ESTRUCTURACIÓN Y CLASIFICACIÓN AUTOMÁTICA DE INFORMACIÓN: APLICACIÓN A UNA COLECCIÓN DE TEXTOS MEDICOS*

Jorge Morato**, José Antonio Moreiro***, Juan Llorens**** y Manuel Velasco*****

RESUMEN

Se describe una herramienta que mediante una aproximación multidimensional permite la estructuración y clasificación de textos. El fin que se persigue es el estudio de las distintas secciones del documento. En el desarrollo del módulo se emplearon algoritmos de filtrado (N-grams) y de clasificación (K-means y Chen). La estructuración de los documentos se realizó mediante marcadores lingüísticos, tipográficos y herramientas estadísticas. Para la evaluación del método se recopilaron de Medline documentos médicos a texto completo y se incorporó una herramienta de comparación, el MeSH. Mediante un análisis estadístico y comparativo, se ha comprobado la necesidad y validez de este tipo de aproximaciones. Por último, se propone la integración del método en un módulo que optimice la asignación de pesos en el diseño de herramientas de clasificación y recuperación documental.

PALABRAS CLAVE: Lingüística estructural, Discurso científico, Lingüística del texto, Documentación automatizada, Documentación científica, Análisis documental, Medicina, Análisis de clusters, Clasificación estadística.

MORATO, J. ***et al. Estructuración y clasificación automática de información: aplicación a una colección de textos médicos. En: Revista Interamericana de Bibliotecología. Vol. 24, No. 1 (ene.-jun., 2001); p. 117-136.***

* El presente estudio ha sido financiado por la Consejería de Educación de la Comunidad Autónoma de Madrid, dentro del proyecto titulado «Aplicación de técnicas informétricas a la construcción automática de tesauros». Artículo recibido en abril, revisado y modificado en mayo de 2000, y aceptado en marzo de 2001.

** Doctor en documentación y Profesor Ayudante del Departamento de Informática de la Universidad Carlos III de Madrid, España. E-mail: jmorato@bib.uc3m.es.

*** Profesor e investigador del Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid, España. E-mail: jamore@bib.uc3m.es.

**** Profesor Titular del Departamento de Informática de la Universidad Carlos III de Madrid, España. E-mail: llorens@inf.uc3m.es.

***** Profesor Titular Interno del Departamento de Informática de la Universidad Carlos III de Madrid, España. E-mail: velasco@id.uc3m.es.

ABSTRACT

In this study, an automatic linguistic tool is described. The goal of this tool is to analyse the behaviour of different text structures when they are faced to filtering and classification algorithms. The model structures the text by means of a multidimensional approach. On one hand, text has been divided in sections by means of typographic constraints, semantic labels, and location rules. On the other, vocabulary related to different text structures has been implemented in the database. The text analysis algorithms that have been implemented were the n-grams filter, and the classification algorithms k-means and Chen co-wording. The module has been tested using a collection of full-text documents from Medline. The evaluation of the methodology was accomplished by comparing with the MeSH vocabulary and a statistical analysis. This study had shown some advantages of the context approach. Finally, it is proposed to improve the success of information retrieval and classification algorithms with structuring techniques.

KEY WORDS: Automatic-text-structuring, Discourse-model, Computational-linguistics, Text-analysis-methods, Automatic-classification, Cluster-analysis, Medicine, Linguistics.

MORATO, J. *et al. Automatic Structuring and classification of information: an Applied Case of a Collection of Medical Texts.* In: *Revista Interamericana de Bibliotecología. Vol. 24, No. 1 (jan.- jun., 2001); p. 117-136.*

INTRODUCCIÓN

Entre los primeros trabajos que relacionan la lingüística con la recuperación documental, destacan los estudios de Garfield (1992) en la década de los cincuenta. Aunque desde un primer momento se detecta un numeroso grupo de estudios que tratan de aprovechar las distintas características del discurso en la indización documental (Slype, 1991; Wormell, 1985), no es hasta varias décadas más tarde cuando surgen los primeros ensayos encaminados a la realización de un análisis automático del discurso (Pêcheux, 1978).

A principios de los 70, adquiere un gran auge la lingüística del texto, que propugna un análisis del texto que no esté limitado a las proposiciones de manera individual (Dijk, 1996). Esta corriente lingüística sugería un análisis integral del texto en oposición a la limitación de la frase empleada hasta ese momento. Se puede observar la necesidad de este enfoque cuando se analiza la forma en que se redacta un texto. Es evidente que al escribir un artículo se necesita una macrosemántica que nos informe de una forma más eficaz que un conjunto de frases aisladas. Mediante la lingüística del texto, las anáforas que suponen algunos pronombres, artículos, conjunciones, adverbios y fenómenos como la presunción o la coherencia, son desambiguadas al considerar el conjunto del texto. Algunos trabajos, sobre todo dentro de la traducción automática y la documentación,

han surgido a raíz de esta escuela. Ejemplos que muestran este proceso podrían ser los trabajos que estudian la influencia del número de pronombres en la aplicación de herramientas informáticas (Mitkov, 1997; Llorens, 1998). Simultáneamente, en el campo de traducción, pero con un razonamiento similar, han surgido un gran número de estudios en los que se propone realizar el análisis de textos mediante un estudio previo de la estructura, tipología y/o del registro del texto (Abaitua, 1997).

Siguiendo esta idea, diversos autores (Dijk, 1996; Moreiro, 1993 , Salton, 1997) han sugerido la localización de las macroestructuras y superestructuras para la indización y la generación de resúmenes. La forma en que las distintas secciones del documento se reflejan en el texto se debe a las superestructuras. Estas juegan un papel fundamental en la producción y comprensión del discurso. Una breve revisión de la literatura sobre el tema nos muestra el gran esfuerzo empleado en esta área. Por ejemplo, la aplicación orientada a la localización de distintas macroestructuras parciales para la delimitación de la semántica global en la recuperación automática ha sido estudiado por Hearst (1993). Acorde con esta línea de pensamiento, Leydesdorff (1997), utilizó un análisis discriminante para diferenciar entre las distintas secciones de un texto, mediante la tipología textual propia de cada sección. Con una perspectiva diferente, Seglen (1996) ha utilizado diferentes medidas como la densidad de información en epígrafes y tablas, y su relación con el factor de impacto.

