UNIVERSIDAD CARLOS III DE MADRID

FACULTAD DE HUMANIDADES, COMUNICACIÓN Y DOCUMENTACIÓN

TESIS DOCTORAL EN DOCUMENTACIÓN

ANÁLISIS DE RELACIONES CIENCIOMÉTRICAS Y LINGÜÍSTICAS EN UN ENTORNO AUTOMATIZADO

CODIRECTORES

JUAN BAUTISTA LLORENS MORILLO

JOSE ANTONIO MOREIRO GONZALEZ

AUTOR: JORGE MORATO LARA

Abril, 1999

TABLA DE CONTENIDO

1 INTRODUCCIÓN

1.1 INTRODUCCIÓN

La enorme rapidez en la sucesión de los descubrimientos científicos, potenciado por los grandes avances tecnológicos, la mayor inversión y el aumento del número de científicos, ha permitido un desarrollo exponencial de la producción científica.

A medida que este crecimiento aumentaba, también se produjo un rápido envejecimiento de los documentos y la inaccesibilidad de muchos otros. Al tiempo, aparecían redes de comunicaciones más rápidas y se incrementaba el número de publicaciones electrónicas sin revisar.

Para poder tratar este volumen de información se han desarrollado herramientas que permiten el estudio de las conexiones entre documentos, las pautas de publicación, la representación del contenido y la optimización de la recuperación. La realización de estudios por medios manuales es costosa, lenta y de calidad variable. Así se ha ido haciendo cada vez más imperiosa la creación de una herramienta rápida y flexible para evaluar y gestionar esta información, siendo patente la necesidad de la informática para realizar estos estudios (Cleveland, 1990).

Algunas de estas herramientas, como la indización es necesaria para que un sistema de información sea eficiente, ya que la eficacia de un sistema está en función de su organización (Granda, 1990). Otras herramientas como la generación de tesauros o los índices cienciométricos, también, tienen tendencia a presentar una automatización complicada, principalmente debido a lo poco extrapolables que han sido los experimentos de un tema a otro.

Gran parte de esta falta de aplicación entre diferentes campos se debía a no haber evaluado con anterioridad la influencia del género donde se producía la comunicación. Durante la década de los ochenta aparecieron en lingüística una serie de metodologías que permitían analizar los distintos contextos y estructuras de los textos (Swales, 1990). A mediados de los años 90 el análisis de género empieza a adquirir importancia en la recuperación y análisis de la información (Losee, 1996).

En la etapa actual, que tiende a la clasificación automática valiéndose de la aplicación de umbrales, no pueden existir buenos resultados sin tomar en consideración características propias del subcampo (o subgénero) en el que nos movemos.

Por otro lado, la necesidad de un modelo integrador es el objetivo último de los procesos de documentación e información para resolver una necesidad de conocimiento (García-Marco, 1995). Para ello se expone una metodología que analiza globalmente todas estos factores. Por otro lado, se describen una serie de herramientas que se han desarrollado para poder aplicarla.

En este trabajo, se pretende presentar una visión integradora de todos estos aspectos. A lo largo de la tesis se entremezclan conceptos de Psicología cognitiva, Lingüística, Cienciometría, Documentación, Estadística, Clasificación e Informática, en concreto en sus vertientes más relacionadas con el tratamiento, organización y caracterización de la información textual.

No hay que olvidar, que aunque criticados, existe una gran variedad de estudios que avalan la utilidad de los indicadores cienciométricos (Garfield, 1998). Aunque con severas limitaciones, los métodos expuestos por Callon (1995) deben de complementar a los indicadores, no sustituirlos.

El objetivo final es analizar la influencia que tiene el análisis de género en la caracterización de los parámetros cualitativos y cuantitativos, y en concreto, de las herramientas que se encargan tradicionalmente de estos estudios, como los indicadores cienciométricos y la clasificación de términos.

1.2 ESTRUCTURA DE LA TESIS

El contenido de la tesis está estructurado en siete capítulos que se detallan a continuación:

- Capítulo uno: Introducción, estructura de la tesis y agradecimientos.
- Capítulo dos, se describe un estado de la cuestión de los diferentes aspectos teóricos. Este capítulo esta subdividido en varios apartados donde se explican diversos aspectos de la cienciometría, lingüística, filtrado, indización y clasificación.
- Capítulo tres, se exponen los objetivos que se pretenden alcanzar en la tesis.
- Capítulo cuatro, se muestra el entorno de trabajo donde se han desarrollado las distintas aplicaciones que se han programado.
- Capítulo cinco, se detalla la metodología empleada.
- Capítulo seis, se expondrán los resultados.
- Capítulo siete, se comentarán los resultados.
- El último capítulo esta dedicado a las referencias bibliográficas, recursos y bases de datos utilizadas.

1.3 AGRADECIMIENTOS

Existe una buena razón para que estos agradecimientos sean extensos, más debido a mi ignorancia que a un carácter agradecido. En efecto, aunque mis conocimientos de programación sean rudimentarios la presente tesis utiliza aproximadamente cincuenta mil líneas de código, diez mil son exclusivas de este trabajo, de las que por supuesto no soy autor. Por lo tanto, quiero agradecer a los miembros, actuales o no, del GTI (cuyas siglas son 'IE') la realización de todo este trabajo. Y en particular:

A Luís, que programó Cmetría, lo cual es de agradecer. Y por haberse quedado bastantes tardes, después de acabar su jornada, ayudándome con las bases de datos y con los programas que no funcionaban. Para hacerse una idea de la complejidad del programa, basta decir que si durante el último año hubiera tenido un accidente el responsable, sin duda, sería Luís. En cualquier caso se lo tenía merecido por hacer la mili.

A Alda. Si Luís es el responsable del 60% de la tesis, Alda ha tenido que ver con el restante 55% (en sus propios términos). Alda me ha ayudado con la documentación, las reglas de normalización y, sobre todo, con su amistad incontables veces.

A Inés, que desarrolló el módulo de cálculo de variables. Siguiendo con el porcentaje, creo, le sería atribuible otro 50%. Durante varios meses me preguntó, con escaso éxito, características de la base de datos y de los datos en sí. La realización de su proyecto fue especialmente difícil dada la escasez de datos en aquél momento, lo cual es una pesadilla para cualquiera que le guste supervisar su trabajo.

A JM. Durante meses le transmití las preguntas de Inés. Astutamente, creo, y para evitar tal avalancha de preguntas, me explicó, con todo detalle, todo lo que tenía que ver con la estructura de la base de datos y la utilización de programas que me pudiesen ser de utilidad (SQL Talk, Access, Macros,...). Como todo buen plan, falló, durante los siguientes meses ha sido la primera persona a quien preguntarle algo, dado lo detallado de las explicaciones.

A Emi y a Raúl, por todos los problemas que me han resuelto durante meses. Lo malo de tener varias bases de datos de más de

ciento cincuenta Megas es, aparte de tener la cuenta en red más voluminosa del labo, la cantidad de tiempo que se obliga a perder a los administradores cada vez que hay un problema. A Emi, por todo lo que me ha ayudado con Access.

A Irene, por haberme ayudado siempre que se lo he pedido. En concreto, por la programación del Indizador junto con Manu, Juan y JM.

A Víctor, por arreglarme, siempre en poco tiempo, incontables programas para funcionar a nivel de colección y por tabular el MeSH.

A Manu, por todas las dudas que me ha resuelto, ya que sobre el funcionamiento de todos los algoritmos es quién sabe como funcionan en la práctica. Además, esta tesis se beneficia bastante de todos los desarrollos derivados de su tesis.

A Javi, por todos los papeleos que ha tenido que realizar por mi causa. Pero sobre todo por tantas buenas sobremesas junto con Gonzalo y Manu.

A Asún, que me ha resuelto incontables dudas sobre el n-grams y que realizó uno de los programas que mejores resultados y menos problemas me ha dado.

A José, por sus conocimientos sobre los programas más insospechados y por todas sus anécdotas.

A Julio, por tener siempre la respuesta cuando todo falla. Y junto con José y Luis por el normalizador.

