ARTÍCULO TEÓRICO-METODOLÓGICO

Requerimientos para mejorar la normalización de datos en software de análisis métricos de la información

Requirements to improve the normalization of data in software for metric analysis of information

Ramón Orlando Albo Hernández,[I] María Victoria Guzmán Sánchez,[I] Ivet Álvarez Díaz,[I] Jesús Francisco Bouza Figueroa,[I] Romel Calero Ramos[II]

[I] Instituto "Finlay". Dpto. Inteligencia Empresarial. La Habana, Cuba.
[II] Centro de Ciencias de la Complejidad C3. México.

RESUMEN

Por la diversidad de formas en la entrada de los campos de autor-afiliación, la normalización de los datos bibliográficos es uno de los problemas que limitan los análisis de información métrica en tiempo de ejecución, fiabilidad de los indicadores y tamaño del *corpus* de datos. Este trabajo tiene como objetivo proponer los requerimientos para el mejoramiento de la normalización de datos en software de análisis métricos. Para lograr el objetivo se realizó un diagnóstico de los principales métodos y técnicas que son empleados a nivel mundial en este tipo de estudio. Como resultado principal, se relacionan los requerimientos para una aplicación de preprocesamiento automatizado de datos con fines métricos. Se proponen la base de datos, las tareas, los pasos y los algoritmos que contendrá esa aplicación. Se debe usar una combinación de algoritmos para desambiguar los campos afiliación y autor.

Palabras clave: procesamiento de datos; minería de datos; bibliometría; descubrimiento basado en la literatura; análisis de datos.

ABSTRACT

Due to the diversity of methods used to enter author-affiliation information, the resulting lack of standardization of bibliographic data has become one of the problems limiting analysis of metric information in terms of execution time, reliability of

indicators and size of the data corpus. The purpose of the study was to propose requirements to improve data normalization in metric analysis software. To achieve this objective, a diagnosis was made of the main methods and techniques used worldwide in this type of study. The main result is the presentation of requirements to be met by an application for automated pre-processing of data for metric purposes. A proposal is made of the database, tasks, steps and algorithms that this application will contain. A combination of algorithms should be used to disambiguate author and affiliation fields.

Key words: data processing; data mining; bibliometrics; literature based discovery; data analysis.

INTRODUCCIÓN

Existen tres tipos de conceptos científicos: clasificatorios, comparativos y métricos. Los dos primeros son cualitativos, mientras que los métricos son cuantitativos. La obtención de conceptos cuantitativos se basa en la medición de las magnitudes correspondientes.[1] En ese sentido, formar conceptos cuantitativos en ciencias sociales, y específicamente en la Ciencia de la Información, radica en llevar los conceptos cualitativos a cuantitativos; es decir, en buscar unidades de medición para las diferentes problemáticas asociadas a las actividades informativas, como lo son los análisis asociados al estudio de la actividad científica y tecnológica a diversos niveles de complejidad. Es por eso, que se trata de identificar el desarrollo científico de un país, representar un dominio del conocimiento, caracterizar los flujos en la colaboración, etc., a partir de indicadores métricos como conteo de artículos, conteo de patentes o la cantidad de colaboración que registra una institución, entre otras magnitudes.[2]

La obtención de conceptos métricos tiene en la actualidad varias acepciones, como bibliometría, informetría, cienciometría, patentometría, etc. en función del objeto y del tema de estudio de cada una de ellas. Sin embargo, de forma general se podría plantear que las metrías son una disciplina instrumental, que aplica indicadores métricos a la información registrada en diferentes soportes, empleando técnicas provenientes de cualquier algoritmo de análisis y visualización; es decir, implica la aplicación de un algoritmo a cualquier conjunto de datos con significado. Este conjunto de datos está contenido en un soporte y proviene de una fuente determinada como bases de datos (BD) digitales o en papel. Igualmente, estos soportes tienen diversas estructuras y propósitos. En este trabajo, se abordarán las problemáticas asociadas a las BD digitales de tipo bibliográfico, independientemente de su contenido (biológicas, de patentes, prensa, etcétera).

Estas BD tienen una gran probabilidad de registrar datos "sucios" o con "ruido", por la manera en que se recoge la información, los diferentes formatos de las citas, las violaciones de las restricciones de integridad y de los estándares, los nombres de autores muy frecuentes o ambiguos, las abreviaturas de los nombres de las fuentes de publicación y los grandes volúmenes de datos de las citaciones, etcétera. La ambigüedad en los nombres de los autores en las bases de datos bibliográficas ha sido reconocida desde hace tiempo como un problema importante.

El resultado de los análisis métricos a partir de estos valores con "ruido" puede llevar a interpretaciones erróneas y poco realistas. Igualmente, puede conducir a encubrir patrones de comportamiento útiles que están escondidos en los datos, así como a un bajo rendimiento y a una baja calidad en las salidas. Todos estos elementos son causas importantes para concentrar esfuerzos en la preparación de datos. Esta problemática coincide con lo reportado en la literatura consultada.[3-6] Según *Spinak*[7] y *Lardy,*[8] estos problemas constituyen un importante inconveniente para la explotación métrica de las BD.

Para nombrar la etapa de preprocesamiento de datos suelen utilizarse indistintamente términos como preprocesamiento, normalización y preparación de datos. En resumen, el preprocesamiento de los datos contenidos en las BD, hasta el momento actual, tiene las siguientes limitaciones y particularidades:

1. Existen numerosas causas que provocan "suciedad" en los registros de los sistemas, lo que trae como consecuencia que haya gran cantidad de datos almacenados que carecen de la calidad adecuada para ser utilizados de forma confiable y se hace necesario tratarlos de diferentes formas.

2. La limpieza de datos se divide en varios pasos: separar elementos, estandarizar, verificar, comparar, agrupar y documentar.

3. La corrección de datos incluye la eliminación de registros duplicados o con valores inválidos. En muchos casos, la información y el conocimiento disponibles son insuficientes para determinar las transformaciones necesarias para eliminar las anomalías; solo nos queda la eliminación de esos registros como única solución práctica, a pesar de que puede conducir a la pérdida de información.