La aplicación de estas técnicas a los artículos científicos parece especialmente indicada dada su estructura. En concreto, el discurso científico está estructurado según ciertas pautas de organización retórica, aunque con una cierta libertad individual de variación estilística (Weissberg, 1990). Los documentos científicos, por su condición argumentativa, necesitan una serie de categorías, como es el caso de las premisas y de las conclusiones. La macroestructura del articulo de investigación, conocida como IMRD, fue propuesta por Bruce en el año 1983. El nombre de este esquema se corresponde con las secciones que Bruce considera principales: Introducción, Métodos, Resultados y Discusión. Aunque otros modelos han sido propuestos posteriormente, se han rechazado por no tener una relación tan estrecha con el razonamiento inductivo, éste es el motivo por el cual en el presente trabajo se utilizará el esquema IMRD en la estructuración del documento. En cualquier caso, si se ha demostrado en diversos estudios que existen ciertas dife-

rencias en la distribución de las características lingüísticas y retóricas a lo largo de este esquema, por ejemplo, el 90% de los tiempos verbales en presente están entre la introducción y la discusión (Heslot, 1982).

Entre los experimentos para analizar la estructura documental, conjuntamente con el género documental, destacan los estudios realizados por Berri (1996). Este autor ha propuesto un método para estructurar los artículos científicos a través de la exploración contextual, por medio de las posibilidades que ofrecen los caracteres con formatos especiales y mediante la localización de un conjunto de etiquetas argumentativas y descriptivas en el texto; aunque, quizás el autor que haya tenido más impacto en el análisis de la terminología y estructuras científicas sea Swales (1990). Swales, en una larga serie de trabajos ha propuesto la localización del vocabulario específico y la delimitación de los micromovimientos propios de cada superestructura en el discurso científico. Otros estudios se han centrado en la identificación de las unidades estructurales por medio de etiquetas funcionales centrado en campos más específicos, por ejemplo, Nwogu (1997) en el campo médico. También, Gilyarevsky (1997), estableció un método para diferenciar entre distintos tipos de artículos científico-técnicos de agricultura a través de recuentos de términos en los títulos de cada documento. Con una aproximación algo más compleja, Loose (1996) ha realizado un análisis sobre la variación morfológica entre las distintas estructuras de los diferentes géneros. Por último, se han realizado experimentos que utilizan una serie más amplia de parámetros junto a un análisis estadístico para analizar las diferentes estructuras (Morato, 1999).

En las siguientes secciones se expondrá cuáles son los objetivos que se persiguen en este estudio. Una vez expuesto el marco teórico y los fines que se persiguen, se describirá cuáles son las herramientas cuyo comportamiento se pretende estudiar a la luz de las distintas estructuras documentales. En esta sección se describen detalladamente los algoritmos de filtrado y clasificación empleados. A continuación, se propondrá un desarrollo experimental que incluya el modelo propuesto en los apartados anteriores. Finalmente, se analizarán los datos obtenidos en el experimento.

1. OBJETIVOS

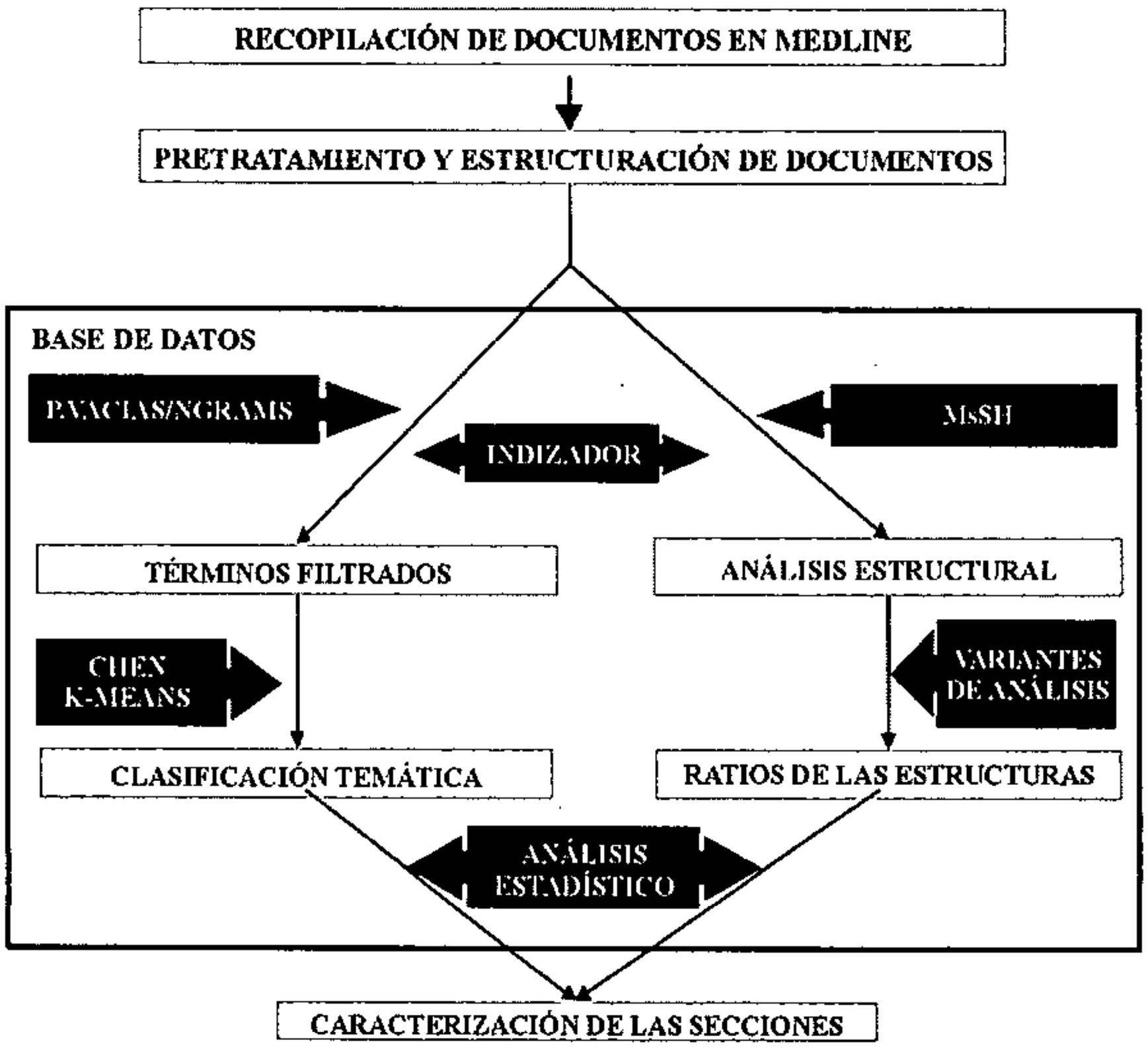
El objetivo principal que se pretende alcanzar en este estudio, es comprobar si existen variaciones entre las superestructuras del texto que puedan tener interés desde el punto de vista de las herramientas documentales. Para conseguir este

objetivo se ha recurrido a una doble vía. Por un lado se han identificado una serie de variables que nos permitan afirmar o negar estadísticamente el hecho; y por otro, se ha realizado un análisis mediante una comparación entre un vocabulario plenamente aceptado por la comunidad científica, el MeSH y la clasificación automática de las distintas secciones del documento. Para este fin, se han seleccionado dos algoritmos complementarios de clasificación automática.

