "Dudosa", en la que presenta un centro de trabajo diferente. Los documentos de una y otra entrada serán comparados por el algoritmo, y se obtendrá un valor de similaridad a partir del cual se decidirá si aceptar o rechazar que la segunda entrada pertenece al mismo investigador.

4. Resultados

4.1. Resultados del análisis del algoritmo de detección de variantes de firma similares

Para este análisis se utilizó como comprobación de las parejas de firmas que se obtienen con el algoritmo, la tabla de control mostrada en la Figura 2, dado que si el algoritmo funciona apropiadamente debería ser capaz de detectar la mayor parte de las firmas identificadas en el estudio anterior.

Se han obtenido 1.176 firmas de autores sobre las cuales se ha ejecutado el algoritmo de identificación de variantes y a partir de las cuales se han obtenido 220 parejas de firmas (véase ejemplo Figura 6).

AUT1	AUT2
Brandle JL	Brandle J
Bustillo MA	Bustillo M
Bustos AR	Bustos A
Castillo VM	Castillo M
Castillo VM	Castillo V
Danobeitia JJ	Danobeitia J
Delgadohuertas A	Delgado A
Delgadohuertas A	Huertas AD
Diazpaniagua C	Díaz C
Duarte CM	Duarte C
Duarte MC	Duarte C

Fig. 6. Combinaciones de firmas obtenidas a través del algoritmo

Se analizó si las 86 firmas de control se encuentran entre las 220 combinaciones obtenidas a través del algoritmo, y resultó que 58 cadenas de la Tabla de control coinciden con alguna de las combinaciones obtenidas a través del algoritmo, de modo que el 67% de las parejas de firmas de control son detectadas por el algoritmo. El 33% restante se corresponde bien con errores tipográficos, o bien con combinaciones de firmas que no son detectables por el algoritmo (véase ejemplo Figura 7).

Nombre	Firma1	Firma2
BRAZA LLORET, FRANCISCO	Braza F	Braza P
CARRILLO ESTEVEZ, MANUEL	Carrillo M	Carillo M
CASIMIRO-SORIGUER ESCOFET, RAMÓN	Soriguer R	Soriguer C
DELGADO HUERTAS, ANTONIO LUIS	Huertas AD	Delgado A
DÍAZ CUSI, JORDI	Cusi JD	Díaz J
DOBLAS LAVIGNE, MIGUEL MANUEL DE LAS	Doblas M	Doblas M

Fig. 7. Ejemplo de parejas de variantes de firmas no detectadas por el algoritmo

4.2. Resultados del análisis del algoritmo de cuantificación de la similaridad entre variantes de firma

Se obtuvo la tabla Autor-centro, en la que un total de 141 autores presentan una entrada "Cierta" y, como mínimo, una entrada "Dudosa" que hay que verificar. En total hay 748 entradas en la Tabla Autor-Centro, de las cuales 153 son "Ciertas" (un autor puede tener más de una entrada "cierta") y 595 "Dudosas".

Las variantes de firma que aparecen ligadas al centro de trabajo real de un autor son las más fáciles de detectar y validar. El mayor problema se refiere a identificar aquellas variantes asignadas a distintos centros, que en ocasiones corresponden a investigadores diferentes, pero que también pueden corresponder a un mismo investigador que ha cambiado su lugar de trabajo.

Se ejecutó el algoritmo, y se compararon sus entradas "ciertas" con las "dudosas" para cada autor, y se obtuvieron los VS de dichas comparaciones, para posteriormente decidir qué entradas "Dudosas" pertenecen efectivamente a los investigadores del estudio.

En la Tabla 3 se puede observar que 461 entradas autor-centro (62% del total) obtuvieron un VS=0, lo que sugiere que estas entradas no pertenecen al investigador asignado. Gracias a los resultados del estudio anterior se observa que el 97% de las entradas con un VS=0 efectivamente no pertenecían al investigador analizado.

Por otra parte existen 134 entradas con un VS>0 (Tabla 3). De ellas, 99 (74%) pertenecían efectivamente a los investigadores en estudio, mientras

que 35 (26%) no pertenecían a éstos. Sin embargo se observa que todas las entradas con un VS>=20 corresponden efectivamente a variantes de un mismo investigador.