4. Es la etapa en la que el analista invierte mayor cantidad de tiempo y esfuerzo.

5. Es altamente costoso en tiempo de cómputo.

Actualmente, los problemas en las BD pueden ser reconocidos y resueltos de dos maneras:

a) Manualmente con la participación de un especialista. Gran parte de las tareas de esta etapa se realizan de forma manual o con muy bajo nivel de automatización por expertos en el tema en cuestión, ya que se requiere conocimiento previo del tema y determinada pericia y habilidades.[9]

b) Automatizadamente, en menor o mayor grado, con la utilización de herramientas para la detección de los valores particulares que están en contradicción con algunas dependencias funcionales implícitas en la BD.[10] Sin embargo, las herramientas de software que existen orientadas al análisis métrico no tienen suficientemente soportado el preprocesamiento.

Ante el contexto explicado anteriormente, este trabajo se ha trazado como objetivo proponer los requerimientos para mejorar la normalización de datos en software de análisis métrico.

FUNDAMENTACIÓN DEL MÉTODO PROPUESTO

DIAGNÓSTICO

Se realizó un diagnóstico de los métodos y las tendencias empleadas en el preprocesamiento de datos. Con este fin se hizo una búsqueda en la BD multidisciplinar Scopus y se obtuvieron 486 registros en el período 1985-2015. La estrategia seguida fue buscar los posibles términos asociados al objeto de estudio, ya que no hay un descriptor único definido para esta área del conocimiento. Este análisis sirvió de base para identificar el campo de investigación y los algoritmos que pueden utilizarse para el preprocesamiento.

Los 486 documentos fueron analizados usando la técnica de análisis documental clásico. Sobre la base de los métodos usados a nivel internacional para el preprocesamiento y los algoritmos que se describen, se obtuvieron como resultados los requerimientos fundamentales que debe poseer una aplicación informática para el preprocesamiento de datos. Se explican el diseño de la base de datos, las tareas que la aplicación debe realizar y cómo se articulan.

El mapa obtenido fue procesado siguiendo la metodología ViBlioSOM, la cual permitió, a partir del algoritmo de los Mapas Auto-Organizados (SOM o *Self-Organizing Maps*), organizar la información de entrada de forma automática y visualizar relaciones importantes entre los datos. En este caso se representan los descriptores asociados a los artículos analizados.

El mapa representado en la figura 1 permite apreciar que las técnicas de preprocesamiento de datos van desde los árboles de decisión (*Decision Trees*, *cluster* C8) hasta los algoritmos genéticos (*Genetic Algorithms*, *cluster* C2). Se pueden identificar otras técnicas como *Artificial Neural Networks, Association Rule Mining, Bayes Theorem, Clustering, Discretizations, Heuristic Methods, Hierarchical Systems, Feature Selection, Learning Algorithms and Machine-Learning* . Estas técnicas se enmarcan en el campo de la inteligencia artificial y se aplican principalmente para la desambiguación de los datos de autor y afiliación, así como para el procesamiento del lenguaje natural en el tratamiento de textos.

En el mapa no se aprecian de forma explícita aquellas técnicas que se enmarcan dentro de la estadística y que fundamentalmente se aplican para el tratamiento de errores y solución de conflictos; es decir, para el tratamiento de faltantes, valores atípicos, duplicados y datos con ruido, aunque todas estas técnicas se encuentran muy interrelacionadas entre sí.

En la BD analizada no se encontraron estudios en Cuba que abordaran las problemáticas relacionadas con el preprocesamiento de datos enfocado a las metrías. Esto contrasta con el aumento que se reporta en la literatura, en el uso de técnicas métricas y su nivel actual de aplicabilidad (estudios de inteligencia, vigilancia científico-tecnológica, evaluación de proyectos, etcétera).[2]

Fig. 1. Mapa de las líneas de investigación relacionadas con el preprocesamiento de datos.

CAMPOS DE LAS BASES DE DATOS BIBLIOGRÁFICAS A NORMALIZAR

En el cuadro 1 se detallan los campos a normalizar y los tipos de estudios métricos relacionados. Los campos procedentes de las BD bibliográficas que tienen alta relevancia y mayor necesidad de preprocesamiento para los estudios métricos son los autores, la afiliación de procedencia del autor o signatario de una patente, así como los datos de carácter temático (descriptores, términos MeSH o Medical Subject Headings, temas, etcétera).[11]

Cuadro 1. Campos a normalizar y estudios relacionados

Campos de las bases de datos bibliográficas	Tipos de estudios métricos
Autor (es)	Productividad científica individual, colaboración científica, frentes de investigación, autocitación, importancia de un proyecto (cantidad de autores que firman artículos), etc.
Afiliación	Productividad científica institucional, por país, dependencia tecnológica y/o científica, colaboración.
Descriptores o temas	Líneas de investigación, representación de dominios del conocimiento, frentes de investigación.
Fecha de publicación	Obsolescencia, dinámica de un campo de investigación, potencial científico y tecnológico.
Títulos de revista	Revistas periféricas.
Citas	Citación relativa, visibilidad, impacto, transferencias de la I+D a la tecnología, etcétera.
Palabras/título o resumen	Representación de dominios temáticos, frentes de investigación, líneas estratégicas, etcétera.

MÉTODOS GENERALES PARA EL PREPROCESAMIENTO DE LOS DATOS

Los mejores resultados de los métodos y algoritmos de preprocesamiento dependen de la naturaleza de cada conjunto de datos. El papel que juega la experiencia del analista de datos también es relevante. A continuación, se explican los métodos generales identificados que fueron apropiados y factibles de emplear en este estudio.