2. METODOLOGÍA

Como se ha comentado en el apartado anterior, se utilizó una metodología que utiliza una doble vía paramétrica y clasificatoria, y que confronta los datos de salida mediante un análisis multivariante (Gráfico 1). En los siguientes apartados se describirán los elementos principales del método, y la toma de datos que se ha realizado para validar el método.

Gráfico 1.
Metodología de extracción de Información



2.1 Creación de las colecciones

Como se aprecia en el gráfico uno, el método comienza con la recopilación de documentos para la colección. Para crear la colección se seleccionaron un total de 300 documentos en formato electrónico. La búsqueda se centró en artículos de investigación en biomedicina. La elección del tema biomédico fue motivado principalmente, por dos razones. La primera fue la gran normalización presente en el vocabulario, en las superestructuras y en los micromovimientos del discurso médico (Nwogu, 1997); en segundo lugar, la gran accesibilidad de los documentos, en todos los registros estudiados, consecuencia natural de la existencia de un gran número de documentos en formato electrónico.

Los documentos procedían de la base de datos MEDLINE, indizada mediante el MeSH (Lowe, 1994), contando con 10 millones de registros y, siendo así, la mayor y más utilizada base de datos en medicina. En Medline, se seleccionaron documentos de investigación de las siguientes publicaciones: New England Journal of Medicine, British Medical Journal, Lancet, Journal of Clinical Investigation y Aids Care. La selección de estas publicaciones estuvo motivada por tener la mayoría de ellas un elevado factor de impacto dentro de sus grupos respectivos, del JCR (SCI, 1997). Hay que hacer notar que Aids Care no figura en este catálogo y que la publicación Journal of Clinical Investigation se encuentra en un grupo diferente al del resto (está en "Medicine Research Experimental", mientras que el resto está en "Medicine General and Internal").

Se realizó una búsqueda por documentos del año 1996 sobre SIDA, Hepatitis y Scrapie. A su vez, se hizo una búsqueda paralela con el fin de comparar posteriormente los resultados en el Journal of Clinical Investigation sobre bioquímica de proteínas en medicina clínica. El objetivo fue que, en la comparación, hubiera cierto grado de relación en los temas, aún en campos distintos.

El motivo de elegir estas tres enfermedades era el gran volumen de artículos generados y, por otra parte, la comparación del componente temporal en los tres campos, es decir entre una enfermedad con larga trayectoria en investigación médica (Hepatitis), otra con una historia más breve (Creufeltz-Jacob) y, por último, una en pleno auge, el SIDA/HIV.

2.2 Pretratamiento de los documentos

Según se desciende en el esquema mostrado en el gráfico uno, se experimenta la necesidad de pretratar los documentos previamente a su procesado por el mó-

dulo. Los caracteres extraños, subíndices, erratas, etiquetas de determinados interfaces, etc., se suprimieron con la finalidad de disminuir errores a causa de factores ajenos al experimento. Las características de los documentos, como: temática, tipología, autores, instituciones, fuente y palabras clave, se insertaron manualmente en las propiedades de Microsoft Word del documento para así, poder contar con información adicional que pueda mejorar la clasificación de la información documental.

Por último, con el fin de observar el comportamiento de las diferentes estructuras, toda la colección de documentos se estructuró en las diferentes secciones (resumen, título, conclusiones, métodos, resultados, bibliografía, fuente y leyendas).

2.3 Base de datos

Como se puede observar en el gráfico uno, a partir de esta etapa todos los resultados se almacenan en una base de datos relacional. De esta manera se permitirá no sólo una mayor flexibilidad en la creación de asociaciones entre los datos, si no la también la incorporación de diferentes recursos que posibiliten la evaluación del módulo de análisis.

Para posibilitar la evaluación de la clasificación automática mediante un vocabulario universalmente aceptado en el campo médico, se incorporó el MeSH a la base de datos. El MeSH es el vocabulario controlado más extendido en el ámbito médico. Mediante esta herramienta es posible tener un referente del comportamiento en la indización de las distintas subcolecciones. En concreto, se utilizaron dos productos: el Medical Subject Headings Tree Structures (MeSH-Tree Structures), con la estructura jerárquica, y el Medical Subject Headings-Annotated Alphabetic List. Este último, consiste en una versión ampliada del MeSH, con 18.000 descriptores y 100.000 términos más entre sinónimos y diferentes grafías de cada uno de ellos; además, existen códigos de términos relacionados, definiciones en cada descriptor indicando lo que abarca y cambios en su historia, cuándo se incorporó y si sustituyó a otros. También, contiene anexos geográficos y de compuestos químicos.

Varios vocabularios específicos se añadieron a la base de datos, como el listado de palabras vacías utilizado en el programa SMART. En algunos casos, como con los pronombres o los verbos, se añadió la categoría gramatical. Por otra parte, se

ha agregado un listado de términos relacionados con el registro científico-técnico, tomado de Weissberg (1990) y Swales (1990).

A partir de este punto se sigue una doble vía: el análisis de indicadores obtenidos a partir de las estructuras de los textos y, por otro lado, la clasificación automática de la información contenida en los mismos. El objetivo será complementar los resultados mediante la combinación y análisis estadístico de ambas vías.