A Marcos, por todas las direcciones útiles.

A Víctor, por programar el integrador de todos los clasificadores.

A Josué, por lo del Chen (y a Mercedes, por lo mismo, claro).

Por todo lo anterior les estoy, de verdad, agradecido, pero entre todo, lo que más valoro es su amistad.

A Isa y Paolo por ayudarme con toda la lingüística.

A Raquel y a mi familia, decir por qué sería poner un límite innecesario e ilógico, a todo lo que les debo.

A mis codirectores, Jamore y Juan. A los dos les debo el haberme ayudado siempre que los he necesitado. A Jamore gracias por lo de Polanco. A Juan por lo de la beca y por todos los medios que he tenido durante el último año y medio.

Una vez me comentó Juan su creencia en que, según se acercaba el fin de una tesis, el doctorando tendía a contradecir al director. No creo que esto sea verdad, aunque no estoy seguro si con esto le quito o le doy la razón. En cualquier caso, si hubiera que dividir qué parte de esta tesis es mérito suyo sería sencillo, todo lo que merezca la pena, el resto es mío.

2 ESTADO DE LA CUESTIÓN

2.1 LINGÜÍSTICA

2.1.1 ANÁLISIS DEL DISCURSO

Según Van Slype (1988), 'Una expresión no debería caracterizarse solamente en cuanto a su estructura interna y su significado si no también en función del acto realizado al producir tal expresión'. Es con este objetivo con el que vamos a emprender un análisis del discurso. El discurso, según Van Dijk (1996) lo define, es una disciplina que estudia el texto y el habla desde todas las perspectivas posibles. Existen dos dimensiones:

- Textual, estudia las estructuras del discurso a diferentes niveles de descripción.
- Contextual, analiza la dimensión textual con las diferentes propiedades del contexto, como los procesos cognitivos y las representaciones o factores socioculturales.

Por otro lado, Swales (1990) describió el discurso como "...a recognisable communicative event characterised by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs. Most often it is highly structured and conventionalised with constraints...[1]".

Existen algunas dimensiones del discurso de interés para el presente estudio:

- Estilo. El estilo son las variaciones en el discurso que el hablante o escritor realiza según el rol del contexto, pero sin implicar un cambio en la semántica. Estas variaciones pueden ser de tipo léxico, entre diferentes términos equivalentes, o la elección de diferentes estructuras semánticas.
- Pragmática. Las reglas pragmáticas determinan el uso sistemático de las expresiones por parte de una comunidad. Uno

---

[1] 'Un acto comunicativo caracterizado por un conjunto de objetivos comunicativos que son identificados y comprendidos por una comunidad académica o profesional. Frecuentemente, [el discurso] esta muy estructurado y obedece a ciertos convencionalismos.' (N.A.)

de los objetivos de la pragmática es relacionar las expresiones con un contexto y formular que expresiones son satisfactorias y en que condiciones.

- Superestructuras. La caracterización semántica de las estructuras discursivas se puede hacer a un nivel de organización global. Las macro-estructuras juegan un papel principal en la producción y comprensión del discurso. Estas estructuras del texto son un tipo de macrosintaxis que caracterizan de forma global al discurso. Para describir los capítulos o apartados de un discurso escrito es necesario una macrosemántica, que nos indique las relaciones imposibles de describir mediante frases aisladas. Toda clase de textos argumentativos tienen categorías globales como premisas y conclusión y, posiblemente, sub-categorías como condición o garantía.

  El discurso científico es del tipo introducción, problema, solución y conclusión, con estructuras argumentativas incrustadas en varias clases (para una descripción detallada de cada estructura ver: R. Weissberg (1990). La tarea de una teoría general del discurso es clasificar y definir tales categorías, reglas y funciones textuales; debiéndose incluir reglas específicas de ciertos contextos y situaciones sociales.

- Retórica. Las estructuras retóricas no son necesarias en el discurso, a diferencia del estilo. El hablante utiliza la retórica para intensificar la organización y, así, mejorar el almacenamiento y la recuperación de la información por parte del receptor.

La retórica, junto con la estilística y la investigación literaria, sirven para diferenciar tipos de discurso y determinan efectos específicos de comunicación discursiva (Van Dijk, 1996).

2.1.1.1 EL DISCURSO CIENTÍFICO

El genero de artículo científico se desarrolló a partir de las relaciones epistolares de la comunidad científica (Ard, 1983). La evolución de estos documentos y, posteriormente, de los artículos científicos es interesante. Una revisión somera de ciertos datos nos enseñan como el discurso, lejos de ser un sistema rígido, tiene su propia dinámica.

En un principio, las cartas dirigidas en primera persona fueron evolucionando a un estilo cada vez más impersonal. También, la extensión del artículo sufrió variaciones, desde las 5.000 palabras de media del 1900 hasta las 10.000 del 1980. Por otro lado, las referencias comienzan a ser cada vez más concretas y abundantes, mientras que sus citas tienden a seguir las macroestructuras del texto. Las frases han mantenido su longitud pero las frases causales, nominales y temporales han crecido en frecuencia. Los términos tienden a una mayor abstracción, al tiempo que los dibujos de instrumental casi desaparecen y aumentan los esquemas. Respecto a la macroestructura actual se generaliza a partir de los años 30. Otras variaciones son el incremento en la coautoría y en la utilización de estadísticas en el texto (Swales, 1988).

El lenguaje científico se caracteriza por poseer un vocabulario objetivo, normalizado y universal. El discurso científico esta estructurado según ciertas pautas de organización retórica, aunque con una cierta libertad individual de variación estilística (Widdowson, 1979). La macroestructura del artículo de investigación propuesta por Bruce (1983) consiste en la división de secciones normalizadas, en concreto: Introduction, Method, Result, Discussion (conocido como esquema IMRD) ya que sigue la línea general del razonamiento inductivo. Otros modelos han sido propuestos, pero al no existir una evidencia clara en los documentos que reflejen esta estructura no han sido aceptados. Para una revisión de los estudios sobre la estructura de artículos científicos ver Swales (1990).

La variación temática y las distintas apariciones de ventanas de texto en el texto han sido estudiadas por Jacquemin (1994, 1996). El concepto de ventana, según este autor, es 'un abanico de palabras contiguas dentro del documento'. En un estudio sobre las ventanas más significativas estadísticamente (Losee, 1996), se

observó que difería dentro del campo académico y dentro de las distintas secciones del documento.

2.1.1.1.1 CARACTERISTICAS DE LAS SECCIONES EN TEXTOS CIENTÍFICOS

Swales (1990), concentró parte de su análisis en desarrollar un método que cuantificara las características lingüísticas de las estructuras del texto. Existen diferencias en la distribución de las características lingüísticas y retóricas a lo largo del texto (ver tabla), por ejemplo, según Heslot (1982) el 90% de los tiempos verbales en presente están entre la introducción y la discusión.

CARACTERISTICA	INTROD.	METODOS	RESULT.	DISCUSION	AUTORES
Comentarios del autor	alto	muy bajo	muy bajo	alto	Adams Smith
Voz pasiva	Bajo	alto	variable	variable	Heslot
Pasado	Bajo	alto	alto	bajo	Heslot
Presente	alto	bajo	bajo	alto	Heslot
Declaración Inform.	alto	bajo	bajo	alto	West

Tabla 1. Características entre las secciones de los textos científicos (tomado de Swales, 1990)

Dentro de cada sección también aparecen subsecciones. Los investigadores en sus escritos hacen, continuamente, referencia al contexto de su disciplina donde se sitúan. Por ejemplo, la introducción esta subdividida en objetivos, problema, solución, criterios y capacidad de análisis (Zappen, 1983). Swales (1990), identificó listas de frases y términos que indican varios micromovimientos dentro de cada sección, que al mismo tiempo están subdivididos en varios pasos.

