Aula Abierta 2012, Vol. 40, núm. 2, pp. 65-74
ICE. Universidad de Oviedo
ISSN: 0210-2773

Normalización automática de registros obtenidos de la Web of Science

Antonio E. Serrano-López y Carmen Martín-Moreno
Universidad Carlos III (Madrid)

Los lenguajes de programación orientados al tratamiento automático de textos ya han demostrado anteriormente su utilidad para procesar y gestionar los registros bibliográficos obtenidos de diferentes bases de datos. Concretamente el lenguaje Perl ha sido utilizado en numerosas metodologías relacionadas con la bibliometría, para extraer los diferentes elementos de los registros bibliográficos, incorporarlos a bases de datos relacionales, procesarlos y obtener de ellos múltiples indicadores. Este trabajo tiene como objetivo el desarrollo de un sistema de normalización de datos, basado en scripts de Perl y en algoritmos de similaridad, que nos permitan realizar una normalización básica sobre los nombres de autor, direcciones y palabras clave procedentes de los registros obtenidos a través de la Web of Science. El procedimiento posee un margen de error muy pequeño y es especialmente eficiente en los nombres de instituciones, permitiendo eliminar más del 60% de la redundancia en este campo, un 10% en el caso de los nombres de autor y alrededor de un 50% en cuanto a las palabras clave.
Palabras clave: Normalización, autores, direcciones, palabras clave, Perl.

Automatic normalization of records from the Web of Science. Programming languages focused on automatic text processing have previously demonstrated their utility to process and manage bibliographic records obtained from different databases. Specifically, Perl has been used in many methodologies related to bibliometrics to extract the different elements of bibliographic records, incorporate them to relational databases, and process them to get multiple indicators. This work aims to develop a normalization system based on Perl scripts and similarity algorithms, which allow us to perform a basic normalization of author names, addresses and keywords from records obtained through Web of Science. The procedure has a very small margin of error and is especially efficient in the names of institutions, eliminating more than 60% of the redundancy in this field, 10% in the case of the author's names and about 50% on the keywords.
Keywords: Normalization, authors, addresses, keywords, Perl.

Tradicionalmente la obtención y tratamiento de datos de forma masiva ha estado limitada por el acceso a equipos informáticos de gran capacidad que, por añadidura eran costosos, de difícil manejo y prácticamente inaccesibles. Sin embargo, con la aparición de los ordenadores personales y su evolución hacia equipos cada vez más potentes y baratos, la obtención y tratamiento de la información, incluso en grandes conjuntos de datos, se ha vuelto un procedimiento mucho más sencillo, barato y accesible.

Cuando se traslada este escenario al área informétrica, se encuentra que inicialmente se realizaban los estudios bibliométricos prácticamente de forma artesanal, analizando manualmente las referencias bibliográficas. Posteriormente se comenzaron a reali-

Fecha de recepción: 30-04-2012 • Fecha de aceptación: 21-05-2012
Correspondencia: Antonio Eleazar Serrano-López.
Universidad Carlos III de Madrid.
Departamento de Biblioteconomía y Documentación. Laboratorio de Estudios Métricos de Documentación (LEMI). C/ Madrid, 126. C.P. 28903, Getafe (Madrid). Correo-e: aeserran@bib.uc3m.es

zar estudios de los trabajos con herramientas informáticas muy básicas, que generalmente, se corresponden con las que proporcionan los principales servicios de bases de datos bibliográficas, tales como el "Essential Indicators©" o el "Citation Report©" de Thomson Reuters. Estas herramientas realizan análisis muy limitados de la bibliografía, basados casi exclusivamente en el conteo de publicaciones en diferentes niveles de agregación y en el análisis de citas a un nivel muy básico. En muchos casos esta metodología es la única posibilidad que tienen aquellos investigadores, que no cuentan con formación y conocimientos informáticos, de desarrollar análisis de la literatura e información científica.

Sin embargo, cuando los investigadores cuentan con desarrollos informáticos, los análisis se vuelven mucho más potentes y son capaces de gestionar grandes conjuntos de datos que de otro modo, serían totalmente inmanejables. En este sentido, a finales de los años 90, Katz y Hicks (1997) acuñaron el término "Desktop Scientometrics", que traducido como "Bibliometría de escritorio" ha sido utilizado para referirse al conjunto de herramientas informáticas que permiten a los investigadores realizar estudios bibliométricos, con grandes conjuntos de datos, de forma sencilla y sin necesidad de poseer conocimientos informáticos avanzados.