2.4 Análisis de las estructuras de los textos

2.4.1 Reconocimiento de estructuras

Para estructurar el texto se utilizó una doble aproximación:

- El primero, similar al ya nombrado de Berri (1996), consistente en la localización de secuencias significativas desde el punto de vista tipográfico y semántico. La aproximación tipográfica se realizó a través de la localización de caracteres en negrita o/y en mayúsculas, párrafos consistentes en una sola frase carente del punto final, etc. El segundo punto, el semántico, se llevo a cabo mediante un vocabulario que tuviera el mayor número de variantes posibles de los términos presentes en las secciones de los subtítulos.
- Otro método utilizado consistió en la localización de macroestructuras parciales en documentos, por medio de agregados representativos del vocabulario de cada sección. Los agregados consisten en grupos de palabras simples o compuestas que suelen presentarse asociadas con una frecuencia elevada en determinadas condiciones, como puedan ser ciertas secciones de los documentos, determinados grupos de documentos o determinados campos del saber.

  Estos agregados se construyeron de dos maneras diferentes. Por un lado, se recurrió a la propuesta de Leydesdorff (1997) de localizar terminología cuya semántica estuviera relacionada con aspectos tales como la observación, la metodología y la teoría. Por otro lado términos y frases relacionados con las diferentes secciones del texto se recopilaron de los trabajos de Nwogu (1997), Estevez (1996), Skelton (1994) y Swales (1990).

2.4.2 Analizador de citas

Para estudiar alguna de las estructuras retóricas utilizadas, se creó un analizador de referencias. El sistema trabaja de un modo similar al desarrollado en ACI (Lawrence, 1999). La selección de párrafos del capítulo de bibliografía y su com-

binación con una serie de argumentos, agregados a la base de datos, posibilitan localizar el autor del documento referenciado, el título, el año y la publicación. A continuación, se ubican en el texto los distintos sistemas de citación (Swales, 1990), como por ejemplo, citas no integrales del tipo: (Sánchez, 91), [SANC 91], [1] o (Sánchez et al., 1991), o integrales, como: Sánchez (1990). El sistema localiza en el texto estas citas y las compara con la sección de bibliografía, validando la ocurrencia de la cita en el texto. Estas ocurrencias se emplearán luego en la generación de indicadores.

Para el analizador de referencias, se han añadido a la base de datos todos los esquemas que cumplían las referencias y citas en el texto en las publicaciones seleccionadas. El esquema consiste en la especificación de los separadores y el orden en que aparecen los distintos campos de la referencia (autor, título, año y publicación). Un proceso similar se sigue para el marcador en el texto al que hace referencia la cita bibliográfica.

Como ya se ha comentado, el analizador incorpora automáticamente todos los autores de las referencias a la base de datos, vinculándolas con el título de la referencia, año y publicación. El Journal Citation Report correspondiente a ese año fue tabulado en la base de datos para el posterior análisis de las revistas de la bibliografía.

2.4.3 Ratios bibliométricas y lingüísticas

Se seleccionaron una serie de variables bibliométricas (Egghe, 1990), para analizar las relaciones con las distintas estructuras. Es necesario descender a distintos niveles de análisis, para valorar estos índices en su justa medida. Tanto las ratios bibliométricas como lingüísticas se calcularon en el ámbito del conjunto de la colección, del documento y en el de las secciones.

En la bibliografía, se calcularon, la frecuencia y el número de referencias; la obsolescencia de las citas y el porcentaje de autocitas.

En cuanto a indicadores procedentes del proceso de indización se ha calculado el número de descriptores procedentes de información gráfica, procedentes de las leyendas de las tablas y los procedentes de la primera frase de cada párrafo; por otro lado, se ha calculado el porcentaje de descriptores procedentes de cada sección del documento.

Por último, en lo concerniente a los documentos tratados en parejas, se calculó el número de cocitaciones en cada subcolección.

2.5 Clasificación automática de la información

2.5.1 Filtrado mediante N-grams

El filtrador N-Gram se empleó con el doble objetivo de aumentar la eficiencia del sistema, al tiempo que se disminuía el ruido (Cohen, 1995). Este algoritmo realiza un filtrado estadístico por medio de la comparación con una serie de cadenas de caracteres. El método es independiente del lenguaje, ya que sólo necesita trabajar con recuentos de las apariciones en el documento y su comparación con un texto más genérico del mismo idioma (background). El esquema desarrollado fue el siguiente (Velasco, 1998):

El método se basa en representar el texto mediante secuencias de cadenas de caracteres de un tamaño fijo. Esta secuencia es la que da nombre al método n-grams, donde n representa el número de caracteres consecutivos cuya frecuencia se va a calcular. El método es simple, el proceso empieza con el primer carácter del texto. El sistema consiste en ir guardando los conjuntos de caracteres, tanto de texto como especiales, que aparecen en una ventana cuya anchura se corresponde a un determinado valor de n. La ventana se va desplazando un único carácter antes de cada lectura. Con este sistema conseguimos todos los conjuntos de n caracteres que se pueden encontrar en el texto. Posteriormente, en una segunda lectura se calcula la frecuencia de cada cadena de caracteres. Las palabras, compuestas o no, que contienen estos caracteres en una proporción mayor a la esperada en ese idioma, son seleccionados como descriptores.

La estimación de cuales son los conjuntos de caracteres que presentan una frecuencia significativamente mayor en nuestro corpus que en el lenguaje usual, se realiza mediante un background. Un background es una colección de documentos generales del idioma, que tienen un escaso nivel de solapamiento con el dominio que vamos a estudiar. Los documentos de background están usualmente integrados por novelas, textos de divulgación o conversaciones transcritas. Las ventajas que tiene la utilización de un background respecto a un listado de palabras vacías es su mayor flexibilidad y adaptación a determinados corpus.

En este trabajo se han tomado cadenas de cinco grams, para poder tener un carácter central en el n-gram. Respecto a la manera en que los documentos eran

procesados, los mejores resultados se obtuvieron cuando se analizaban los documentos en lotes de cincuenta documentos solapados cada veinticinco. El background fue elegido tras una serie de pruebas, y consistió en la unión de novela histórica junto con diferentes artículos de geología. Estos campos de conocimiento fueron seleccionados dado su escaso solapamiento con la disciplina médica.

2.5.2 Análisis léxico e indización

Para continuar con lo propuesto en el gráfico uno, se alcanza un estado previo a la clasificación que consiste en relacionar los documentos con los descriptores obtenidos en etapas previas. Previo a este paso es necesario un análisis léxico. El análisis léxico (Frakes, 1992) se realiza con el propósito de transformar los términos a una forma canónica, todas las familias de palabras procedentes de los distintos textos y los términos obtenidos en el n-gram. El módulo de normalización trabaja localizando palabras que no estén identificadas como vacías. A continuación coteja su terminación con una tabla. Cuando la terminación coincide con un registro de la tabla, aquella es sustituida por una terminación normalizada, obteniendo de esta manera un término candidato. Los términos resultantes se comparan con un vocabulario controlado del mismo discurso que los documentos con los que se está trabajando, validando, o descartando en su caso, el candidato normalizado.