Métodos para la desambiguación de los datos "autor" y "afiliación del autor"

El problema de asociar los nombres con las entidades reales es conocido como desambiguación de nombres. La desambiguación de los nombres de los autores es un proceso que pretende simultáneamente separar los casos de nombres ambiguos referidos a individuos diferentes y fusionar los casos de variantes de nombres referidos a un mismo individuo. El problema de la desambiguación de los nombres de autores comprende:

1. *Sinonimia:* un mismo individuo puede publicar con múltiples nombres. Esto incluye:

a) Variantes ortográficas y cambios de letras.

b) Errores mecanográficos.

c) Cambios de nombre en el tiempo como ocurre por matrimonios, divorcios, conversión religiosa o cambio de sexo.

d) Uso de seudónimos, alias y variantes de los nombres y apellidos.

2. *Homonimia:* muchos individuos diferentes tienen el mismo nombre.

3. Los metadatos necesarios a menudo están incompletos o faltan.

4. Muchas publicaciones tienen varios autores, quienes además representan múltiples instituciones.

El campo "afiliación institucional" presenta una ambigüedad análoga. Muchas afiliaciones pueden aparecer con variantes distintas en la BD. También se producen ambigüedades producto de la jerarquía, ya que algunas instituciones pertenecen a otras y pueden interpretarse como si fueran distintas. El amplio uso de acrónimos y siglas para identificar las instituciones también es origen de ambigüedades.

La desambiguación de los nombres de los autores y afiliaciones es un paso fundamental para la identificación de los dominios del conocimiento y para otros análisis métricos.[12] En este estudio se comprobó que no existe un método de desambiguación que deba ser tomado como paradigma. Cada tarea de investigación, cada base de datos, cada conjunto de datos tiene sus particularidades propias. Debe buscarse la flexibilidad del método y el balance conveniente entre exactitud, escalabilidad y tiempo de cómputo. Además, se apreció cómo cada uno de los distintos autores consultados experimenta diversas variantes y enfoques, combinando diferentes funciones y algoritmos de desambiguación en distintas BD y luego comparan la eficiencia y los resultados obtenidos por otros autores con los propios. El cuadro 2 incluye un resumen de los métodos que aparecen en la literatura sobre la desambiguación de nombres.

Una de las propuestas más usadas es la de *Torvik,*[12] quien plantea que la mayoría de los métodos de desambiguación resumen las puntuaciones de todas las características en un solo número, que indica el grado de similitud de un par de artículos. *Torvik*[26] elaboró un modelo para generar automáticamente los conjuntos de datos para entrenamiento y posterior estimación de la probabilidad, de que un par de artículos de Medline que poseen el mismo apellido y la primera letra del nombre son del mismo autor, basados en otros metadatos (título, nombre de la publicación, MeSH, coautores, afiliación).[11,27]

Han, Zha y Giles[15] consideran que la aplicación "k-way" del método de clusterización espectral con descomposición QR, brinda mejores resultados que los métodos tradicionales de clusterización que, por ejemplo, el k-medias (k-means). *Giles* y otros[16] aplican primero un método de poda por autor y luego una clusterización empleando como función de distancia SVM. Por otra parte, *Bhattacharya*[12] propone una adaptación de Asignación Dirichlet Latente (LDA). Los autores pueden pertenecer a uno o varios grupos de individuos que tienden a escribir juntos. Este método descubre simultáneamente *clusters* de autores-individuos y *clusters* de artículos, lo cual tiene un alto costo computacional. Emplean un método de entrenamiento no supervisado y el algoritmo de "esperanza-maximización" (cuadro 3). Un primer paso de limpieza, estandarización y poda por el campo apellido, es propuesto por *Pino-Mejías*.[20] Posteriormente, se deben realizar comparaciones entre seis campos de cada par de artículos y se calcula un índice de semejanza entre 1 (las cadenas son iguales) y 0 (las cadenas son totalmente distintas) empleando funciones de similitud (cadena exacta, Levenshtein, Jaro, Winkler). El conjunto de datos obtenido se somete a una clusterización.

Cuadro 2. Métodos generales de preprocesamiento de datos

Método	Comentarios
Tratamiento a valores atípicos (*outliers*) o con ruido	
Eliminación	Puede provocar pérdida de información.
Agrupamiento	Se agrupan los valores de los atributos en grupos, se detectan y luego se eliminan o imputan (sustituyen) los *outliers* por un valor.
Muestreo	Se considera solamente una parte de los valores de los atributos y luego se estima un valor representativo.
Ordenamiento	Consiste en ordenar los valores de atributos, dividirlos en intervalos y luego escoger para cada intervalo un valor representativo.
Data *scrubbing*	Consiste en limpieza de errores tipográficos.
Tratamiento a datos faltantes	
Ignorar y descartar datos faltantes	Se eliminan del conjunto de datos los registros que contengan valores faltantes de un atributo. Es apropiada para los conjuntos de datos con aleatoriedad de la clase MCAR (*Missing Completely At Random*), ya que no se introduce sesgo en los datos.
Imputar los valores faltantes	Consiste en sustituir los valores faltantes mediante alguno de los métodos de imputación.
Imputación de valores atípicos y datos faltantes	
Regresión	Se remplazan todos los valores faltantes o atípicos con un estadígrafo y se asume una relación lineal entre los atributos.
Hot-deck	Se remplazan todos los valores faltantes o atípicos con una distribución estimada de los datos reales (o sea, no imputados previamente).
Cold-deck	Es similar al *hot-deck*, pero el valor que se imputa tiene que ser tomado de una fuente de datos diferente.
Estimación de parámetros	Mediante procedimientos que usan variantes del algoritmo esperanza-maximización se estiman valores para sustituir los faltantes o atípicos.
k-NN	Consiste en remplazar todos los valores faltantes o atípicos con el k-vecino más próximo (kNN) mediante un algoritmo que determina la similitud de dos instancias con el empleo de una función de distancia.
Imputación múltiple	Se remplaza cada valor faltante por diferentes valores y posteriormente, con el uso de todo el conjunto de datos,
	se calcula el promedio de los valores y así se obtiene el conjunto de datos completo.
Imputación personalizada	Se remplazan los valores faltantes o atípicos por un valor de los existentes en la distribución.
Tratamiento a textos	
Limpieza	Se eliminan los caracteres no significativos.
Eliminación de *stopwords*	Se eliminan las palabras carentes de significación.
Lematización y *stemming*	Consiste en la representación de las palabras por su raíz.
Detección de frases	Consiste en la extracción de términos significativos compuestos por 2 o más palabras.