Por ejemplo, en el paso 3 de la Introducción ("Reviewing items of previous research"), se analizan cómo son las pautas de citación, según ciertas pautas, una citación integral es cuando el nombre del investigador aparece como un elemento de la frase, mientras que una citación no integral es en la que aparece, entre paréntesis, el nombre del autor y el año. A su vez, estos dos tipos pueden estar asociados a verbos informativos (show, establish, claim, etc.) o no informativos y concurrir o no con identificadores de negaciones o cuasinegaciones. Asociándose, por último, estas formas al tiempo verbal.

Los identificadores negativos o cuasinegaciones, pueden ser de distintos tipos (Swales, 1990) desde cuantitativos (no, little, none, few), verbos (fail, lack, overlook), adjetivos (inconclusive, complex, misleading, elusive, scarce), sustantivos (failure, limitation) u otros (without regard of).

Losse (1996), señala a los resúmenes de los artículos experimentales como la estructura más útil para analizar, ya que, a diferencia de los artículos más teóricos, sus ventanas tienen una intersección mayor con el texto completo.

El análisis de genero se ha aplicado a una gran variedad de campos. Entre estos está el de sus aplicaciones para la recuperación de información. Por ejemplo, aplicado a Internet (Amitay, 1998; Losee, 1996).

2.1.1.1.2 ANÁLISIS DE GENERO EN ARTÍCULOS DE INVESTIGACIÓN MÉDICA

La estructura de la información en documentos de investigación en medicina ha sido analizada mediante el análisis de género, propuesto por Swales. En el caso de Skelton (1994) se trata de ampliar el trabajo de Weissberg (1990) para enseñar a escribir artículos con una estructura científica, pero esta vez enfocado exclusivamente a la medicina. En un nivel más general, mediante la cuantificación y análisis estadístico de palabras, temas e ilustraciones ha sido estudiado por Seglen (1996).

Más interesante es el trabajo, poco conocido, de Estevez (1996), en este artículo partiendo de un corpus de ocho artículos procedentes de 6 revistas, se estudia la frecuencia de términos frecuentes en cada micromovimiento de la sección. Además, tabuló la frecuencia de los verbos (tanto en activa como pasiva, para el presente, pasado, modales, presente perfecto y continuo) de cada sección y micromovimiento.

Nwogu (1997), con un corpus de 30 textos procedentes de Lancet, British Medical Journal, the New England Jounal Of Medicine, The Journal of Clinical Investigation y el Journal of the American Medical Association, llegó a conclusiones bastante similares al describir los micromovimientos de las IMRD del documento. También, amplía la lista del vocabulario que emplea cada sección.

2.1.1.2 ANÁLISIS DE GENERO EN ARTÍCULOS DE DIVULGACIÓN

Un campo mucho menos conocido es el de la estructura de los artículos de divulgación. La estructura típica es de Problema-Solución. Nwogu (1991) esbozó nueve movimientos para artículos de divulgación en medicina. Posteguillo (1996) comparó varias características de estos artículos con otros de divulgación en el campo de la informática. Algunas de las conclusiones a las que se llego en estos artículos fueron:

- Los estilos no académicos presentan una gran disminución en la aparición de marcadores del discurso (in this section, first, secondly, here, etc.) frente a los artículos de investigación. Por otro lado los artículos de divulgación tienen muchas más fuentes tipográficas.

- La utilización de verbos en pasiva y presente simple también es mucho menor en los géneros no académicos. En los no académicos está presente, sobre todo, el pasado simple.

- Utilización de acrónimos. Los artículos de divulgación utilizan más acrónimos no científicos que los de investigación. Más curioso es el hecho de que dentro de los acrónimos científicos, los artículos de divulgación, especifiquen en menor medida cuál es su significado.

2.1.1.3 ANÁLISIS DE GÉNERO EN PRENSA

En Van Dijk (1996), se muestra una metodología para analizar artículos periodísticos mediante la estructuración temática en diversas jerarquías. Este método puede incluir categorías fijas (p.e. causas, actor principal, antecedentes, consecuencias). El discurso se puede definir en términos de un esquema de categorías jerárquicamente ordenadas. Con un enfoque similar se pueden ver los trabajos de Pinto (1996).

El discurso periodístico, como el científico, puede tener una forma convencional, que a menudo es un esquema argumentativo (premisas que llevan a una conclusión). Las categorías del esquema periodístico: titular, encabezamiento, episodio (acontecimiento principal y antecedente), consecuencias y comentario. De entre éstos Van Dijk, señala como un resumen del esquema informativo a los dos primeros. Siguiendo este esquema, se consideró oportuno en nuestro estudio considerar el encabezamiento como equivalente al abstract.

Van Dijk (1996), también realizó estudios sobre la semejanza entre el número de palabras por frase en el género de prensa y en artículos de investigación. En el último, las frases son más breves, en prensa es fácil encontrar oraciones subordinadas, algo que es menos probable en investigación.

2.1.2 LINGÜÍSTICA DOCUMENTAL

La estructura de la documentación como organización de contenidos codificables es el objeto de la lingüística documental, desde un doble objetivo: la estructura de la producción de información, la formación y presentación de ideas por el autor y la estructura de representación por parte del productor (García-Gutierrez (1990).

Existe un paralelismo entre la lingüística y la documentación. Como ejemplo están los hiperónimos e hipónimos con respecto a los genéricos y específicos de los tesauros. Esta relación se puede ver claramente en WordNet[2].

Este hecho también se puede apreciar en García-Gutierrez (1990), mediante la descripción de scriptores y axiemas. Un axiema es la unidad mínima de articulación o de posible relación paradigmática y sintagmática en el seno de los lenguajes documentales (p.e. pre-, pos-). Algunas de sus combinaciones pueden formar scriptores, que son la unidad mínima de combinación en el plano del enunciado documental (p.e. pre y guerra para formar preguerra).

Por último, García-Marco (1995) ha estudiado las relaciones del contenido documental con los niveles de comunicación lingüística, de la siguiente manera:

- Nivel pragmático: entendido como nivel de comunicación documental, es el análisis de las necesidades de los usuarios o la organización temática.

- Nivel conceptual: análisis de las entidades y las relaciones de conocimiento. Detección de conceptos y de las relaciones entre estos.

- Nivel semántico: validación del sistema de signos. Correspondencias entre el sistema de signos, el conjunto de conceptos y el de referentes.

- Nivel significante: elección de los significantes que codifican el sistema de conceptos y traducción al lenguaje documental.

---

[2] WordNet es el programa más difundido en lingüística computacional y análisis de textos, para una revisión ver Leacock (1998) o Fallbaum (1998). Se puede conseguir gratuitamente en http://www.cogsci.princeton.edu/~wn.

Control de relaciones entre vocabulario controlado y lenguaje natural. También esta fase se corresponde con la recodificación condensativa.

- Nivel señalizante: operaciones de representación y recuperación del documento.

2.1.3 LEGIBILIDAD

La legibilidad es un indicador lingüístico que expresa la dificultad de un texto para su lectura. En Pinto Molina (1994), ya se propuso este indicador lingüístico como una herramienta útil para la gestión de los procesos analítico documentales.

El concepto de legibilidad nace de la necesidad de medir de manera estadística la facilidad o dificultad de lectura de una texto. Comprensión y legibilidad son distintos conceptos, pero están relacionados. A mayor legibilidad mejor comprensión.

Los primeros estudios sobre legibilidad se deben a los trabajos de Thorndike (1944) sobre la longitud de las palabras en inglés. Más tarde se fue ampliando el concepto a grupos más amplios: grupos nominales y verbales y frases. Esta ampliación se debió a que la comprensión implica un conocimiento sintáctico y gramático por parte del receptor. El índice de legibilidad se ha aplicado en ocasiones para medir la legibilidad de textos de científicos (Johnson, 1987). Los dos índices más conocidos son:

2.1.3.1 NIVEL DE FACILIDAD DE LECTURA DE FLESCH

El índice de legibilidad de Flesch, es el de más utilizado y fue realizado por Rudolf Flesch, en 1948. Es el índice que calcula el procesador de textos Word97, en el menú opciones ortografía. Considera dos variables principales: la longitud de la palabra en sílabas y la longitud media de las palabras de cada frase. Se calcula según la siguiente fórmula:

$$\text{Indice de Flesch} = 206.835 - (0.864 \times S) - (1.015 \times W)$$

Donde:

S = es el número de sílabas

W = el número medio de palabras por frase

Valora el texto en una escala de 100 puntos; cuanto más alto sea el resultado, más fácil será comprender el documento. Para la mayoría de los documentos estándar, el objetivo es un resultado comprendido entre 60 y 70 aproximadamente (Lucisano, 1988).