A continuación se realiza la indización. Es decir, se guarda en la base de datos la localización y las ocurrencias de todos los términos, para su posterior utilización en la etapa de clasificación y como componentes de indicadores bibliométricos. La localización incluye además del documento, la sección en el mismo donde se ha encontrado.

El Medical Subject Headings, dada su reputación en indización de literatura médica, se tomó como herramienta de referencia en todo el proceso, para lo cual, se realizó una indización de las subcolecciones mediante el Medical Subject Headings.

2.5.3 Generación de agregados

En esta etapa se realiza la construcción de agregados para obtener un cuasi-tesauro. Por cuasi-tesauro se entiende un conjunto de términos relacionados semánticamente. Comúnmente, la generación de agregados se ha utilizado para identificar agrupaciones de objetos que mantienen un elevado número de caracte-

rísticas comunes. Estos objetos pueden ser documentos semejantes, usuarios similares, bibliografía coincidente o, como es nuestro caso, semejanzas entre agregados de términos formados en diferentes estructuras del documento, y su comparación estadística con otros agregados que denominaremos de referencia. La identificación automática de los agregados que integran el cuasi-tesauro, se logra mediante el análisis de las palabras presentes en los documentos del corpus.

Se seleccionaron dos algoritmos de clasificación: Chen y K-means. Estos clasificadores son los más extendidos en la generación de agregados de textos. El motivo de elegir esta doble vía está fundada en los trabajos llevados a cabo por Velasco (1998) en los que se demuestra como estos métodos, lejos de duplicar la información, la complementan.

- CHEN. El método de Chen (1992) trabaja con las concurrencias de los términos, para así generar para cada par de términos, una medida del grado de relación. Por concurrencia terminológica se entiende dos términos que aparecen en la misma localización, y que nos sirven para realizar una estimación que indicará hasta que punto estos dos términos están relacionados. De manera que una alta frecuencia de concurrencia de los dos términos, o lo que es igual, un alto grado de relación, indicará que la aparición de un término implica, casi forzosamente, la aparición en el mismo documento del otro término. Por el contrario, una frecuencia muy baja de concurrencia, implicará que los dos términos en ese dominio son prácticamente autoexcluyentes.

  Los pasos que se siguen en la aplicación de este método son los siguientes: primero, se toman los términos generados en el análisis léxico. Posteriormente, se procede a realizar el análisis de concurrencias para todos los documentos de la colección. Es decir de todos los pares de términos posibles en nuestro vocabulario. Con el propósito de establecer un peso a las relaciones asociativas que existen entre los descriptores tomados dos a dos, se calcula un peso que será comparado con un umbral de significación. Los términos más precoordinados y más específicos, alcanzan mayores puntuaciones gracias al método de asignación de pesos (Chen, 1992).

  Concretamente, la asignación de pesos está basada en la frecuencia documental inversa y la frecuencia de aparición del descriptor en el documento. La frecuencia documental inversa se utiliza en la literatura para identificar aquellos descriptores que por aparecer en la mayoría de los documentos de la colección, tienen un poder discriminatorio muy bajo, es decir, la utilización del

descriptor en determinada consulta nos devolvería prácticamente la totalidad de los documentos existentes.

- KMEANS. Se eligió este algoritmo por dos motivos, el primero su popularidad, lo cual hace que exista una abundante literatura experimental, y el segundo es que pese a su sencillez, este método ha conseguido en las pruebas previas mejores resultados que métodos más complejos. Existen muchas variantes del algoritmo, la utilizada ha sido la propuesta por Lelu (1993). El algoritmo actúa mediante centros móviles, es decir, el centro de cada agregado es recalculado en cada nueva entrada de datos. El método de agregación empleado presenta el valor añadido de la obtención de jerarquías (Velasco, 1998). Los pasos seguidos para trabajar con este clasificador es el siguiente:

Se parte de las ocurrencias documentales de cada término obtenido en el filtrado, este algoritmo comienza por un número k de centros temporales procedentes de un número de agregados previamente indicado. A medida que se procesan los casos siguientes, se van actualizando reiterativamente los centros. Un caso puede sustituir a un centro si la distancia más pequeña del caso al centro, es mayor que la distancia entre los dos centros más próximos. De esta manera, se sustituye sucesivamente los centros que estén más próximos al caso. El resultado final, es que todos los casos se agrupan en el agregado con el centro más próximo. Las dificultades que supone este método, tanto en la estimación a priori del número de agregados, como por la tendencia a crear árboles mal balanceados a causa de la selección inicial de centros, ha sido estudiada por Velasco (1998).

3. RESULTADOS

3.1 Comparación n-grams y MeSH

Para visualizar el comportamiento del módulo de filtrado y compararlo con un producto que nos pudiera ofrecer una visión objetiva, se comparó el resultado con el MeSH. El vocabulario del MeSH tiene una trayectoria de más de treinta años recopilando la terminología médica, habiéndose convertido en este periodo, en la herramienta con más prestigio mundial en esta disciplina. Prácticamente, un 30% de los 1800 descriptores filtrados por el n-grams, fueron coincidentes con el MeSH, si bien se observaron un mayor número de coincidencias en algunas secciones como en la discusión y el resumen (Tabla 1).

Tabla 1.
Comparación del N-grams y el MeSH

	TÉRMINOS TOTALES NGRAMS	PORCENTAJES ENTRE MESH Y NGRAMS	
		DESCRIPTORES IGUALES	DESCRIPTORES Y SINÓNIMOS IGUALES
ESTRUCTURA			
RESUMEN	381	22	28
MÉTODOS	418	19	24
DISCUSIÓN	483	22	26

Cuando se consideraron también los sinónimos, diferentes temáticas mostraban diferente comportamiento según la sección considerada. Así en el caso del SIDA/ HIV, se comprobó que existía una mayor coincidencia en el resumen y las referencias. Sin embargo, al considerar la sección del resumen se conseguían puntuaciones mayores.