Tabla 3. Análisis del grado de acierto del Valor de Similaridad (VS).

	TOTAL	%	% TOT CAD (748)
Entradas con VS=0	461	100	62
Variantes aceptadas	12	3	2
Variantes rechazadas	449	97	60
Entradas con VS > 0	134	100	18
Variantes aceptadas	99	74	13
Variantes rechazadas	35	26	5

5. Conclusiones

La normalización de los datos de las bases de datos bibliográficas es esencial para mejorar su calidad y optimizar su uso en la recuperación de información, y especialmente en la realización de estudios bibliométricos. Sin embargo, la mayor parte de las bases de datos presentan todavía diversos problemas de normalización, y uno de los más importantes es el relativo al campo autor, que obliga a desarrollar metodologías de trabajo específicas para superar estas limitaciones.

Este trabajo se ha propuesto una metodología que permita identificar variantes de nombre y normalizar dichas variantes con apoyo en la información del campo lugar de trabajo. Tal metodología presenta dos ventajas:

- permite detectar automáticamente posibles variantes de firma de una misma persona, que serían difíciles de identificar en una revisión manual;
- permite cuantificar la similaridad de la producción de dos variantes de firma, a partir de lo cual sería posible realizar la normalización automática de las entradas con un alto valor de similaridad.

Como se ha observado, el algoritmo de identificación de variantes de firma presenta una eficacia notable. Detecta con acierto el 67% de las variantes de firma similares. La mayor parte de las firmas no detectadas corresponde a errores tipográficos o a variantes que no tienen elementos textuales suficientes para ser identificadas automáticamente. En cuanto al análisis de similaridad,

se establece que un VS>=20 es un umbral adecuado para afirmar que dos firmas pertenecen efectivamente a una misma persona, mientras que un VS=0 se corresponde en el 98% de los casos con firmas de autores diferentes.

En cuanto a las limitaciones de los algoritmos se pueden señalar las siguientes:

- no se puede automatizar la normalización de nombres de autores cuyo lugar de trabajo no es conocido, ya que bajo un mismo nombre se podría mezclar la producción de más de un autor;
- si se acepta como "Cierta" automáticamente una entrada de autor-centro incorrecta, todo el proceso puede verse contaminado por esa entrada, por ello es necesario elegir un umbral alto de similaridad para la automatización iterativa (VS>=30).
- la no efectividad del algoritmo en el 100% de los casos hace necesaria una revisión manual en un pequeño número de casos con bajo VS, para lo cual se sugiere consultar el *currículum vitae* de los investigadores, obtener información en Internet o consultar a los propios autores.

Finalmente hay que señalar que el algoritmo de similaridad sería susceptible de algunas mejoras potenciales por medio de la inclusión de nuevos elementos que ayuden a medir el parecido entre documentos (palabras clave, palabras del resumen, materias ISI, referencias, etcétera). Asimismo, también sería factible realizar una ponderación de los diferentes elementos incluidos en el cálculo del valor de similaridad, ya que teniendo en cuenta lo afirmado por Torvik *et al* (2005), y considerando los tres elementos incluidos en el presente análisis (coautores, centros y revistas), se puede afirmar que el número de coautores en común tiene conceptualmente más importancia que los otros dos elementos, lo que hace posible darle un mayor peso a la coincidencia de coautores que a la coincidencia de revistas o centros de trabajo.

En definitiva, los algoritmos propuestos pueden ser de gran utilidad para normalizar los nombres de los autores incluidos en las bases de datos bibliográficas, y serían de gran interés para realizar estudios bibliométricos. La metodología también podría extenderse a otros campos como es el control de autoridades en bases de datos o la normalización de catálogos. Los algoritmos aquí presentados y otros descritos en la literatura son útiles para enfrentarnos al problema de la falta de normalización de nombres vigente hoy en las bases de datos bibliográficas, pero simultáneamente es importante establecer procedimientos que incrementen la normalización de los nombres de autores en los distintos medios que éstos utilizan para difundir sus avances científicos.

6. Bibliografía

Bordons, M.; Zulueta, M.A.; Cabrero, A.; Barrigón, S. (1995). "Identifying research teams with bibliometric tools", en *Proceedings of the fifth Biennial conference of the International Society for Scientometrics and Informetrics.* London: Learned Information, p. 83-92, 1995.