Cuadro 3. Principales enfoques para desambiguación de nombres

Tipo	Autor	Métodos	Comentarios
Supervisado	Han, Giles[13]	Naïve Bayes	Inapropiado para grandes bases de datos (BD) por el alto esfuerzo humano.
No supervisado	Han, Zha, Giles[14]	K-means	Brinda resultados aceptables en la *clusterización*, pero el k-way spectral es superior.
No supervisado	Giles, Han, Zha[15]	K-way spectral	No es viable para grandes BD por la alta complejidad computacional.
No supervisado	Huang, Giles[16]	DBSCAN	Eficiente para grandes BD. Está implementado en Weka.
No supervisado	Song[12]	PLSA (análisis semántico probabilístico latente)	Alta utilización de temas y otros metadatos.
No supervisado	Jordan, Song, Getoor y Bhattacharya[8]	LDA (asignación Dirichlet latente)	Modelo bayesiano jerárquico.
No supervisado	Bhattacharya y Getoor[12]	Resolución de entidades colectivas/ EM (esperanza-maximización)	Adaptación del LDA. Descubre *clusters* de autores y de artículos, con un alto costo computacional.
supervisado	Tejada[17]	Árboles de decisión	Mediante el aprendizaje de reglas de similitud.
Híbrido	Huang, Giles[16]	DBSCAN/LASVM (*Online Support Vector Machines*)	Es un *framework* (entorno de trabajo) para procesar grandes BD. Emplean la funciones de distancia LASVM que es supervisada y Soft-TFIDF es híbrida.
Híbrido	Ferreira[18]	SAND (*Self-training Associative Name Disambiguator*)	En una comparación fue superior a DBSCAN y K-way, a pesar de que utiliza pocas características.
Híbrido	Torvik, Smalheiser[19]	Author-ity	Procesaron Medline completo en el año 2006. La precisión se estima en el 98 %.
			http://arrowsmith.psych.uic.edu/
Híbrido	Pino-Mejías[20]	Clasificadores No supervisados: fellegi-sunter, *farthest first*, esperanza maximización. supervisados: árboles de decisión, bosques de árboles de decisión aleatorios.	Es un modelo flexible que admite distintos clasificadores y funciones de distancia. Disponible en la plataforma FEBRL. (Programada en R System y Python).
Otro	Jijkoun[21]	*Named Entity Normalization*	Algoritmo enfocado a textos y no emplea metadatos. Usado en Wikipedia.
No supervisado	Treeratpituk y Giles[17]	Emplean bosques de árboles de decisión aleatorios.	Empleado en Medline con mejores resultados que SVM (Support Vector Machines) y Adaboost.
No supervisado	Bolikovski[22]	Emplean Single-linkage Hierarchical Agglomerative Clustering (SLHAC).[23]	Muy flexible, admite cualquier función *hash*, de distancia y clasificador. Programado en Java.
Otro	Bordons y Costas[11]	Algoritmos para la detección de firmas similares para ISI.	No se ajusta al problema de este estudio, ya que se conocen los nombres reales de los autores.
No supervisado	Magnani y Montesi[24]	*Company name matching*	Algoritmo para la desambiguación de entidades aplicado a BD de patentes, valorando la significación de los términos (*tokens)*.
No supervisado	Ferreira[25]	Emplean clusterización jerárquica basado en heurística (HHC, por sus siglas en inglés).	No requiere entrenamiento y tuvo un mejor desempeño en las pruebas realizadas que otros métodos supervisados y no supervisados tales como SVM, *Naive Bayes Clasiffier*, *K-way Spectral* y DBScan.

Al analizar los métodos resumidos en el párrafo anterior, se corrobora que esos autores aplican principios comunes, los cuales deben ser tenidos en consideración en las propuestas de este estudio, como son entre otros: aplicar una "limpieza" previa, un mecanismo de poda para reducir la complejidad, funciones de distancia y algoritmos de clusterización.

Ordoñez[28] ha experimentado que la mejor forma de realizar el preprocesamiento es aprovechando las ventajas de los sistemas de gestión de bases de datos (SGBD), como es el lenguaje SQL (*Structured Query Language*). Este es un aspecto importante a considerar para la aplicación propuesta, pues se considera más conveniente contar con un entorno integrado soportado sobre un SGBD a emplear aplicaciones externas conocidas como ETL (*Extract-Transform-Load*) concebidas para más amplio espectro de problemas y entornos específicos.

PROPUESTA DE LOS PROCEDIMIENTOS Y MÉTODOS A EMPLEAR EN EL PREPROCESAMIENTO

A partir del análisis de cada uno de los métodos se establecieron las siguientes premisas, las cuales podrían ser pautas a seguir en el diseño de la aplicación:

1. El algoritmo para la desambiguación debe poseer al menos tres componentes principales, que se ejecutan secuencialmente:

- Un mecanismo de selección o poda (*blocking*) mediante una función *hash,* para dividir el conjunto de datos y así reducir el costo computacional. Un ejemplo de esto es el propuesto por *Bilenko,*[12] que separa en grupos los autores que tienen igual el apellido y la inicial del primer nombre.

- Una función de comparación de similitud para analizar pares de registros o cadenas. Esta función debe determinar si dos registros o cadenas se refieren a una misma entidad basados en algunos atributos o características, por lo que su salida debe ser una decisión binaria (sí o no) o un índice de semejanza, que generalmente es entre 0 y 1. Esta función puede ser basada en *tokens* o en distancias de edición.