Al observar la temática dentro del MeSH, se puede percibir que el epígrafe 'técnicas de investigación' tiene, lógicamente, más prevalencia en la sección de metodología. Sin embargo, resulta más curioso el hecho de que las primeras posiciones temáticas en el resumen y en la discusión sean frecuentemente coincidentes. La bibliografía, en el dominio de medicina interna, hace frecuentemente mención a datos geográficos, seguramente por la relación de los títulos de los artículos con un carácter pandémico de las enfermedades.

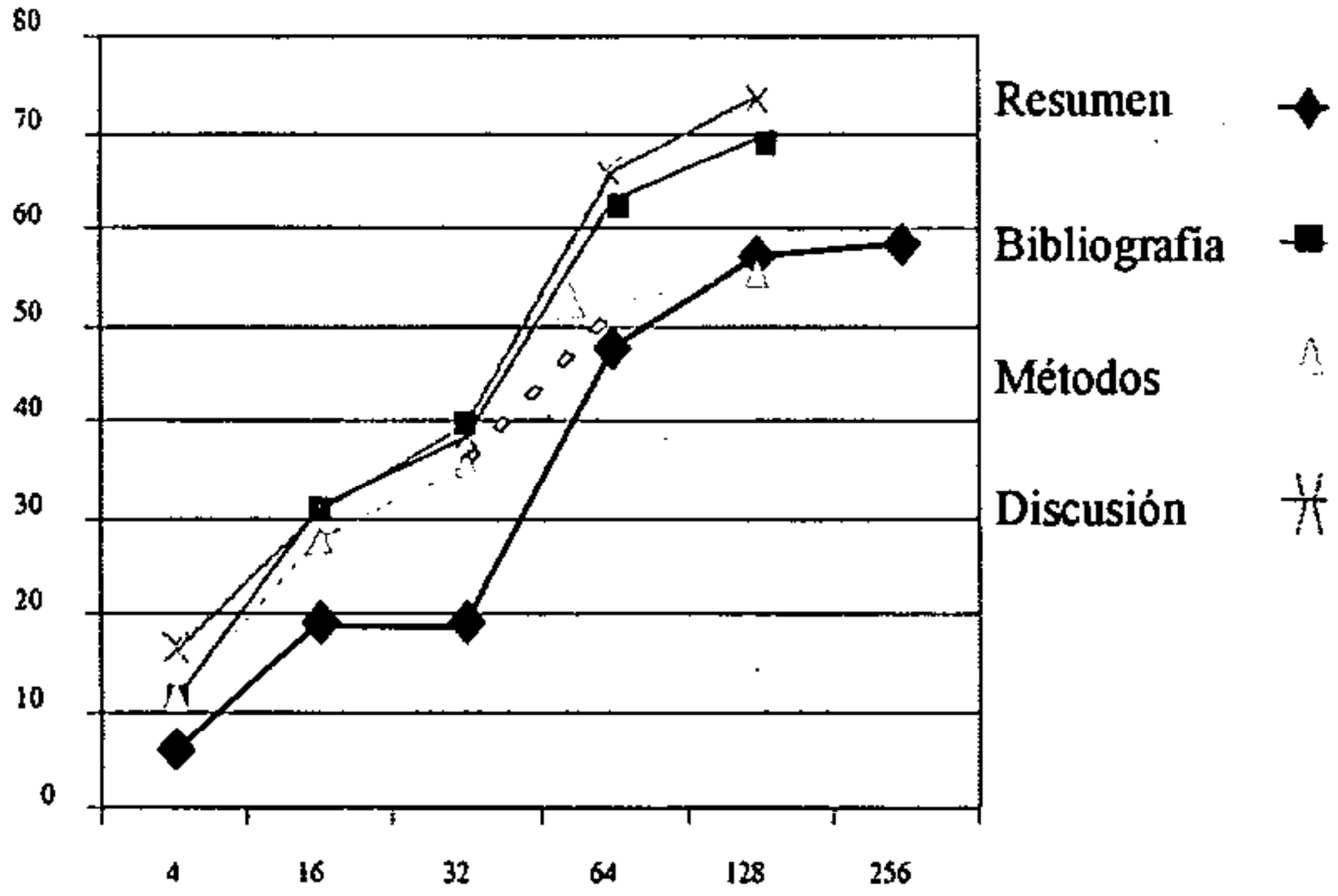
3.2 Comparación K-means y MeSH

Para comprobar el grado de concordancia, se realizó una comparación entre los términos agrupados en el MeSH y en el K-means. Estos agrupamientos incorrectos se pueden deber a dos fenómenos: por un lado, a un sobreagrupamiento, es decir, que el número de clases en el K-means sea insuficiente, agrupándose términos poco relacionados en un determinado nivel de especificidad. Por otro lado, tendríamos un infraagrupamiento, es decir un número de clases demasiado elevado, en el cual términos que deberían estar agrupados en determinado nivel de especificidad no lo están.

El comportamiento, conforme aumenta el número de clases en el K-means, fue comparado con el número de términos incorrectamente relacionados según el MeSH (gráfico 2). Como se puede observar, no todas las estructuras muestran la

misma dinámica. Coincidiendo con Losee (1996), algunas secciones como el resumen tienen un mejor comportamiento en la clasificación.

Gráfico 2.
Comparación K-means y MeSH. Número de términos relacionados en el MeSH y no en K-means, según aumenta el número de clases en K-means, según la localización en el documento.



3.3 Comparación Chen y MeSH

La comparación con el algoritmo de Chen, resulta en muchos puntos coincidente con la suministrada en los dos apartados anteriores (Tabla 2). En lo referente al tamaño de las secciones documentales, se observa que los resultados mejoran sensiblemente cuanto menor es la estructura objeto de examen.