Katz y Hicks (1997) desarrollaron en sus investigaciones un método que combinaba, sobre una arquitectura de software basada en UNIX (Linux/GNU), dos herramientas fundamentales, en primer lugar el lenguaje de programación Perl (Practical Extraction and Report Language), que les permitía procesar los registros obtenidos de las bases de datos Thomson Reuters (entonces ISI). Este lenguaje de programación, combinado con el editor de textos programable "Emacs", utilizado para normalizar las diferentes variantes de nombre en las instituciones analizadas, les permitió manipular, unificar y analizar un conjunto de 500.000 registros, publicados durante 14 años y obtenidos de la base de datos Science Citation Index (SCI).

Posteriormente, otros trabajos han profundizado en esta metodología. Tal es el caso de trabajos como el de Sanz-Casado, Suarez-Balseiro, García-Zorita, Martín-Moreno, y Lascurain-Sánchez (2002), que desarrollaron una metodología para el tratamiento de registros bibliográficos obtenidos de diversas bases de datos (ISI©, Medline©, EconLit©, INSPEC©, ICYT©, IME©, ISOC©, etc.). Esta metodología, que desarrolla el concepto de Bibliometría de Escritorio, se basaba en la utilización de gestores bibliográficos como Procite 5.0©, para la obtención de los datos brutos procedentes de las diferentes bases de datos, de forma que pudieran ser procesados posteriormente mediante diferentes herramientas ofimáticas (Microsoft Word©, Microsoft Excel©, WordPerfect©, etc.) y exportados a sistemas de análisis estadísticos como SPSS© o Statgraphics©, que permitiera la obtención de los indicadores bibliométricos pertinentes.

Por otro lado, el lenguaje Perl de cuyo uso se beneficiaron Katz y Hicks (1997) ha servido para el desarrollo de metodologías más avanzadas como la de Wouters (1999), que desarrolló un conjunto de scripts de Perl que permitían obtener diferentes indicadores bibliométricos a partir de los registros obtenidos de la base de datos SCI.

Basándose en la metodología de Katz y Hicks (1997) y Wouters (1999), García Gonzalez (2010) desarrolló a través de su tesis doctoral, un método de tratamiento de registros obtenidos de bases de datos bibliográficas, que con la ayuda del sistema de bases de datos MySQL, y Perl, permitiera diseñar, alimentar y gestionar una base de datos relacional en la que normalizar, extraer y procesar los datos bibliográficos con el objetivo de calcular cada vez un número mayor de indicadores con un volumen también mayor de registros procesados.

Actualmente, estas metodologías han permitido el desarrollo de herramientas importantes para la evaluación de la actividad científica en todo el mundo, que permiten manejar un gran volumen de registros bibliográficos y que, además, permiten combinar los datos puramente bibliométricos con

otros indicadores sociales, aportándole un valor añadido al análisis de la actividad científica e investigadora.

Objetivo

Este trabajo tiene como objetivo el desarrollo de un sistema de normalización de datos basado en scripts de Perl y en algoritmos de similaridad, que nos permitan realizar una normalización básica sobre algunos de los campos de los registros obtenidos a través de la Web of Science©, tales como los nombres de autor, las filiaciones institucionales o las palabras clave, sin perjuicio de que la normalización "manual" sea, con diferencia la más efectiva, aunque en algunos casos resulta inabarcable, debido al volumen de los datos con los que se trabaja. En estos casos se hace necesario el tratamiento automático de los registros, con el fin de minimizar el esfuerzo normalizador que debe realizar cualquier investigador en el área de la bibliometría. Para ello se han diseñado tres procedimientos diferentes, uno para cada uno de estos campos.

Metodología

En el caso de los autores el proceso es muy complejo, debido a dos factores fundamentales: En primer lugar que las cadenas de caracteres son muy cortas y en segundo lugar que un mismo nombre de autor puede pertenecer a diferentes personas. Para realizar la normalización de los nombres de autor se ha tomado como base el algoritmo n-grams (Egghe, 2000), separando y comparando los diferentes elementos que los componen. Este proceso se complementa con la comparación entre las filiaciones institucionales de los autores, así como los autores con los que firman habitualmente, con el objetivo de discriminar si dos nombres de autor con un gran número de elementos en común, se corresponden con el mismo autor o no. El proceso se compone de cuatro fases diferenciadas.