- Un algoritmo de clasificación que puede ser supervisado, no supervisado o híbrido. Algunos clasificadores supervisados como SVM y los árboles de decisión también pueden ser usados como función de comparación.[17]

2. El algoritmo para la desambiguación debe resumir las puntuaciones de todas las características para indicar el grado de similitud de un par de registros, pero teniendo en cuenta que estas sean independientes entre sí.[12]

3. Los errores de cambios de letra en los nombres y apellidos de los autores pueden ignorarse. *Torvik* y *Smalhaiser*[22] demostraron que aparecen en aproximadamente 1,8 % en la BD Medline.

4. La selección de características es el aspecto más importante en el diseño de un modelo de desambiguación, porque determina el límite superior de precisión. Un buen criterio es emplear la mayor cantidad posible de características útiles disponibles, porque utilizar solamente una o pocas características probablemente limiten los resultados del método.[12]

5. El algoritmo de desambiguación debe ser lo suficientemente flexible como para que el usuario pueda ajustarlo a las necesidades del análisis que está realizando, las características del conjunto de datos, etcétera.

El preprocesamiento de los campos "afiliación" y "autor" es más complejo que el tratamiento de otros como "año" y "MeSH", y puede depender de estos. Por eso, no basta con seleccionar los métodos a emplear, sino que se requiere una secuencia lógica, en su uso, para obtener mejores resultados. Una propuesta aparece en el cuadro 4.

Cuadro 4. Propuesta de la combinación de métodos a emplear para el preprocesamiento

Transformación		Metadatos				
		Fecha	Fuente	MeSH	Afiliación	Autor
Orden de procesamiento del campo		1	2	3	4	5
1	Remplazar caracteres ajenos al juego de caracteres	√	√	√	√	√
2	Remplazar caracteres no acordes con el formato del campo	√	√	√	√	√
3	Remplazar espacios múltiples	√	√	√	√	√
4	Imputar valores faltantes	√	√	√	-	-
5	Imputar valores atípicos (*outliers*)	√	√	-	-	-
6	Eliminar palabras vacías	-	√	-	√	-
7	Realizar remplazos globales	√	√	-	√	-
8	Realizar remplazos globales múltiples mediante una lista de sinónimos	-	√	√	-	-
9	Extraer dato país_afiliacion	-	-	-	√	-
10	Extraer dato email	-	-	-	√	-
11	Desambiguar por método de Magnani[24]	-	-	-	√	-
12	Desambiguar por método de Bolikovski[22]	-	-	-	-	√
13	Eliminar valores faltantes	√	√	√	√	√
14	Eliminar valores atípicos (*outliers*)	√	√	√	√	√
15	Eliminar registros duplicados	√	√	√	√	√

Se recomienda un orden en la aplicación de los métodos por campos de un registro (en este caso algunos campos de la BD Medline). Debe tenerse en cuenta que algunos métodos requieren parámetros de entrada. Por ejemplo, imputar valores faltantes puede ser por la moda, la media, etc., un valor dado por el usuario.

Para realizar la desambiguación se recomienda:

- Emplear diferentes métodos para los campos autor y afiliación del autor por tener características diferentes. Estas diferencias consisten en que el autor está compuesto por tres campos (nombre, apellidos e iniciales), mientras que la afiliación es un solo campo, pero que puede contener los datos de país y email. Estos datos deben ser separados antes. Además, para la desambiguación del campo autor se incluyen metadatos, no así para la afiliación donde solamente es relevante el campo país, y se debe acudir a la similitud entre cadenas. Es evidente que la desambiguación de la afiliación debe acometerse primero, ya que aporta características que pueden contribuir a la desambiguación del autor.

- Emplear procedimientos de limpieza a los datos que van a ser desambiguados -en este caso autor y afiliación- así como al resto de los campos que se van a emplear como características relevantes para el método de desambiguación o sean relevantes para el estudio bibliométrico que esté realizando el usuario.

- Contar con una aplicación que sea suficientemente flexible, ya que no existe un procedimiento ideal, y por eso se requiere que el usuario (que es el experto en el dominio de los datos) pueda configurar lo que necesite.

La selección de los métodos empleados en la aplicación se basó en los siguientes criterios:

- Que sean apropiados, efectivos y eficientes para resolver el problema planteado.

- Que estén suficientemente documentados como para comprenderlos y emplearlos plenamente.

- Que existan en bibliotecas de código abierto y su complejidad de programación no sea muy grande.

- Que sean flexibles y ajustables al problema planteado.

Para la desambiguación de la afiliación del autor se propone emplear el método de *Magnani* y *Montesi*.[24] Estos autores destacan la comparación de nombres de empresas (*Company Name Matching*) como otra forma del problema de la desambiguación, aplicada a BD de patentes como Amadeus y Patstat. Como resultado se obtiene el nombre legal de la empresa. Se realiza una limpieza de los datos eliminando signos de puntuación, palabras vacías, espacios múltiples, etcétera. Posteriormente se eliminan filas duplicadas y se aplican a los nombres de empresas funciones de distancia de edición y basadas en términos o *tokens*. Mediante otra función determinan el peso que tiene cada *token* proporcional a su significación o importancia. Se emplea la técnica de poda mediante el campo "país" para reducir la complejidad, o sea, obteniendo primero el país se puede después dividir en subconjuntos y disminuir el tiempo de cálculo.

A partir de los métodos estudiados y de las problemáticas del usuario, se considera que, para el caso de la desambiguación de nombres de autores, el método más flexible y conveniente es el de *Bolikovski,*[22] que consta de tres pasos:

1. Mediante una poda, los documentos son separados en grupos mediante una función hash.

2. Para cada par de documentos de un mismo grupo se calcula su afinidad (semejanza) total, que es la suma de las afinidades atómicas para cada una de las características (atributos) a considerar. Estas, a su vez, se obtienen como resultado de una función que devuelve un valor entre -1 y 1, que representa el aporte de esa característica en la comparación y que posteriormente se multiplica por el peso correspondiente a esta. El valor 1 indica que según esa función es seguro que los atributos corresponden al mismo individuo. Un valor cero indica que esa función no puede determinar si esas características corresponden o no a una misma persona. El valor -1 indica que esos dos atributos corresponden a dos personas distintas. Algunas características aportan un alto peso cuando coinciden, por ejemplo, el email, pues demuestran que es la misma persona; otras tienen una importancia débil, por ejemplo, el nombre de la publicación. A veces ocurre que una misma característica es fuerte para la coincidencia, pero débil para la diferencia (o viceversa), por ejemplo, el email. Un elemento importante del método de este autor es que los pesos pueden ser ajustados de forma flexible; también pueden determinarse mediante una aplicación informática. El resultado de este paso es una matriz de las afinidades totales.