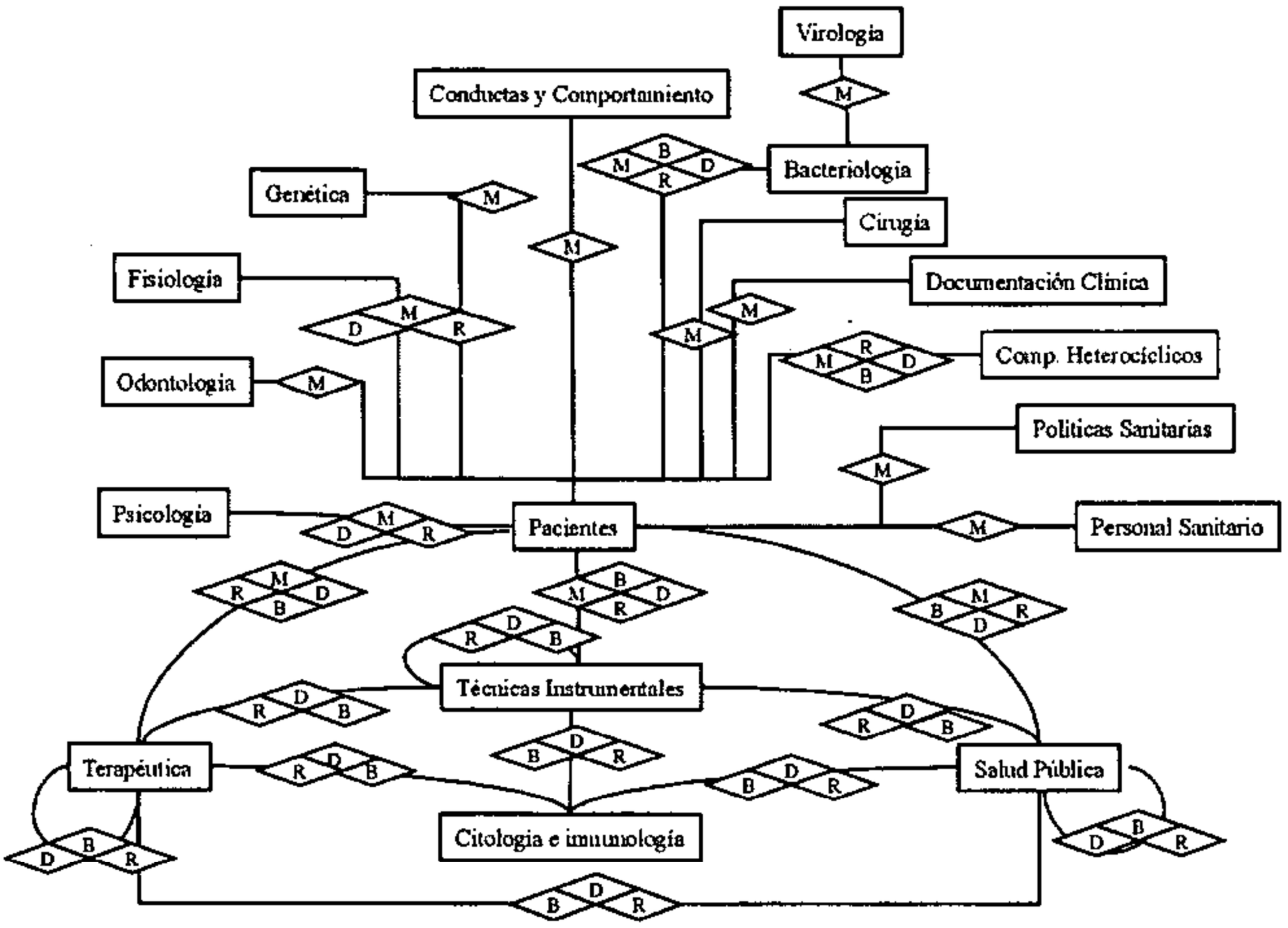
Tabla 2.
Coincidencias Chen y MeSH

		Nº DE TÉRMINOS EN COMÚN CHEN Y MESH	% PORCENTAJE RESPECTO MESH
ESTRUCTURA			
	BIBLIOGRAFÍA	88	10,0
	DISCUSIÓN	118	9,6
	RESULTADOS	136	8,6
	RESUMEN	157	7,1

Se analizaron en qué grupos temáticos del MeSH se agrupan al menos el 50% de los pares de términos suministrados por Chen. Se obtuvo que los mayores pesos y concurrencias se daban dentro del mismo tema. De nuevo, este proceso viene a reafirmar el fenómeno citado por Velasco (1998), consistente en que los términos con mayor peso, tienen tendencia a tener un mayor grado de relación. También, se pueden observar que en las distintas estructuras los pares temáticos suelen ir asociados de la misma manera entre sí. Es más extraño el caso de la metodología, en la que no se observa un solapamiento con las demás estructuras, sino que parece ser obviado cualquier aspecto que tenga que ver con la toma de datos, en las conclusiones y resumen. Discusión y resumen frecuentemente, coinciden no sólamente en la temática general, sino que también ocurre ésto si consideramos descriptores concretos. Los títulos de todas las referencias de la bibliografía no parecen estar relacionados tanto con aspectos metodológicos como sería de esperar.

Gráfico 3.

Pares de Chen, relacionados mediante la temática del MeSH en cada sección del artículo. Se han considerado cuatro secciones (B-bibliografía, D- discusión, R-resumen, M-métodos)



3.4 Variables cienciométricas y lingüísticas

Para el análisis de los resultados se realizó, dada la naturaleza de los datos, un análisis multivariante mediante el paquete estadístico SPSS (Voelkl, 1999). Si bien fue necesario, estudiar los datos previamente con una estadística descriptiva. Se empleó el estadístico de Kruskal-Wallis para analizar si las variables entre las distintas secciones del documento contenían diferencias relevantes, observándose que eran significativamente diferentes ($p<0.05$). A continuación, se realizó un análisis de las componentes principales. Observándose que las dos primeras componentes estaban relacionados con aspectos tales como las superestructuras y la extensión del texto.

4. CONCLUSIONES

La literatura sobre concentraciones de información para la indización citan el resumen, conclusiones, leyendas, principios de párrafos y títulos (Van Slype, 1991). Este hecho se refleja en el presente estudio, las estructuras, en concordancia con la literatura, presentan diferencias significativas al ser tratadas en los distintos módulos. En concreto, el resumen y las conclusiones justifican de este modo su indización preferente. Sin embargo, contrariamente a lo afirmado por Gilyarevsky (1997) o los procedimientos utilizados en algunas bases de datos de recuperación por términos del título, los títulos en documentos médicos no parecen contener una densidad informativa lo suficientemente alta.

Las secciones con menor extensión tienen frecuentemente mejor comportamiento en los distintos módulos. En cualquier caso, por el método de filtrado del n-gram, muchos términos como los relacionados con el nombre de revistas y autores en la bibliografía, muy raramente alcanzarán el nivel que les permita ser identificados como términos valiosos desde el punto de vista discriminatorio. Es el caso de la gran abundancia de términos geográficos en la bibliografía del HIV, en el que se aúna el hecho de que solamente los términos relacionados con el título de la referencia, pasan el umbral, y por otro lado la gran prevalencia de esta enfermedad en Africa, hacen que la terminología geográfica sobresalga.