La primera fase consiste en preparar los datos procedentes de los registros de la Web of Science© para normalizarlos de la forma más eficiente posible. En primer lugar, para insertar los registros en un modelo de bases de datos SQL se utiliza la metodología propuesta por García Gonzalez (2010). Una vez que han sido cargados los registros bibliográficos en nuestra base de datos el siguiente paso consiste en determinar cuáles son los campos que se van a necesitar para conseguir la información más completa posible de cada autor para obtener las similaridades entre ellos.

El primer problema a resolver es que los registros cuentan con dos campos diferentes en los que existen nombres de autor. Por un lado el campo "author" contiene el nombre abreviado de los autores y está presente en todos los registros, mientras que en el campo "full author" presenta el nombre completo de los autores, pero sólo se encuentra disponible en los registros más recientes. Para solventar este problema se decidió crear un sistema de tablas intermedias en nuestra base de datos, de forma que más adelante se pudiera generar un fichero único de autores que contuviera todas las formas posibles de cada autor junto con sus datos de identificación principales.

Para crear las tablas intermedias se procede a separar los dos tipos de nombres de autor, de forma que por un lado se cuenta con una tabla con los autores procedentes del campo "author" asociados mediante el ID de cada registro con las instituciones del campo "reprint", que es el único que permite asociar algunos de estos nombres de autor con su dirección institucional. Mientras que por otro lado se genera una tabla con los nombres de autor completos, procedentes del campo "full author" así como sus filiaciones institucionales procedentes del campo "C1". Éste se caracteriza por incluir antes de cada dirección institucional, y entre corchetes, los nombres completos de todos los autores afiliados a ella. Por último se crea una tabla que pudiera contener los registros conjuntos de ambas tablas intermedias y que permitiera también asociar a los autores con sus direcciones.

La segunda fase consiste en construir la lista de autoridades a partir de las tablas inter-

medias de la fase anterior. Una vez determinada la estructura de las tablas que van a contener los nombres de autor con las direcciones asociadas, así como la tabla de autoridades, que servirá para hacer las comparaciones y detectar las variantes de nombres de autor junto con sus instituciones, el siguiente paso consiste en identificar, dentro de cada uno de los registros bibliográficos, a que institución pertenece cada uno de los autores firmantes y construir con esos datos la versión de la lista de autoridades que más tarde nos servirá para llevar a cabo el proceso de normalización.

Como ya se ha comentado anteriormente, en la base de datos contamos con dos campos que contienen nombres de autor (author y full author) así como dos campos que contienen direcciones institucionales (reprint y C1). Para asociar cada autor con su dirección correspondiente se siguió el siguiente procedimiento:

En primer lugar se aprecia que los autores del campo "full author" aparecen en la dirección del campo "C1" entre los signos [y]. Tomando como ejemplo la siguiente dirección:

[Gast, Marie-Christine W.; Beijnen, Jos H.] Slotervaart Hosp, Netherlands Canc Inst, Dept Pharm & Pharmacol, Amsterdam, Netherlands.

[van Tinteren, Harm] Antoni Van Leeuwenhoek Hosp, Netherlands Canc Inst, Dept Biometr, Amsterdam, Netherlands.

Esta dirección indica que "Gast, Marie-Christine W." y "Beijnen, Jos H." comparten la misma dirección institucional (Slotervaart Hosp, Netherlands Canc Inst, Dept Pharm & Pharmacol, Amsterdam, Netherlands) mientras que la institución de "van Tinteren, Harm" sería "Antoni Van Leeuwenhoek Hosp, Netherlands Canc Inst, Dept Biometr, Amsterdam, Netherlands". En este caso es sencillo asociar cada autor con su dirección institucional, basta con indicarle a la base de datos que si el nombre de autor procedente del campo "full author" se encuentra contenido en el texto de una institución quiere decir que el autor pertenece a ella.

Sin embargo, aunque los registros más recientes cuentan con los nombres de autor completos y la dirección institucional recogida en este formato, los registros más antiguos tan solo incluyen el nombre de autor abreviado y la dirección institucional del primer autor en el campo "reprint".

Esto dificulta enormemente la labor de asociar nombres de autor y direcciones, y solo se puede estar seguro de conseguir una buena asociación para el primer autor, siguiendo el mismo método que para el caso anterior, pues el formato de las direcciones en el campo "reprint" es el del siguiente ejemplo:

Sekowska, M, UBLG 1 LF UK, Albertov 4, Prague, Czech Republic.