Aunque es sencillo de calcular, a la hora de programarlo tiene la dificultad del cálculo de las sílabas de cada palabra.

2.1.3.2 INDICE GULPEASE

Fue creado en 1988 por IBM y el GULP (Gruppo Universitario Lingüístico Pedagógico de la Universidad de Roma). Se consideran dos variables lingüísticas: las palabras (según su longitud en caracteres) y la frase (medida por la longitud media de las palabras de la frase). La fórmula es la siguiente:

Indice Gulpease = 89 - (Lp / 10) + (3 x Fr)

Donde:

Lp = (100 x total de letras) / total de palabras

Fr = (100 x total de frases) / total de palabras

A más alto el valor, más legible. Los valores del índice de legibilidad varían de un idioma a otro (Luciano, 1992).

2.2 INFORMETRIA

La informetría trata de la medición, incluyendo el análisis mediante herramientas matemáticas, de todos los aspectos de la información, abarcando su almacenado y recuperación (Egghe, 1990). En un principio el término cienciometría se utilizó en EE.UU., mientras que el de informetría se usaba en los países del Este para referirse a temas similares (Egghe, 1990). Tras hacer una búsqueda bibliográfica en LISA, se puede ver que los términos más utilizados actualmente son bibliometría y cienciometría, con una tendencia a excluirse mutuamente y matizar su semántica. Aunque hay que precisar que no faltan ejemplos en la literatura en que estos términos se utilizan como sinónimos, como una jerarquía (el más genérico, informetría) o como relacionados.

Tradicionalmente, ha sido la Cienciometría "la ciencia de medir la ciencia" Bookstein (1995), la encargada de identificar y medir las diferentes agrupaciones de documentos científicos, ya sea mediante análisis de referencias, citas o coaparición de palabras (Callon, 1995).

Aunque estos indicadores cienciométricos tienen una amplia difusión, la literatura sobre indicadores cualitativos continua siendo escasa (Rip, 1997). Algunos de los objetivos de la cienciometría son conocer las características implicadas en la comunicación científica, evaluar las actividades científico-investigadoras y determinar los procesos ligados al consumo de información científica. Para ello se utilizan indicadores cienciométricos, que son los datos que se extraen de los documentos que publican los investigadores o que utilizan los usuarios.

Un término muy relacionado es la bibliometría, definida por Pritchard (1969) como los estudios orientados a la cuantificación de los procesos de la comunicación escrita. En general, persiguen un doble objetivo, estudios estadístico descriptivos y de las relaciones de la literatura, y por otro lado el análisis de estudios sobre los aspectos sociométricos de los grupos que producen esos documentos. Según Moed (1989), los estudios cuantitativos de la actividad científica se corresponden con los estudios cuantitativos de la literatura de la ciencia. Los estudios bibliométricos se han sucedido desde los trabajos de Cole y Eales en 1917.

2.2.1 CALIDAD DE LA INVESTIGACIÓN

El inconveniente que tiene medir la calidad de la investigación es la subjetividad de lo que se pretende medir, por lo que todos los indicadores tienen sus desventajas. Normalmente, se utilizan los siguientes parámetros para medir la calidad de la investigación:

2.2.1.1 ANÁLISIS DE CITAS

Se trata de la remisión bibliográfica obtenida por un documento a partir de otro publicado posteriormente (Ferreiro, 1993). Se ha utilizado para evaluación científica desde que Cole y Clark (1967) encontraron una correlación positiva con la calidad, aunque no han faltado trabajos de signo contrario (Lancaster y Pontigo). En cualquier caso es un indicador de impacto (Martín e Irvine, 1983) y un indicador de la visibilidad. Es el método más conocido y criticado.

2.2.1.2 OPINIÓN DE EXPERTOS

La supervisión de la documentación, por parte de los revisores cualificados, es la garantía tradicional de calidad. El 'peer review' es el método más fiable en la selección de artículos de revistas, pero es demasiado lento y muy subjetivo.

2.2.1.3 CENTRADOS EN LOS AUTORES

Ya sea individualmente o como grupo. Se refiere a premios científicos obtenidos, su curriculum vitae, los proyectos aprobados, la financiación que han recibido, el prestigio académico, etc.

2.2.2 FUENTES DE DATOS EN BIBLIOMETRÍA

Los objetos de estudio bibliométrico son tanto el mensaje como su soporte. Estos soportes se denominan unidades documentales bibliométricas. Pueden ser documentos primarios, secundarios, terciarios o de consulta. Cuando se trabaja con conjuntos de documentos, los conjuntos deben de poseer características comunes para hacer posible su tratamiento cuantitativo.

Como ejemplo de unidades bibliométricas, están (Ferreiro, 1993):

- Publicaciones no periódicas: mapas, monografías, patentes, tesinas, etc.

- Publicaciones periódicas primarias.
- Publicaciones periódicas secundarias (revistas de resúmenes, etc.)

Los instrumentos y métodos cienciométricos están concebidos para identificar y tratar informaciones contenidas en publicaciones científicas y técnicas. Éstas son esencialmente artículos, ponencias, monografías o patentes. Dentro de cada grupo se pueden hacer subdivisiones, por ejemplo, existen artículos científicos que trabajan con investigación básica o aplicada, etc.

2.2.2.1 TEMATICA

La asignación del tema se puede basar en los descriptores, la temática de la publicación, en la clasificación ofrecida por la base de datos, etc. Cuando se trabaja en entornos de clasificación automática palabras que tendrían un bajo poder discriminante según la distribución de Zipf o con el método Inverse Document Frequency, pueden pasar a ser consideradas temas (Velasco, 1998).

2.2.3 LEYES BIBLIOMÉTRICAS

Existen ciertas pautas comunes que siguen los documentos en bibliometría.

2.2.3.1 LEY DE CRECIMIENTO EXPONENCIAL DE LA INFORMACIÓN CIENTÍFICA

Si se representa gráficamente la literatura científica frente al tiempo obtendremos una curva exponencial. Se prevé que pueda existir un límite de saturación que transforme la curva exponencial en logística (López, 1996).

2.2.3.2 LEY DE OBSOLESCENCIA DE LA LITERATURA CIENTÍFICA

Según se multiplica por dos la literatura científica cada 10 a 15 años el número de citas que reciben las publicaciones se divide por dos cada 12 años.

2.2.3.3 LEY DE DISPERSIÓN DE LA LITERATURA CIENTÍFICA

La Ley de Bradford (1948), estudia la distribución de la literatura científica. Bertram C. Bradford trabajó como bibliotecario y químico en el Museo de Ciencias de Londres. Según Bradford, cada tema científico sería tratado preferentemente por un conjunto reducido de publicaciones periódicas. Cuando se estudia cualquier tema, se puede constatar que la mayoría de los artículos son publicados por un pequeño número de revistas (revistas del núcleo). A partir de esta zona hará falta un número muy superior de revistas para recuperar el mismo número de artículos (zona dos de Bradford), y así sucesivamente con el resto de las zonas. El número de zonas a determinar es arbitrario, aunque originariamente Bradford definió tres zonas. Se suele utilizar este indicador para determinar las fuentes más importantes y su evolución.

Para ilustrar esta Ley se puede ver la cobertura que ofreció Medline sobre el HIV/AIDS en el período 1982-1987. De los 8500 artículos de la base de datos distribuidos en 165 revistas, la tercera parte, casi 3000 artículos se encuentran en tan sólo 15 publicaciones. De estas revistas las diez primeras fueron Lancet, JAMA, New England Journal of Medicine, Nature, Science, British Medical Journal, MMWR, American Journal of Medicine y Journal of Infectious Diseases. Por lo tanto, el rendimiento de un fondo documental sobre este tema, estará en función de se encuentre en dicho fondo dichos títulos y mejorará con los de zonas adyacentes.