Borgman, C.L.; Siegfried, S.L. (1992). "Getty's Synoname and its cousins: a survey of applications of Personal Name-Matching Algorithms", en *Journal of the American Society for Information Science,* 43 (7), 459-476, 1992.

Camps Paré, R. (2003). *Búsqueda aproximada de antropónimos en las bases de datos de los Sistemas de Información, en presencia de errores.* [Tesis Doctoral]. Barcelona: Universitat Politècnica de Catalunya, 2003.

Costas Comesaña, R. (2003). *Desarrollo metodológico para la realización de estudios bibliométricos en el nivel micro: estudio de caso del Área de Recursos Naturales del CSIC.* [Tesina de doctorado]. Madrid: Universidad Carlos III, 2003.

Costas-Comesaña, R. y García-Zorita, J.C. (2003). "Indicadores de rendimiento en bases de datos bibliográficas: la tasa de filtrado del campo autor. Una aplicación al caso de los nombres de autores españoles", en *II Jornadas de Tratamiento y Recuperación de la Información (JOTRI)*, Getafe, Universidad Carlos III de Madrid.

Fernández, E.; García, A.M. (2003). "Accuracy of referencing of Spanish names in Medline", en *The Lancet,* 361(9369), 351-352, 2003.

Fernández, M.T.; Cabrero, A.; Zulueta, M.A.; Gómez, I. (1993). "Constructing a relational database for bibliometric analysis", en *Research evaluation,* 3 (1), 55-62, 1993.

Frías, J.A.; Romero Gómez, P. (1998). "¿Quiénes son y qué citan los investigadores que publican en las revistas españolas de biblioteconomía y documentación?", en *Anales de Documentación*, 1, 29-53, 1998.

Gross, A.D. (1991). "Getty Synoname: the development of software for Personal Name Pattern Matching", en *RIAO 91 conference proceedings.* Condé-sur-Noireau: Centre des Hautes Estudes Internationales d'Informatique, p. 754-63, 1991.

Harman, D. (1992). "Ranking algorithms", en Frakes, W.B. y Baeza-Yates, R. Eds. *Information retrieval: data structures and algorithms.* New Jersey: Prentice Hall, p. 363-392, 1992.

Katz, J.S.; Hicks, D. (1997). "Desktop scientometrics", en *Scientometrics,* 38 (1), 141-153, 1997.

Lardy, J.P.; Herzhaft, L. (1992). "Bibliometric treatments according to bibliographic errors and data heterogeneity: the end-user point of view", en *16th international online information meeting.* (London), Oxford, New Jersey: Learned Information.

Lee, D.L.; Chuang, H.; Seamons, K. (1997). "Document ranking and the Vector-Space Model", en *IEEE software*, 14(2), 67-75, 1997.

Patman, F.; Thompson, P. (2003). "Names: a new frontier in text mining", en *Intelligence and security informatics. Proceedings lecture notes in computer science.* (2665), 27-38, 2003.

_____, (2005). "Text mining, names and security", en *Journal of Database Management*, 16 (1), 54-59, 2005.

Ruíz-Pérez, R.; Delgado López-Cózar, D. y Jiménez Contreras, E. (2002). "Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies", en *Journal of Medical Library Association*, 90 (4), 411-30, 2002.

Spinak, E. (1995). "Errores ortográficos en el ingreso en bases de datos", en *Revista española de documentación científica*, 18, (3), 307-319, 1995.

Thompson, P. y Dozier, C.C. (2003). "Name searching and information retrieval", en Arxiv.org, 13 p. Accesible en: http://arxiv.org/html/cmp-lg/9706017. [Consulta 24-6-2005]

Torvik, V.I.; Weeber, M.; Swanson, D.R.; Smalheiser, N.R. (2005). "A probabilistic similarity metric for Medline records: a model for author name disambiguation", en *Journal of the American Society for Information Science and Technology,* 56(2), 140-158, 2005.

Zulueta, M.A.; Cabrero, A.; Bordons, M. (1999). "Identificación y estudio de grupos de investigación a través de indicadores bibliométricos", en *Revista Española de Documentación Científica.* 23(3), 333-348. 1999