3. El último paso consiste en la clusterización a partir de las matrices obtenidas en el paso anterior y empleando el algoritmo "clusterización aglomerada jerárquica con enlace simple" (*Single-Linkage Hierarchical Agglomerative Clustering*, SLHAC),[23] que compara con un umbral establecido.

Una ventaja significativa en este método es su flexibilidad, ya que la función *hash*, la función de similitud y el algoritmo de clusterización pueden ser sustituidos por otros a conveniencia. Esto permitiría adecuar el método a las necesidades, experimentar diferentes variantes, nuevos algoritmos e ir creando una base teórico-práctica propia en el tema del preprocesamiento.

DISEÑO DE LA BASE DE DATOS

Estas recomendaciones y exigencias son específicas para el caso de nuestro estudio. Son las más importantes para el diseño de una aplicación informática para el preprocesamiento, lo cual no excluye que se deberá elaborar un documento más detallado y que cumpla con los requisitos de la Ingeniería de Software para diseñar, programar y explotar dicha aplicación. Es previsible que en ese proceso y con las experiencias de trabajo con la aplicación se detecten aspectos que deban ser perfeccionados, métodos de programación más eficientes, rutinas y funciones existentes que pueden emplearse y otros.

El alcance de este trabajo debe contemplarse por etapas. En la primera etapa la aplicación:

- Se limitará al preprocesamiento de datos obtenidos de la base de datos Medline en formato XML a través del ViBlioSOM Software, que posee un diseño de base de datos propio.

- Se limitará al preprocesamiento (normalización según la metodología ViBlioSOM) de los datos, sin inmiscuirse en la creación de la BD.

- Funcionará como una aplicación concebida para un único usuario y una base de datos en una red de área local. No es relevante si se programa como una aplicación web o de escritorio.

- No requerirá de medidas de seguridad contra el acceso no autorizado a la información, pero sí para prevenir la posible destrucción de la información, tales como la realización de salvas, implementación de chequeos de consistencia en las transacciones y la recuperación (*rollback*) en caso de que no puedan ser finalizadas exitosamente, derechos limitados para los usuarios y otras. De esta forma se garantizará una mayor confiabilidad en los resultados de las tareas que se realizan en la aplicación.

- Contará con una interfaz que debe ser de fácil comprensión, intuitiva y amigable, acorde con los estándares actuales, considerando que el usuario de la aplicación es un trabajador de la información, no necesariamente experto en informática.

- Deberá brindar mensajes claros ante los errores de la aplicación o del usuario y soluciones que permitan a este último continuar trabajando y no perder la información. Todos los errores deben quedar registrados.

- Deberá interactuar con el ViBlioSOM Software de forma modular, sin interferir en modo alguno con su funcionamiento actual, o sea, importará los datos desde la base de datos del mismo, los preprocesará y posteriormente los restaurará. En dependencia de los resultados que se obtengan con esta aplicación se harán sugerencias para que en el diseño de las nuevas versiones de ViBlioSOM este módulo de preprocesamiento esté totalmente integrado.

- Se programará en código abierto. Existen repositorios de algoritmos y código que pueden ser valorados para emplearlos en la programación de la aplicación. Entre estos se encuentran:

SecondString: Paquete de código abierto en Java con técnicas de comparación aproximada de cadenas.

Simmetrics: Biblioteca de código abierto en Java con técnicas de similitud.

Febrl: Biblioteca de código abierto en Python con técnicas de desambiguación.

ATLaS: Extensión al lenguaje SQL para potenciar funciones de agregado y la minería de datos.

- El SGBD a emplear será PostgreSQL v.8.4, por ser donde está soportado el ViBlioSOM y que garantiza que el mayor procesamiento sea realizado en el lado del servidor. Este posee el lenguaje PL/pgSQL, que es también de código abierto y con amplias posibilidades, como es su extensibilidad con Java, Python y otros lenguajes. Además, entre sus ventajas cuenta con facilidades para manipular arreglos, que en nuestra aplicación son necesarios para crear las matrices de similitud y procesamiento de tokens.

- Aunque no tendrá exigencias críticas en cuanto al desempeño, se tratará de disminuir lo más posible los tiempos de ejecución, emplear las potencialidades de cálculo del servidor, racionalizar la escritura en disco y tomar otras medidas en la programación (Fig. 2).

- La aplicación tendrá como objetivo principal mejorar el preprocesamiento de los datos en una metodología tomada como marco de referencia como el ViBlioSOM. Otro objetivo será obtener una plataforma de desarrollo y experimentación que permitan evaluar las potencialidades y resultados de los diversos algoritmos y métodos para el

preprocesamiento. Este trabajo permitirá desarrollar una concepción propia sobre este tema.