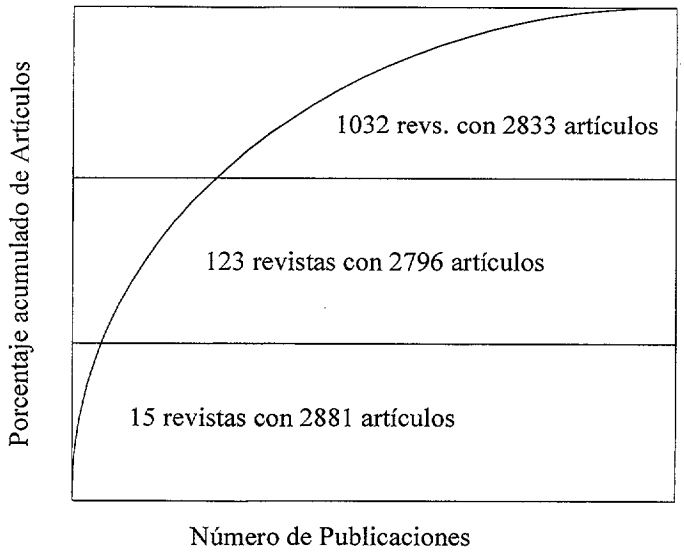


Figura 1: Representación del número de publicaciones frente al número de artículos para el tema SIDA en el periodo 1982-1987 (Tomado de Lancaster, 1998)

Ha habido varios estudios que relacionan los artículos del núcleo con un alto número de citaciones, y que los artículos más interesantes según los expertos están en este núcleo (Lawani, 1982). Según Lancaster (1998), la mayor productividad de una publicación no implica que tengan más calidad que otras.

La Ley de Lotka (1926) muestra la misma relación de dispersión pero con los autores. Lotka observó que el número de trabajos científicos que publicaban los autores (sólo teniendo en cuenta el primer autor) cumplía la siguiente relación: $a_n = k/n^2$. Donde a es el número de autores que producen n trabajos; y k la constante característica de cada materia (determinada por el número de autores que elaboran un solo trabajo).

2.2.3.4 EXPLICACION DE LAS LEYES

El por qué de estas distribuciones ha intentado ser explicado de varias maneras (Egghe, 1990). La más popular es la SBS. El éxito engendra más éxito (success- breeds- success o SBS), es un argumento probabilístico, consiste en que cuantos más objetos tiene una fuente, mayor es la probabilidad de que esta fuente atraiga a otro objeto, si la fuente carece de objetos la probabilidad de que crezca es baja. Por ejemplo, los autores más productivos

tienden a ser los que más publiquen o las revistas más citadas en la bibliografía tienden, aún, a ser más referenciadas.

Existen otras teorías menos conocidas como el argumento de combinación de fractales de Mandelbrot o la función de argumentos de Bookstein (para una revisión detallada ver Egghe (1990)).

2.2.4 INDICADORES BIBLIOMÉTRICOS

Aunque existe gran variedad de indicadores, su tendencia es a un incremento a partir de los indicadores secundarios (Callon, 1996). Algunos de estos nuevos índices podrían ser, los ya citados, del número de términos en los títulos, número de tablas, o de ventanas (Seglen, 1996; Gilyarevsky, 1997; Losse, 1996). Existen abundantes revisiones de indicadores, por ejemplo, Osareh (1996). Dentro de la investigación en medicina López Piñero (1992) realizó una revisión. A continuación se expone una revisión somera de los más clásicos:

2.2.4.1 INFLUENCIA DE LAS FUENTES

Influencia que tienen las publicaciones dentro de su disciplina, es relevante para evaluar publicaciones y realizar la política de adquisiciones. La influencia de las fuentes es igual al total de citas que recibe la publicación de otras dividido por el total de referencias que la publicación hace a otras.

2.2.4.2 ENVEJECIMIENTO DE LA CIENCIA

El envejecimiento obedece a razones como que la información esté anticuada o haber perdido actualidad el tema o el campo científico que se estudia. La literatura técnica envejece más rápidamente que las ciencias sociales, por ejemplo la fisiología se puede considerar vigente entorno a los 7 años y la botánica entorno a los 10 años.

Normalmente se mide mediante la vida media o según el Índice de Price, (Price, 1965) que es el tanto por ciento de referencias con menos de 6 años de antigüedad.

2.2.4.3 PALABRAS ASOCIADAS

Callon (1995), divide los indicadores en tres tipos:

- Indicadores de actividad: número de publicaciones, número de referencias, la producción de artículos científicos de un país, etc.

- Indicadores relacionales de primera generación: las coautorías, las colaboraciones institucionales, cocitaciones, etc.

- Indicadores relacionales de segunda generación o el análisis de las palabras asociadas.

Basado en los estudios de Zipf (1949). Busca los documentos que contienen los mismos descriptores relacionándolos temáticamente (Callon, 1993).

Ofrece una gran cantidad de ventajas por trabajar directamente con el texto íntegro y no con una sección de él, como las referencias o la filiación. Además, permite la realización de mapas de la ciencia, puede utilizar una gran diversidad de fuentes y posibilita un tratamiento de la información más rápido y eficiente que otros métodos.

Se ha de tener en cuenta que los otros indicadores, por ejemplo, el método de citas conjuntas, a pesar de haber tenido buen predicamento, ignoran sistemáticamente las publicaciones sin referencias, lo mismo se puede decir de los índices de coautoría, para memorias de empresas o instrucciones de aparatos, etc. Es decir, estos indicadores a diferencia del de palabras asociadas tienen un campo de aplicación limitado.

2.2.4.4 INDICE DE PRODUCTIVIDAD

Representa el logaritmo de los documentos publicados por país, institución o grupo de investigación.

2.2.4.5 ANÁLISIS DE REFERENCIAS

Son las menciones que un autor hace a documentos anteriores. Se suelen analizar factores como la obsolescencia de la antigüedad documental, tipología documental, coautoría, autocitas, temática, índice de aislamiento (porcentaje de documentos referenciados que pertenecen al mismo país que las publicaciones / autores citantes), capacidad idiomática (idiomas en los que pueden trabajar los investigadores), índice de dispersión (documentos más utilizados), etc.

2.2.4.6 ANÁLISIS DE CITAS

Parece conveniente comenzar por establecer el significado exacto de "cita" y "referencia". Narin et al. (1976) los definen diciendo que

una referencia es la confirmación de que un documento lleva (conduce) a otro; sin embargo el reconocimiento que un documento recibe por parte de otro se llama cita. Por ejemplo cuando el documento (A) aparece en la lista de referencias del documento (B), esto significa que el documento (A) ha sido citado por el documento (B) como fuente de información en apoyo de una idea o un hecho. En este caso, el documento (A) no sólo es una referencia del documento (B) sino que también ha recibido una cita del documento (B) (Garfield 1991). Resumiendo, de acuerdo con la terminología bibliométrica el documento (B) es el documento citaste y el documento (A) es el documento citado (Diodato 1994).

Una cita representa la decisión tomada por un autor que quiere mostrar la relación existente entre el documento que está escribiendo y el trabajo de otro autor (en un punto concreto) (Sandison, 1989). Así mismo, el análisis de las citas nos daría una muestra de la relación entre autores que es un índice de la extensión de su relación indirecta a través de la literatura.

Varios autores han utilizado diferentes tipos de citas bibliográficas en publicaciones científicas a fin de establecer relaciones entre documentos, p. ej.: la cita directa (la cita de un documento anterior por parte de un documento nuevo), método de referencias bibliográficas comunes, término introducido por Kessler (1963), del MIT que vinculan documentos fuente que están bibliográficamente emparejados y método de citas conjuntas (cocitas) que vinculan documentos citados.