Como se puede observar la dirección incluye únicamente el nombre del primer autor firmante, en este caso "Sekowska, M". Para aplicar éste método de asociación e insertar los nombres de autor junto con sus direcciones identificadas y el resto de datos bibliográficos, se diseñó un script en PERL que asociaba al autor con una dirección institucional siempre y cuando dicha dirección contuviera el nombre del autor. De esta forma, para las direcciones completas se tienen en cuenta los nombres completos de los autores y para las direcciones del campo "reprint" se tienen en cuenta los nombres abreviados.

En la tercera fase se aborda el proceso de comparación y normalización de los nombres de autor propiamente dichos. Una vez asignados a cada autor todas las posibles direcciones con las que aparece en los registros de la base de datos e incluidos en el fichero de autores preliminar, se pone en marcha el método de detección de duplicados, por medio del uso de n-grams de tres caracteres o trigramas, para obtener los valores de similaridad necesarios para determinar si dos o más nombres de autor se corresponden con un mismo autor.

Previamente debemos especificar que en este método, los umbrales de similaridad que se han establecido como los más óptimos para identificar duplicados, se han obtenido por medio de la observación directa de los registros. Utilizando un conjunto más reducido de registros y calculando sus similaridades por medio del mismo sistema, se

pudo determinar cuáles eran los valores que debía alcanzar dicho cálculo para considerar dos variantes de firma como pertenecientes al mismo autor.

Debido a que se trata de un proceso en el que intervienen diferentes medidas de similaridad, y que para las diferentes combinaciones de cada una de ellas es necesario aplicar distintos umbrales de similaridad, hemos creído necesario dividir el proceso en los pasos que se detallan a continuación.

Paso preliminar: En este momento se realizan tres funciones básicas. En primer lugar se llevan a cabo las consultas SQL sobre el listado de autoridades de nuestra base de datos. Posteriormente se preparan los datos extraídos para su procesamiento (escapando caracteres, separando los valores por campos y renombrando las variables) y por último se calculan los valores de similaridad para cuatro variables diferentes: apellidos, nombres de pila, direcciones institucionales y coautores.

Paso 1. Similaridad entre apellidos y nombres: Una vez que los datos se encuentran listos para ser procesados, comenzamos a incluir las estructuras de control o condicionales que determinan qué registros se considerarán duplicados y cuáles se mantendrán. Puesto que mantener la integridad de los datos es fundamental, se ha optado por no eliminar en ningún caso los registros duplicados de forma permanente, por lo que cada vez que un registro se identifica como duplicado, los datos que contiene, pasan a formar parte de los campos que incluyen todas las posibles variantes de autor, dirección, coautores, etc., del registro que se consideró como el más completo entre los que hacen referencia a un mismo autor.

En este primer paso las condiciones que se establecen son muy simples. Basta con que la similaridad entre los apellidos de dos nombres de autor sea superior a 0,4 y que la similaridad entre sus nombres de pila sea superior a 0,5 para que se consideren como el mismo autor. Sin embargo para que a uno de ellos se le asigne la etiqueta de "duplicado" es necesario que la otra variante de firma sea más completa.

Paso 2. Similaridad entre apellidos, nombres y direcciones: Esta segunda fase va orientada a detectar los duplicados que no se han detectado en la primera, o que habiéndose detectado, no se ha podido determinar cuál de ellos es el más completo (al procesarse únicamente los nombres de autor). Para ello, van a utilizarse unos umbrales de similaridad más bajos, tanto en los apellidos como en los nombres, que se filtrarán por medio de los valores de similaridad obtenidos entre las direcciones.

Asimismo esta fase se encuentra divida en dos partes. La primera comparte el objetivo de la fase anterior, de detectar las variantes de firma que sean lo más parecidas posibles, mientras que la segunda permite identificar dos variantes de firma con apellidos prácticamente idénticos, pero en los que al menos uno de los nombres de pila está en forma de iniciales. Para conseguir esto se aumenta el umbral de similaridad para los apellidos y se reduce casi al mínimo el del nombre de pila, al mismo tiempo que se exige como condición que la primera inicial del nombre sea idéntica. Además se reducen los umbrales de similaridad de las direcciones, pues se encontraron autores que cuando firmaban con su nombre completo incluían la dirección institucional en inglés, mientras que cuando firmaban con sus iniciales, el nombre de la institución aparecía en su lengua materna. Así pues, para la fase 2 van a utilizarse los umbrales de similaridad presentes en la tabla 1.