El diferente estadio de las pandemias estudiadas seguramente pueda explicar el hecho del diferente comportamiento en sus respectivas estructuras. El HIV, con una historia más reciente que la hepatitis, presenta una menor normalización

terminológica, lo cual repercute en sus comparaciones con el k-means y el n-grams.

Se ha comprobado que los indicadores bibliométricos clásicos son realmente responsables de una alta parte de la varianza, y de ahí su alto poder discriminatorio, en los artículos científicos. Si bien, para la extensión a otro tipo de documentos sin bibliografía, existe una necesidad de generar nuevas variables.

BIBLIOGRAFÍA

1. GARFIELD, E. The relationship between mechanical indexing, structural linguistics and information retrieval. En: First Symposium on Machine Methods for Scientific Documentation. (1: march 1953: Johns Hopkins University). Journal of Information Science, 18: 343-354. 1992. p 343-354.

2. SLYPE, Georges Van. Los lenguajes de indización: Concepción, construcción y utilización en los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez, Piramide, 1991.

3. WORMELLl, I. Subject Access Project (SAP). Lund, 1985. Improved Subject Retrieval for Monographic Publications. Ph.D. Thesis. Lund University.

4. PECHEUX, M. Hacia el análisis automático del discurso. Madrid: Gredos, 1978.

5. MITKOV, R. The latest in anaphora resolution: going multilingual. En: Revista de Procesamiento del Lenguaje Natural. Vol. 23 (1998); p 1-7.

6. LLORENS, J. *et al.* Características textuales como medida cualitativa de la información en la generación semiautomática de tesauros. En: Revista de Procesamiento del Lenguaje Natural. Vol. 23 (1998); p 61-68.

7. VELASCO, M. Generación automática de representaciones de Dominios. Madrid, 1998. Tesis Doctoral Univ. Politécnica de Madrid.

8. ABAITUA ODRIOZOLA, Joseba K.; CASILLAS RUBIO, Arantza y MARTÍNEZ UNANUE, Raquel. Segmentación de corpus paralelos para memorias de traducción. En: Revista de Procesamiento del Lenguaje Natural. Vol. 21 (1997); p 17-30.

9. DIJK, Teun A. Van. La noticia como discurso: comprensión, estructura y producción de la información . Barcelona : Paidós, 1996.

10. MOREIRO GONZÁLEZ, José Antonio. Aplicación de las ciencias del texto al resumen documental. [Getafe] : Universidad Carlos III de Madrid ; Madrid : BOE, 1993.

11. SALTON, G; SINGHAL A., Mitra M. y BUCKEY, C. Automatic text structuring and summarization. En: Information Processing and Management. Vol. 33, No. 2(1997); p 193-207.

12. LOOSE, R.M. Text windows and phrases differing by discipline, location in document, and sysntactic structure. En: Inform. Processing & Manag. Vol. 32, No. 6; p 747-767.

13. HEARST, M.y PLAUNT, Christian. Subtopic structuring for full-length document access. En: Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh. New York: ACM, 1993.

14. LEYDESDORFF, Loet. Why words and cowords cannot map the development of the sciences. En: JASIS. Vol. 48, No. 5(1997); p 418-427.

15. SEGLEN, P.O. Quantification of scientific article contents. En: Scientometrics. Vol. 35, No.3 (1996); p 335-366

16. WEISSBERG, R. y BUKER, S. Writing up research. Englewood Cliffs (NJ): Prentice Hall Regents, 1990.

17. BRUCE, N.J. Rhetorical constraints on information structure in medical research report writing. En: ESP in the Arab World Conference. (agosto, 1983: Univ. Aston, UK).

18. HESLTOT, J. Tense and other indexical markers in the typology of scientific texts in English. Hedt, 1982. p 83-103.

19. BERRI, J. *et al.* A linguistic method for text filtering. En: Revista de Procesamiento del lenguaje natural. Vol. 19 (1996); p 159-165.

20. SWALES, J.M. Genre analysis: English in academic and research settings. Cambridge [UK]: Cambridge University Press, 1990.

21. NWOGU, K.N. The medical research paper: structure and functions. En: English for Specific Purposes. Vol. 16, No. 2 (1997); p 119-138.

22. GILYAREVSKY, R; UZILEVSKY, G y MOUDROV, E. An automatic statistical classification of different types of journals. En: Int. Forum on Information and Docum. Vol. 22, No. 3 (1997); p 24-35.

23. MORATO, Jorge. Análisis de las relaciones cienciométricas y lingüísticas en un entorno automatizado. Madrid, 1999.Tesis Doctoral.Universidad Carlos III.

24. LOWE, H. J. y BARNETT, G. O. Understanding and using the medical subject headings vocabulary to perform literature searches. JAMA. 13 271(14): 1103-8

25. A BIBLIOMETRIC analysis of science jounals in the ISI Database. En: SCI Journal Citation Reports. Editor in chief Eugene Garfield, Inst. Scientific Inform. Inc., Philadelphia.

26. LAWRENCE, Steve y BOLLACKER, Kurt. Digital libraries and Autonomous Citation Indexing. En: IEEE Computer. Vol.32, No.6(1999); p 67-71.

27. ESTEVEZ, N. y MARTÍNEZ-PELEGRIN, P. An approach to the linguistic structures of health science articles. Lenguas para fines especificos (V). p 301-309.

28. SKELTON, J. Analysis of the structure of original research papers: An aid to writing original papers for publication. En: British J. of General Practice.Vol. 44 (1994); p 455-459.

29. COHEN, J. Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting. En: JASIS. Vol. 46, No. 3 (1995); p 162-174.

30. FRAKES, W. B. y BAEZA-YATES, R. Information Retrieval. Data Structures and Algorithms. Prentice Hall PTR. Upper Saddle River, New Jersey : 1992.

31. CHEN, H. y LYNCH, K. J. Automatic Construction of Networks of Concepts Characterizing Document Databases. En: IEEE Transactions on Systems, Man and Cybernetics. Vol. 22 (1992); p 885-902.

32. LELU, C. Modéles neuronaux pour l'analyse de données documentaires et textuelles. Paris, 1993. Ph. D. Université de Paris.

33. EGGHE, L y ROUSSAU, R. Introduction to informetrics. Quantitative methods in Library, Documentation and Information Science. Amsterdam: Elservier Science, 1990.

34. VOELKL, K.E. y GERBER, S. Using SPSS for Windows : data analysis. New York: Springer, 1999.