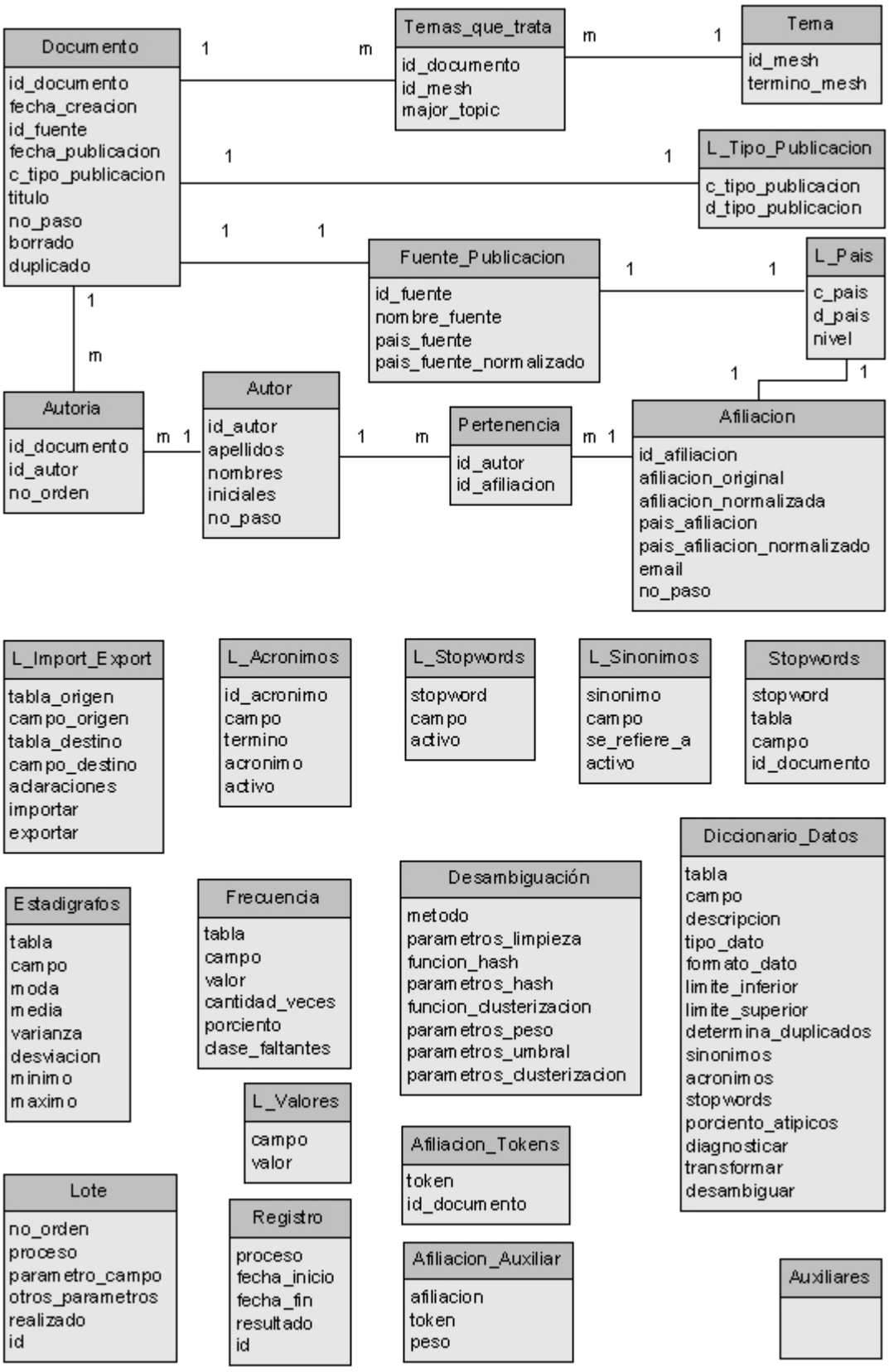

Fig. 2. Estructura de la base de datos.

CONCLUSIONES

La propuesta de un algoritmo para el preprocesamiento de los campos "autor-afiliación" con un enfoque métrico y que además pudiera ser implementado en un sistema de análisis métrico de la información, no es un problema trivial. Sobre este tema se tiene poco conocimiento acumulado a nivel internacional y en Cuba no se encontraron documentos publicados sobre el tema. Esto permitió constatar que, a pesar del aumento en el uso de las metrías, el desarrollo de aplicaciones que solucionen los problemas de preprocesamiento de datos no ha sido abordado, por lo que las técnicas para el preprocesamiento de datos en su sentido más amplio, no están siendo aprovechadas suficientemente en Cuba para fines métricos.

El estudio de la literatura sobre preprocesamiento ha permitido establecer que esta puede definirse como un proceso dirigido a las transformaciones óptimas de los datos, destinados a la obtención de un conocimiento significativo y puede estar compuesta por uno o varios métodos. Aunque el preprocesamiento está asociado con la aplicación informática de algoritmos matemáticos, también está vinculada a un proceso cognitivo en cuanto a su propósito de “descubrir un conocimiento nuevo”.

Los análisis métricos pueden mejorar sus resultados al emplear estadígrafos antes de tomar decisiones de transformación o eliminación de los datos, así como aplicar métodos para la desambiguación de los nombres de autores y afiliaciones, como solución a uno de los problemas que más deterioran la calidad de los estudios métricos.

Un enfoque viable para optimizar el preprocesamiento de datos para los estudios métricos puede ser como el que aquí se propone: modular, basado en las potencialidades del SGBD, aprovechando el código abierto existente, configurable y flexible y donde el especialista en información no pierda el control de lo que está sucediendo.

Se logró establecer un grupo de requerimientos al diseño de una aplicación informática para el preprocesamiento de los datos que puede estar conformada por un conjunto de tareas, pasos y algoritmos. Se sugiere usar no un solo algoritmo, sino una combinación de estos para llegar a desambiguar el campo afiliación y autor. Estos pueden ser seleccionados en función de los datos y las necesidades del usuario.

Para otras etapas de desarrollo de la aplicación se recomienda:

- Considerar la necesidad de procesar otras bases de datos bibliográficas.

- Considerar el procesamiento paralelo (por ejemplo, Google's Map Reduce).

- Considerar la posibilidad de trabajo en grupo.

- Extender el preprocesamiento a otros campos (por ejemplo, títulos, resúmenes, etcétera).