Tabla 1. *Umbrales de similaridad (Fase 2)*

Variables	Umbrales (Fase 2.1)	Umbrales (Fase 2.2)
Apellido	0,4	0,9
Nombre	0,4	0,05
Dirección	0,2	0,15

Por otra parte, del mismo modo que en la fase 1 se determinaba cuál era el registro más completo a partir de la longitud de los nombres de autor, en este caso se utiliza el mismo criterio, salvo que ambos nombres de autor tengan la misma longitud, en cuyo caso se toma como firma más completa aquella que incluya la dirección institucional de mayor longitud, sin olvidar que la dirección descartada se incluirá como variante, por lo que en cualquier caso seguirá asociada al nombre de autor normalizado.

Paso 3. Similaridad entre apellidos, nombres, direcciones y coautores: Cómo último paso para determinar las variantes de firma, se ha diseñado una tercera fase orientada a detectar nombres de autor con variaciones significativas en las direcciones que no pudieron identificarse en las fases anteriores, debido a que siempre se ha utilizado como filtro el hecho de que el campo ID no fuera idéntico en ambos registros. De este modo se evitaba comparar el mismo registro consigo mismo. Pero por el contrario, en los casos de autores que firman con diferentes instituciones en un mismo trabajo se van a encontrar duplicados.

Por poner algún ejemplo, algunos autores utilizan indistintamente su idioma materno y el inglés para reflejar su dirección institucional, por lo que un umbral de similaridad tan alto no permite identificarlos. En un principio se trató de ajustar el umbral de similaridad en las direcciones para detectar estos casos. Sin embargo, utilizando este método se encontraría un gran número de autores mal identificados. Por tanto se hacía necesario incluir una variable adicional que nos permitiera eliminar el filtro de ID's idénticas, sin que la precisión del sistema se viera afectada.

Después de realizar pruebas con diferentes variables (categorías, revistas y coautores), y teniendo en cuenta que las variables de categorías y revistas son demasiado parecidas cuando se trata de autores que trabajan en áreas muy cercanas, se optó por utilizar la variable coautores como filtro, otorgándole un umbral de 0,3 para filtrar los errores en la asignación. Por último para asegurar que en esta última fase no se comparaban registros idénticos se incluyó la condición de que las direcciones de ambos registros fueran diferentes.

En la fase final del desarrollo se procedió a optimizar el sistema con el fin de que fuera más eficiente. En este sentido, el desarrollo del sistema de normalización por fases se hacía necesario debido a que, para realizar los ajustes en los umbrales de similaridad, necesitábamos saber en que punto del proceso se identificaba cada autor concreto. Por este motivo y a efectos de agilizar el proceso, una vez que se contó con una versión definitiva del código fuente por fases, se decidió implementar en un nuevo script, que recogiera todas las condiciones que se había insertado en las diferentes fases del proceso, en una única estructura condicional.

Al mismo tiempo, se planteó la posibilidad de cambiar algunos de los procesos a la hora de realizar las consultas a la base de datos, para evitar la realización de tantas consultas como registros se encontraran en el fichero de autoridades. Pero esto suponía perder una funcionalidad importante, como es evitar la comparación de registros ya identificados como duplicados, y apenas suponía una mejora en cuanto a tiempo de ejecución y consumo de recursos.

Por este motivo, se optó por optimizar el sistema incluyendo dos nuevas condiciones, relativas a la similaridad de los nombres de autor antes de comenzar a calcular el resto de similaridades.

De esta forma, sólo para aquellos pares de autores que cuenten con un índice de similaridad en su apellido superior a 0,4 se calculará el índice de similaridad del nombre. Y si éste es superior a 0,05 (el valor mínimo en las condiciones posteriores), entonces se calcularán los índices de similaridad para las direcciones y los coautores. Con este proceso logramos reducir el tiempo de ejecución del script hasta una media aproximada de 2 horas para un fichero de autoridades compuesto por 4.393 registros.

Además, en la fase de optimización se detectaron ciertos errores a la hora de identificar las variantes de firma en las que en el

apellido sólo variaba la inicial (por ejemplo Konrad y Conrad), por lo que se incluyó una nueva condición para que no fueran consideradas como variantes de firma de un mismo autor, salvo que las iniciales del nombre y del apellido fueran idénticas.