2.2.4.7 FACTOR DE IMPACTO

El Journal Citation Reports (JCR) es uno de los índices que publica el Institute for Scientific Information (SCI; Garfield, 1991), que mide afinidades temáticas, impacto y visibilidad. El factor de impacto (FI) se mide de la siguiente manera:

El factor de impacto de un determinado año es igual al número de citas recibidas durante el año por trabajos publicados en los dos años anteriores dividido por el número de trabajos publicados durante los dos años anteriores.

Inconvenientes del FI procedente del JCR:

- Alto factor de impacto de publicaciones especializadas en revisiones.

- Hábitos de citación variables según los campos científicos (lo cual hace que los factores de impacto no sean comparables entre distintos campos).
- Incluso las publicaciones más prestigiosas pueden publicar algunos artículos mediocres.
- Sólo trabaja con un número reducido de publicaciones.
- Sesgos de tipo idiomático.
- El asociar factor de impacto a calidad puede resultar engañoso, pues el primero se basa en citas anteriores y la segunda es actual.

2.2.4.8 COCITACIÓN

Citación obtenida por dos documentos procedentes del mismo documento citante (Small, 1973). La frecuencia de cocitación es definida como la frecuencia con la que los documentos son citados juntos.

Bellardo (1980-1981), apunta que el análisis de la cocitación constituye un indicador cuantitativo para identificar las relaciones entre las ideas en los trabajos científicos en especial puede utilizarse como herramienta para representar la estructura de la ciencia.

Si se utiliza como una parte del proceso automático de agrupación supone asumir primero que los artículos más citados son los más importantes y en segundo lugar que los artículos cocitados están relacionados por el tema con otro. Cawkell (1976).

2.2.4.9 AUTOCITACIÓN

Según Lawani (1982), mide el número de veces que el autor o la publicación se citan a sí mismos. La autocitación puede medir la continuidad de un autor en las mismas líneas de investigación. En temas novedosos la autocitación es baja, en temas muy especializados alta.

2.2.4.10 INDICE DE INMEDIATEZ

Tiempo que tarda un trabajo científico determinado en ser utilizado en otros trabajos. Sirve para evaluar la importancia de la fuente

aunque tiende a dar valores más bajos cuanto menor es la periodicidad de la publicación.

El índice de inmediatez equivale al número de citas recibidas por los artículos publicados en una publicación durante un año determinado dividido por los artículos publicados en la publicación durante un año determinado.

2.2.4.11 FRENTES DE INVESTIGACIÓN

Son los compuestos por aquellos científicos más significativos dentro de un campo científico, como norma, son los autores que concentran el 50% de las referencias.

2.2.4.12 COLABORACIÓN

Existen beneficios en la colaboración como una correlación positiva entre colaboración (de autores o institucional) y productividad. Además a mayor coautoría habrá más citas. Por otro lado es un indicador de madurez en la investigación y es variable en ciencias experimentales.

2.2.4.12.1 COAUTORÍA

Es la colaboración entre autores. En ésta la colaboración está influida por la disciplina (más alta en ciencia y tecnología), mayor con un tipo de investigación aplicada que con una básica, y variable según el ámbito de publicación. La calidad de las publicaciones elaboradas por múltiples autores suele ser mayor, recibiendo además, mayor número de citas. La media de firmas por trabajo es de 2.5 a 3.5.

Índice de coautoría: promedio de firmas por trabajo.

Grado de colaboración: porcentaje de documentos con autoría múltiple

2.2.4.12.2 COLABORACIÓN ENTRE INSTITUCIONES

La colaboración internacional suele ser de tipo básico y de interés más general. Las colaboraciones nacionales suelen ser de interés local y de tipo aplicado.

Tasa de colaboración: número de instituciones que firman los documentos * 100 / número total de documentos

Índice de colaboración nacional (alta en investigación aplicada; útil si se compara con el índice de colaboración internacional): número de documentos en colaboración nacional/ número total de documentos en colaboración

Índice de colaboración internacional (alta en investigación básica; útil si se compara con el índice de colaboración nacional): número de documentos en colaboración internacional/ número total de documentos en colaboración.

2.2.4.12.3 COLABORACIÓN FINANCIERA

Las decisiones contables para destinar los fondos que se basan en la opinión de expertos pueden no ser las más idóneas si son únicas. Por ello se ha hecho realmente importante el contar con herramientas analíticas adecuadas para fijar los aspectos relevantes de la investigación científica. Además de indicadores de "inversión" basados en datos referidos a gastos, número de investigadores, equipos... se requieren datos como indicadores de la actuación científica en términos de "resultados" y del impacto de la investigación científica. Se ha puesto el énfasis en la estructura y la transferencia de la información, representados como contribuciones manifiestas al conocimiento científico, normalmente bajo la forma de comunicaciones escritas, esto es, documentos de investigación como actas de conferencias y publicaciones en revistas científicas en el caso de ciencia básica y publicaciones de patentes para tecnología.

2.2.5 LIMITACIONES DE LOS INDICADORES BIBLIOMÉTRICOS

A través de los años, se han ido poniendo objeciones a algunas de estas medidas. Rosa Sancho (1990) realizó una revisión de los más clásicos. Desde el argumento de Haitun (Brookes, 1984) de que las ciencias sociales no podían ser tratadas estadísticamente, ya que no cumplen los supuestos que se presupusieron al desarrollar la mayoría de los estadísticos, hasta el ya mencionado de Callon (1993). A continuación se revisan algunos de estos problemas:

La revisión por expertos (peer-review). Entre los puntos objetables a estas revisiones estarían la parcialidad de los científicos que las realizan y su lealtad a campos en decadencia lo que puede suponer un mayor reconocimiento a disciplinas antiguas frente a las nuevas.

El crecimiento de la ciencia es medido por el número de trabajos de investigación publicados; se basa en dos suposiciones que no son ciertas: todo el conocimiento científico se encuentra en esos trabajos y cada trabajo contiene igual proporción de conocimientos. Igualmente hay que considerar la gran motivación para publicar que tienen los científicos del ámbito académico, motivación que no se encuentra entre los investigadores industriales.

El mero cómputo de las publicaciones no proporciona información sobre su calidad, no considera métodos no formales de comunicación, no tiene en cuenta que los hábitos de publicación cambian con el tiempo.

A todo ello habría que añadir los defectos formales de las bases de datos bibliográficas. Habría que normalizar los contenidos de algunos campos documentales.

En relación con el análisis de citas, si bien es cierto que es un indicador útil tampoco está libre de deficiencias, como por ejemplo, el que si bien el impacto de un trabajo demuestra su eficacia y quizá también su valor, la falta de dicho impacto no es necesariamente síntoma de inutilidad, ya que para ser citado un trabajo necesita estar disponible, esto es suficientemente difundido, y ello no tiene que ver con su calidad.

El JCR esta presente en un gran número de trabajos bibliométricos. Por lo que otro punto a tener en cuenta es la alta selectividad del SCI, en su elección de las revistas fuente. Como además éstas cambian continuamente el repertorio no puede ser considerado un

conjunto homogéneo de revistas. Entre las revistas fuente que analiza hay un claro predominio de las anglosajonas, sobre todo norteamericanas. Con ello, artículos no publicados en inglés obtienen menos citas.

Hay grandes diferencias entre disciplinas, subdisciplinas y países. En cuanto a modelos de citación, siempre hay que contar con un elemento de incertidumbre.

De todo ello se deriva que el método más objetivo de valoración, el recuento de publicaciones, es el menos relevante. Siendo el más relevante, el juicio de los expertos, el menos objetivo. Entre estos dos extremos hay un gran número de técnicas bibliométricas que deberían poder analizar la ciencia de manera satisfactoria, pero los actuales indicadores deben ser utilizados con cautela y aplicados a conjuntos homogéneos de científicos trabajando en una misma especialidad.