Contribución de los autores

Ramón Albo determinó el tema, diseñó el estudio, analizó los datos, propuso la solución, redactó las versiones; *María Victoria Guzmán* determinó el tema, diseñó el estudio, redactó y revisó las versiones; *Romel Calero, Ivet Álvarez y Jesús Bouza*

Revisaron la solución propuesta. Todos los autores revisaron la redacción del manuscrito y aprueban la versión finalmente remitida. El artículo no ha sido publicado previamente ni está siendo considerado actualmente por otra publicación.

Conflicto de intereses

Los autores declaran que no existe conflicto de intereses en el presente artículo.

REFERENCIAS BIBLIOGRÁFICAS

1. Omelianovsky ME. Los métodos de la matemática contemporánea y la matematización del saber. En: Omelianovsky ME, editor. La dialéctica y los métodos científicos generales de investigación (Tomo I). La Habana: Ciencias Sociales; 1981. p. 179-243.

2. Guzmán MV. Vibliosom: Metodología para la visualización de información métrica con mapas auto-organizados [Tesis Doctoral]. La Habana: Universidad de La Habana; 2009.

3. Kimball R. Dealing with dirty data. DBMS. 1996;9(10):55-60.

4. Müller H, Freytag JC. Problems, methods and challenges in comprehensive data cleansing. Berlin: Professoren des Inst. Für Informatik; 2005:23.

5. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE DEBU. 2001;23(4):3-13.

6. Ontalba-Ruipérez J. Normalización de campos en bibliometría: acciones de la Fecyt. Prof Inf. 2007;16(4):381-3.

7. Spinak E. Errores ortográficos en el ingreso en bases de datos. Rev Esp Doc Cient. 1995;18(3):307-19.

8. Lardy J, Herzhaft L. Bibliometric treatments according to bibliographic errors and data heterogeneity: the end-user point of view. En: 16th international online information meeting. London. Oxford: Learned Information; 1992. p. 547-56.

9. Anguita A, Pérez D, Crespo J, Maojo VM. Automatic generation of integration and preprocessing ontologies for biomedical sources in a distributed scenario. En: Proceedings of 21st International Symposium on Computer-Based Medical Systems (CBMS-2008). Washington DC: IEEE Computer Society; 2008. p. 336-41.

10. Zimei S. KDDML: Estensione alla fase di Preprocessing [Tesis de Grado]. Pisa: Universidad de Pisa; 2004.

11. Bordons M, Costas R. Algoritmos para solventar la falta de normalización de nombres de autor en los estudios bibliométricos. Investig Bibliotecol. 2007;21(42):13-32.

12. Smalheiser NR, Torvik VI. Author name disambiguation. Annu Rev Inform Sci. 2009;43(1):1-43.

13. Han H, Giles CL, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. En: Proceedings of Joint Conference on Digital Libraries (JCDL 2004). Tucson, EE.UU.: ACM; 2004. p. 296-305.

14. Han H, Zha H, Giles CL. A model-based k-means algorithm for name disambiguation. En: Proceedings of 2nd International Semantic Web Conference (ISWC-03) Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data. Sanibel Island FL, Alemania: Springer; 2003.

15. Giles CL, Han H, Zha H. Name disambiguation in author citations using a K-way spectral clustering method. En: Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries (JCDL '05). Denver, New York: ACM; 2005. p. 334-43.

16. Huang J, Ertekin S, Giles CL. Efficient name disambiguation for large-scale databases. En: 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin: Humboldt-Universität zu Berlin; 2006. p. 536–44.

17. Treeratpituk P, Giles CL. Disambiguating authors in academic publications using random forests. En: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL-09). New York: ACM; 2009. p. 39-48.

18. Ferreira AA, Veloso A, Gonçalves MA, Laender AHF. Effective self-training author name disambiguation in scholarly digital libraries. En: Proceedings of the 10th annual joint conference on Digital libraries (JCDL'10). Gold Coast, New York: ACM; 2010. p. 39-48.

19. Torvik VI, Smalheiser NR. Author name disambiguation in Medline. ACM Trans Knowl Discov Data. 2009;3(3):1-29.

20. Pino R, Cubiles MD, Caballero E. A comparison of probabilistic record linkage techniques in the Institute of Statistics of Andalusia (ISI' 2011). En: 58th World Statistics Congress of the International Statistical Institute. Dublin: ISI; 2011.

21. Jijkoun V, Khalid MA, Marx M, Rijke M. Named entity normalization in user generated content. En: Proceedings of the second workshop on Analytics for noisy unstructured text data (AND'08). Singapore, New York: ACM; 2008. p. 23-30.

22. Bolikowski L, Dendek PJ. Towards a flexible author name disambiguation framework. En: Sojka P, Bouche T, editores. Towards a digital mathematics library. Brno: Masaryk University Press; 2011. p. 27-37.

23. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. New York: Cambridge University Press; 2008:482.

24. Magnani M, Montesi D. A study on company name matching for database integration. Bologna: University of Bologna. 2007. Technical Report: UBLCS-07-15.

25. Ferreira AA, Laender AHF, Gonçalves MA, Cota RG, Santos RLT, Silva AJC. Keeping a digital library clean: new solutions to old problems. En: Eighth ACM symposium on document engineering (DocEng '08); 2008 16-19 Sep, Sao Paolo. New York: ACM; 2008. p. 257-62.

26. Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for Medline records: A model for author name disambiguation. J Am Soc Inf Sci Technol. 2004;56(2):140-58.

27. Costas R. Análisis bibliométrico de la actividad científica de los investigadores del CSIC en tres áreas: Biología y Biomedicina, Ciencia de Materiales y Recursos Naturales. Una aproximación metodológica a nivel micro (Web of Science, 1994-2004) [Tesis Doctoral]. Madrid: Universidad Carlos III; 2008.

28. Ordonez C. Data set preprocessing and transformation in a database system. Intell Data Anal. 2011;15(4):613-31.

Recibido: 15 de noviembre de 2017.
Aprobado: 16 de noviembre de 2017.

Ramón Orlando Albo Hernández. Instituto "Finlay". La Habana, Cuba. Correo electrónico: ralbo@finlay.edu.cu