En cuanto a la normalización de instituciones, se utiliza también un algoritmo de similaridad para comparar las cadenas de texto, con la diferencia de que al tratarse de cadenas mucho más largas que las de los nombres de autor, se puede utilizar un algoritmo mucho más eficiente, basado en el método estadístico "String Matching" (Hall y Dowling, 1980; Lehmann, 2007; Navarro, 2001), que compara el número de palabras en común que se encuentran entre dos cadenas de texto, agrupando aquellas direcciones mas similares entre si en una dirección única.

Para realizar este trabajo se programó un script, que se ejecutó en dos tiempos diferentes. En un primer momento para identificar las direcciones completas más similares entre si, y posteriormente, para identificar la institución matriz, tomando tan solo la primera parte de las direcciones institucionales, pues es la que recoge la institución matriz, que viene seguida por los departamentos, secciones, etc., en los que trabajan los autores.

Así pues, se agrupan las instituciones con una similaridad entre ellas mayor o igual al 85%. Debido a que el sistema está basado en algoritmos de similaridad y, al igual que ocurría en el método de normalización de autores, para conjuntos de datos más limitados, especialmente si es por razones geográficas, sería necesario ajustar los niveles de similaridad utilizados para minimizar el error en la asignación.

Debido a que los datos con los que trabajan los autores son fundamentalmente relativos al área de las Ciencias de la Vida y la Salud, era necesario contar con un sistema de normalización que permitiera trabajar con categorías temáticas más amplias que las que facilita la Web of Science©.

Con este objetivo, se decidió programar un método que permitiera asignar los términos del Medical Subject Headings© (MeSH) a los documentos obtenidos de la Web of Science©. Para ello se construyó un sistema en dos fases que, primero permitiera asignar mediante el análisis de las palabras clave, los términos MeSH correspondientes a través del Unified Medical Language System© (UMLS) (NIH, 2011). Y segundo, a través de algoritmos de similaridad, asignar los términos MeSH a aquellas palabras clave que, sin contar con asignación en la primera fase, cuente con una similaridad lo suficientemente alta como para asignarle un término MeSH sin riesgo de hacerlo erróneamente. A continuación se describen ambos procesos de forma pormenorizada.

Fase 1. UMLS y términos MeSH: El UMLS es un conjunto de software y ficheros que unifican diferentes vocabularios y estándares biomédicos con el fin de permitir la interoperabilidad entre sistemas informáticos y los diferentes vocabularios controlados que pueden utilizarse en las ciencias médicas.

El papel de este sistema en el proceso de normalización viene marcado por la disponibilidad de un módulo programado en Perl que permite comparar términos no controlados con el macro-tesauro UMLS y obtener aquellos términos controlados que son equivalentes.

Para llevar a cabo este proceso, se compara cada una de las palabras clave que los autores han asignado a sus propios trabajos en la base de datos SCI, con el macro-tesauro UMLS, a través del módulo de Perl programado por McInnes, Pedersen y Pakhomov (2009), y en el caso de que dicha palabra clave coincida con un término presente en el tesauro (ya sea un término aceptado o no) se le asigna el término aceptado que le corresponda. De este modo, salvo que la palabra clave no se encuentre en el tesauro (algo complicado, debido al enorme tamaño del macro-tesauro UMLS, salvo que exista algún error tipográfico), se le asigna un término aceptado proveniente del tesauro MeSH.

Fase 2. Asignación a través de algoritmos de similaridad: La segunda fase de la normalización consiste en comparar, por