Ante estos problemas, se desarrolla el método de palabras asociadas. Callon utilizó las palabras asociadas por su mayor aplicabilidad y potencial. Posteriormente, el trabajo de Leydesdorff (1997), demuestra como el método de palabras asociadas no puede ser válido para estudiar la evolución de un tema por la propia dinámica de la ciencia. Según el informe de expertos del Comité de las Naciones Unidas de 1984 (Sancho, 1990) hay una falta de base teórica para el desarrollo y análisis de indicadores que suponen un fuerte inconveniente para la validez de los actuales indicadores.

2.3 LENGUAJES CONTROLADOS

Lenguaje designa todo sistema secundario de signos creado a partir de una lengua. De acuerdo con ello lenguaje documental sería 'todo sistema de signos que permita representar el contenido de los documentos pertinentes en respuesta a consultas que tratan sobre ese contenido' (Van Slype, 1991).

Se han clasificado los lenguajes documentales en dos tipos:

- Lenguajes de indización, o, combinatorios, permiten representar el contenido de los documentos y las consultas de forma analítica;
- Lenguajes de clasificación, representan este contenido de forma sintética.

Dentro de los lenguajes de indización se puede realizar una subdivisión. Esta tipología se basa, en el nivel de normalización de su terminología.

- Lenguajes libres, construidos *a posteriori,* basándose en la indización en lenguaje natural de documentos ya registrados de una colección;
- Lenguajes controlados, construidos *a priori,* antes de empezar a indizar los documentos de una colección;
- Lenguajes codificados.

Normalmente, un vocabulario controlado es más que una simple lista, ya que incorpora alguna forma de estructura semántica. En especial esta estructura se define por:

- Control de sinonimias, eligiendo una forma tipificada que servirá de patrón a todas las demás.
- Distinción de homografías.
- Vinculación de términos con significados más próximos. Hay que diferenciar explícitamente dos tipos de relaciones: las jerárquicas y las no jerárquicas (o asociativas).

Se pueden identificar tres tipos fundamentales de lenguajes controlados: sistemas de clasificación bibliográfica, listas de

encabezados de materia y tesauros. Los tres tipos de lenguajes controlan la sinonimia, hacen distinción entre homografías y términos relacionados, pero los tres utilizan métodos diferentes para conseguir estos mismos fines.

Todos ellos intentan presentar los términos tanto alfabética como sistemáticamente:

Clasificaciones bibliográficas, en forma de índice, la ordenación alfabética es secundaria frente a la jerárquica que es la ordenación principal.

Tesauros, la disposición alfabética de los términos es manifiesta pero es patente la construcción de una estructura jerárquica dentro de la lista alfabética mediante el uso de referencias-cruzadas.

La tradicional lista de encabezados de materia es similar al tesauro en su base alfabética. Pero se diferencia de él al introducir una estructura jerárquica menos desarrollada y al fallar a la hora de establecer una clara distinción entre las relaciones jerárquicas y las asociativas.

2.3.1 LISTA DE ENCABEZADOS DE MATERIAS

Charles Ammi Cutter fue el primero en establecer las normas para la construcción de un encabezado de materias alfabético, la primera edición de su "Rules for a Dictionary Catalogue" aparecieron en 1876. El primer lenguaje controlado para catálogos alfabéticos de materias fue la List of Subject Headings for Use in Dictionary Catalogs, que se publicó en 1895 por la American Library Association.

Una lista de encabezamientos de materias es una lista de términos de indización dispuestos en orden alfabético que pueden ser utilizados en un índice, catálogo o base de datos para describir materias (Rowley, 1992). Estas serían las funciones básicas de una lista de encabezados de materia:

- La lista de términos seleccionados se usarán para un catálogo, un índice o una base de datos y muestran la forma en que deberán mostrarse, en esto actuará como una lista de autoridades, por los términos de indización y por su morfología.

- La lista da recomendaciones sobre el uso de las referencias para mostrar las relaciones en un catálogo, un índice o una base de

datos para servir de guía al usuario entre términos conectados o asociados.

- Una lista de encabezados de materia es ante todo una ayuda para el catalogador o indizador. Mayor información sobre los términos y sus relaciones que pueden ayudar al usuario serán transferidas desde la lista al catálogo o al índice.

Los encabezados de materia normalmente son realizados con un propósito específico. Igual que cualquier otro lenguaje de indización debe reflejar los requerimientos de los usuarios y del lenguaje, será por tanto habitual tener que modificar o actualizar una lista para que esta se adapte a los cambios.

Las principales diferencias entre listas de encabezamientos y tesauros, son:

- En un tesauro aparecerán términos más específicos que los que se encuentran en una lista de encabezados de materia.

- Un tesauro evitará los términos invertidos (por ejemplo, 'Escultura, Alemania').

- Los encabezados de los tesauros no están subdivididos. Por ejemplo: Educación–Bibliografías no aparecerá en un tesauro, siendo estos encabezamientos comunes en las tradicionales listas de encabezamientos de materia.

- Las relaciones mostradas en un tesauro son más específicas que las que aparecen en una lista de encabezados de materia.

- Los diferentes tipos de relaciones entre los términos de un tesauro aparecen explicados usando "Términos Relacionados", "Términos Específicos" y "Términos Genéricos" en lugar del "Ver También" que es el más frecuentemente usado en una lista de encabezamientos de materia, para indicar las relaciones entre términos, sin importar la naturaleza de las mismas.

- Las relaciones entre los términos mostradas en un tesauro no suelen transferirse al índice. El catálogo de un diccionario contiene habitualmente, instrucciones de "ver" y "ver también" que vinculan encabezados relacionados.

Como frecuentemente ocurre, las diferencias entre los encabezamientos de materia y los tesauros no está tan definida

como pueda parecer por las normas indicadas más arriba. Este es el caso del Medical Subject Headings (MeSH). El MeSH, aún siendo normalmente confundido con un tesauro no es sino una lista de encabezamientos de materia (Lancaster, 1991). Contiene varios tipos de campos:

- Sinónimos y cuasi-sinónimos
- Cada término tiene una relación de códigos de elementos relacionados
- Fecha de incorporación al MeSH, y término al que sustituyó

Por otro lado, resulta aún más complicado ver las diferencias entre tesauros y encabezamientos de materia cuando analizamos los complementos del MeSH, por ejemplo:

- Medical Subject Headings Tree Structures (MeSH-Tree Structures), con la estructura jerárquica codificada.
- Medical Subject Headings- Annotated Alphabetic List, es una versión del MeSH ampliada, contiene:
- Subencabezamientos permitidos del término
- Términos relacionados
- Términos utilizados en el pasado
- Definición del término
- Posición en el MeSH-Tree, mediante un código organizado jerárquicamente
- Sinónimos y grafías equivalentes
- Anexos de descriptores geográficos y terminología química

Forsyth and Rada (1986), emplearon el MeSH para crear un tesauro más amplio mediante la fusión con el tesauro SNOMED. Según Frakes (1992), esto es posible gracias a que ambos 'tesauros' tienen una estructura jerárquica.