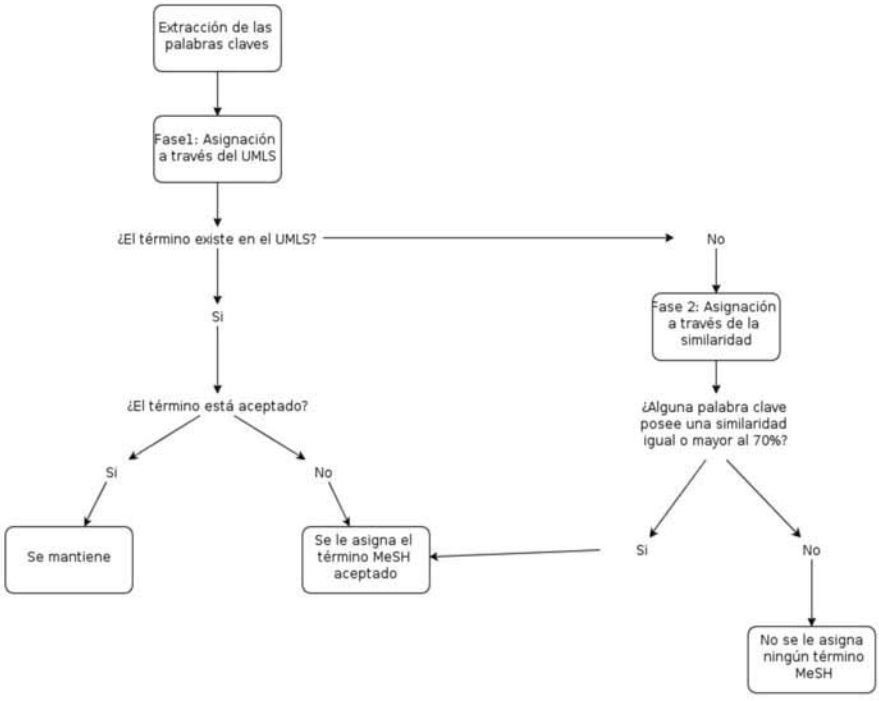

Figura 1. *Método para la asignación de términos MeSH por medio de las palabras clave*.

medio de un algoritmo de similaridad basado en los métodos de "String Matching" y n-grams, las palabras clave que no han recibido asignación en la fase 1 con aquellas que si lo han hecho. El objetivo es subsanar fundamentalmente errores de tipo ortográfico y tipográfico, ya que estas palabras clave han sido asignadas manualmente por los autores.

Tras múltiples pruebas, se encontró que el porcentaje de similaridad ideal era del 70%, por lo que se decidió que si la similaridad entre dos palabra clave supera ese umbral, se le asigna el término MeSH correspondiente, mientras que si no lo supera la palabra clave no recibe asignación. En la figura 1 se encuentra descrito el proceso de normalización en sus dos fases.

El procedimiento de normalización de los nombres de autor es con diferencia el más complejo de los tres. Este sistema ha sido operativo para la normalización de los nombres de autor contenidos en un conjunto de datos compuesto por 65.976 variantes de nombre diferentes. Sin embargo, ha sido necesario mantener ejecutando el procedimiento durante más de 300 horas para completar la normalización, pues al aumentar el número de autores a comparar, aumenta exponencialmente el tiempo necesario para su ejecución.

Por lo tanto, debido a lo costoso del sistema (en términos de ciclos de procesador), se hace necesaria su optimización, con el objetivo de utilizarlo en trabajos futuros en los que el número de autores a normalizar sea mucho mayor.

En cuanto al procedimiento de normalización de instituciones, como se puede observar, el sistema permite agrupar, de forma masiva, las instituciones más similares entre si con un desviación mínima que, en un conjunto de datos limitado y revisado por el autor, y utilizado como banco de pruebas, se encontraba alrededor del 1%.

Tabla 2. *Ejemplo de instituciones normalizadas*.

Institución normalizada	**Institución original**
Fdn Jimenez Diaz	Fdn Jimenez Diaz, Dept Neurol, E-28040 Madrid, Spain. Fundac Jimez Diaz, Dept Genet, E-28040 Madrid, Spain.
Univ Complutense Madrid	Univ Complutense Madrid, Mitochondrial & Hereditary Metab Dis Unit, E-28040 Madrid, Spain. Univ Complutense Madrid, Stat & Epidemiol Unit, E-28040 Madrid, Spain. Univ Complutense, Dept Immunol, Fac Med, E-28040 Madrid, Spain.
Inserm	Inserm, U29, Marseille, France. Inserm, Umr Neurol & Therapeut Expt S679, Paris, France. Inserm, U866, Ctr Rech, Lab Biochim Metab & Nutr, F-21000 Di- jon, France.
Manchester Childrens Univ Hosp	Manchester Childrens Univ Hosp, Dept Resp Med, Manchester, Lancs, England. Manchester Childrens Univ Hosp Nhs Trust, Manchester, Lancs, England. Manachester Childrens Univ Hosp, Nhs Trust, Dept Med Genet, Manchester, Lancs, England.
Appl Biosyst Inc	Appl Biosyst Inc, I-20052 Monza, Italy. Appl. Biosyst Inc, Madrid, Spain. Appl Biosyst Inc., Warrington, Cheshire, England.

En la tabla 2 se muestran algunos ejemplos de instituciones normalizadas, en las que podemos observar que el sistema permite eliminar gran parte de la redundancia en las filiaciones institucionales. En el conjunto de datos utilizado, que cuenta con más de 80.000 direcciones diferentes, el sistema consiguió una reducción de la redundancia de alrededor del 60%.

Por último, los resultados obtenidos mediante el método de normalización temática muestran que, del total de documentos únicos obtenidos a través de la Web of Science© (19.415), al 77,39% (15.025) se les ha conseguido asignar al menos un término MeSH aceptado que describa el contenido del documento de forma más profunda que las tradicionales categorías WOS o las categorías temáticas empleadas por la Web of Science©.

Finalmente es necesario indicar que esta metodología de normalización ha sido aplicada con éxito durante el desarrollo de la tesis doctoral del autor (en proceso), así como en otros estudios bibliométricos (Serrano-López y Martín Moreno, 2009; Serrano-López y Martín-Moreno, 2011).

Conclusiones

En su conjunto, el procedimiento posee un margen de error muy pequeño y es especialmente eficiente en los nombres de instituciones, permitiendo eliminar más del 60% de la redundancia en este campo.

En cuanto a los nombres de autor es menos eficiente, reduciendo aproximadamente un 10% de la redundancia. Además el sistema ha demostrado su validez sobre un gran conjunto de autores internacionales, por tanto, si se quisiera aplicar a un conjunto de autores, una materia, revista, país o institución concreta, sería necesario ajustar los umbrales de similaridad para adaptarlos al conjunto de autores con el que se vaya a trabajar.

Por otro lado su aplicación a las palabras clave no solo reduce la redundancia a valores cercanos al 50%, si no que además per-

mite asignar términos del lenguaje controlado MeSH a registros de la Web of Science©, que no los incorporan por si mismos, aportando un gran valor añadido al análisis temático de los registros bibliográficos en el área de las Ciencias de la Vida y la Salud.

Referencias

Egghe, L. (2000). The distribution of N-grams. *Scientometrics*, *47*(2), 237-252.

García González, P. E. (2010). *Diseño, desarrollo y aplicación de un método para el análisis y tratamiento de la información con fines métricos*. Tesis doctoral no publicada. Universidad Carlos III de Madrid. Madrid.

Hall, P. A. V. y Dowling, G. R. (1980). Approximate string matching. *ACM Computing Surveys (CSUR)*, *12*(4), 381-402.

Katz, J. S. y Hicks, D. (1997). Desktop scientometrics. *Scientometrics*, *38*(1), 141–153.

Lehmann, M. (2007). S*tring:Similarity - calculate the similarity of two strings*. Recuperado el 10 de abril de 2012, de http://search.cpan.org/~mlehmann/String-Similarity-1.03/Similarity.pm

McInnes, B. T., Pedersen, T., y Pakhomov, S. V. S. (2009). UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium*, *2009*, 431-435. American Medical Informatics Association.

NIH. (2011). *Unified Medical Language System*. Recuperado el 10 de abril de 2012, de http://www.nlm.nih.gov/research/umls/

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, *33*(1), 31-88.

Sanz-Casado, E., Suárez-Balseiro, C., García-Zorita, C., Martín-Moreno, C., y Lascurain-Sánchez, M. L. (2002). Metric studies of information: An approach towards a practical teaching method. *Education for information*, *20*(2), 133–144. IOS Press.

Serrano-López, A. E., y Martín Moreno, C. (2009). Producción y consumo de información científica en las Ataxias Raras con causa genética identificada (2003-2007). *A ciencia da informaçao criadora de conhecimento: Actas do IV Encontro Ibérico EDIBCIC 2009* (pp. 485-494). Universidade de Coimbra. Recuperado el 10 de abril de 2012, de http://dialnet.unirioja.es/servlet/articulo?codigo=3098525&info=resumen

Serrano-López, A. E., y Martín-Moreno, C. (2011). Cadasil e Síndrome de Rett: Estudo de caso de dois doenças raras neurológicas. *Ponto de Acesso*, *5*(3), 130-148. Recuperado el 10 de abril de 2012, de http://www. portalseer.ufba.br/index.php/revistaici/article/view/5506

Wouters, P. (1999). *The citation culture*. Tesis doctoral no publicada. Universiteit van Amsterdam. Amsterdam.