La siguiente tabla esta extraída del MeSH-Tree y muestra las ramas principales en que se divide:

DESCRIPTOR	POSC.	DESCRIPTOR	POSC.
BODY REGIONS	A01	NUCLEIC ACIDS, NUCLEOTIDES, AND NUCLEOSIDES	D13
MUSCULOSKELETAL SYSTEM	A02	NEUROTRANSMITTERS AND NEUROTRANSMITTER AGENTS	D14
DIGESTIVE SYSTEM	A03	CENTRAL NERVOUS SYSTEM AGENTS	D15
RESPIRATORY SYSTEM	A04	PERIPHERAL NERVOUS SYSTEM AGENTS	D16
UROGENITAL SYSTEM	A05	ANTI-INFLAMMATORY AGENTS, ANTIRHEUMATIC AGENTS, AND INFLAMMATION MEDIATORS	D17
ENDOCRINE SYSTEM	A06	CARDIOVASCULAR AGENTS	D18
CARDIOVASCULAR SYSTEM	A07	HEMATOLOGIC, GASTROINTESTINAL, AND RENAL AGENTS	D19
NERVOUS SYSTEM	A08	ANTI-INFECTIVE AGENTS	D20
SENSE ORGANS	A09	ANTI-ALLERGIC AND RESPIRATORY SYSTEM AGENTS	D21
TISSUES	A10	ANTINEOPLASTIC AND IMMUNOSUPPRESSIVE AGENTS	D22
CELLS	A11	DERMATOLOGIC AGENTS	D23
FLUIDS AND SECRETIONS	A12	IMMUNOLOGIC AND BIOLOGICAL FACTORS	D24
ANIMAL STRUCTURES	A13	BIOMEDICAL AND DENTAL MATERIALS	D25
STOMATOGNATHIC SYSTEM	A14	SPECIALTY CHEMICALS AND PRODUCTS	D26
HEMIC AND IMMUNE SYSTEMS	A15	DIAGNOSIS	E01
EMBRYONIC STRUCTURES	A16	THERAPEUTICS	E02
INVERTEBRATES	B01	ANESTHESIA AND ANALGESIA	E03
VERTEBRATES	B02	SURGICAL PROCEDURES, OPERATIVE	E04
BACTERIA	B03	INVESTIGATIVE TECHNIQUES	E05
VIRUSES	B04	DENTISTRY	E06

ALGAE AND FUNGI	B05	EQUIPMENT AND SUPPLIES	E07
PLANTS	B06	BEHAVIOR AND BEHAVIOR MECHANISMS	F01
ARCHAEA	B07	PSYCHOLOGICAL PHENOMENA AND PROCESSES	F02
BACTERIAL INFECTIONS AND MYCOSES	C01	MENTAL DISORDERS	F03
VIRUS DISEASES	C02	BEHAVIORAL DISCIPLINES AND ACTIVITIES	F04
PARASITIC DISEASES	C03	BIOLOGICAL SCIENCES	G01
NEOPLASMS	C04	HEALTH OCCUPATIONS	G02
MUSCULOSKELETAL DISEASES	C05	ENVIRONMENT AND PUBLIC HEALTH	G03
DIGESTIVE SYSTEM DISEASES	C06	BIOLOGICAL PHENOMENA, CELL PHENOMENA, AND IMMUNITY	G04
STOMATOGNATHIC DISEASES	C07	GENETICS	G05
RESPIRATORY TRACT DISEASES	C08	BIOCHEMICAL PHENOMENA, METABOLISM, AND NUTRITION	G06
OTORHINOLARYNGOLOGIC DISEASES	C09	PHYSIOLOGICAL PROCESSES	G07
NERVOUS SYSTEM DISEASES	C10	REPRODUCTIVE AND URINARY PHYSIOLOGY	G08
EYE DISEASES	C11	CIRCULATORY AND RESPIRATORY PHYSIOLOGY	G09
UROLOGIC AND MALE GENITAL DISEASES	C12	DIGESTIVE, ORAL, AND SKIN PHYSIOLOGY	G10
FEMALE GENITAL DISEASES AND PREGNANCY COMPLICATIONS	C13	MUSCULOSKELETAL, NEURAL, AND OCULAR PHYSIOLOGY	G11
CARDIOVASCULAR DISEASES	C14	CHEMICAL AND PHARMACOLOGIC PHENOMENA	G12
HEMIC AND LYMPHATIC DISEASES	C15	PHYSICAL SCIENCES	H01
NEONATAL DISEASES AND ABNORMALITIES	C16	SOCIAL SCIENCES	I01
SKIN AND CONNECTIVE TISSUE DISEASES	C17	EDUCATION	I02
NUTRITIONAL AND METABOLIC DISEASES	C18	HUMAN ACTIVITIES	I03

ENDOCRINE DISEASES	C19	TECHNOLOGY, INDUSTRY, AND AGRICULTURE	J01
IMMUNOLOGIC DISEASES	C20	FOOD AND BEVERAGES	J02
INJURIES, POISONINGS, AND OCCUPATIONAL DISEASES	C21	HUMANITIES	K01
ANIMAL DISEASES	C22	INFORMATION SCIENCE	L01
SYMPTOMS AND GENERAL PATHOLOGY	C23	PERSONS	M01
INORGANIC CHEMICALS	D01	POPULATION CHARACTERISTICS	N01
ORGANIC CHEMICALS	D02	HEALTH CARE FACILITIES, MANPOWER, AND SERVICES	N02
HETEROCYCLIC COMPOUNDS	D03	HEALTH CARE ECONOMICS AND ORGANIZATIONS	N03
POLYCYCLIC HYDROCARBONS	D04	HEALTH SERVICES ADMINISTRATION	N04
ENVIRONMENTAL POLLUTANTS, NOXAE, AND PESTICIDES	D05	HEALTH CARE QUALITY, ACCESS, AND EVALUATION	N05
HORMONES, HORMONE SUBSTITUTES, AND HORMONE ANTAGONISTS	D06	GEOGRAPHIC LOCATIONS	Z01
REPRODUCTIVE CONTROL AGENTS	D07		
ENZYMES, COENZYMES, AND ENZYME INHIBITORS	D08		
CARBOHYDRATES AND HYPOGLYCEMIC AGENTS	D09		
LIPIDS AND ANTILIPEMIC AGENTS	D10		
GROWTH SUBSTANCES, PIGMENTS, AND VITAMINS	D11		
AMINO ACIDS, PEPTIDES, AND PROTEINS	D12		

Tabla 2. ***Principales temas del MeSH***

2.3.2 TESAURO DE DESCRIPTORES

Se entiende como tal, una lista estructurada de conceptos, que representan de forma unívoca el contenido de los documentos y de las consultas, como en los encabezados los conceptos se extraen de una lista finita elaborada a priori, y sólo estos términos pueden ser utilizados para la indización. Tiene una estructura semántica que viene dada por las relaciones de equivalencia, de jerarquía y de asociación (Van Slype, 1991).

De igual forma, un tesauro se puede definir como un área de conocimiento, actividad, interés o aplicación, que hace referencia a una doble realidad, por un lado, el tesauro como referencia al concepto en sí mismo y, por otro, el análisis que representa el marco que posibilita la clasificación y la recuperación. Algunos autores han presentado técnicas que posibilitan el desarrollo de un análisis del tesauro (DRACO, DARE, FODA...), y que en el caso de Velasco (1998), incluso se presta a la automatización.

Los tesauros representan un conjunto de atributos y relaciones entre conceptos relativos a un particular campo de conocimiento, que puede ser aplicable a los entornos más variados. Los descriptores tienen que estar ordenados de un modo sistemático, según relaciones jerárquicas y asociativas. La relación de los descriptores con los conceptos debe de ser biunívoca. Según la norma ISO (ISO, 1986) los elementos constitutivos de un tesauro son las unidades léxicas y las relaciones semánticas existentes entre esas unidades.

La génesis de un tesauro suele ser a partir de listas de autoridades o listas de términos construidas a para realizar una indización por extracción (p.e. Information Science Abstracts en el año 84 o Sociofile hasta el 86), poco a poco la lista se sistematiza y se construyen jerarquías para mejorar la recuperación (Palma, 1995). Según Lancaster (1991), en la génesis de los tesauros no hay una relación directa con las listas de autoridades.

2.3.2.1 UNIDADES LÉXICAS

Se pueden distinguir las siguientes:

- Títulos, que encabezan los conjuntos, en el interior de los cuales se agrupan los términos.

- Descriptores, que pueden definirse como los términos (palabra o expresión) que se han escogido a partir de un conjunto de sinónimos, cuasisinónimos y términos emparentados para representar, de forma unívoca, un conjunto susceptible de intervenir en los documentos y en las consultas que se realizan dentro de un sistema documental (Slype, 1991). En clasificación automatizada al descriptor más representativo de un grupo (ya sea por mayor número de ocurrencias, aparición en más documentos o por distancia al centroide) se le suele denominar raíz.

Dicho de otro modo, es cuando una o más palabras incluidas en un tesauro son escogidas entre un conjunto de términos equivalentes para representar sin ambigüedad una noción contenida en un documento o en una petición de búsqueda documental.

No descriptores: son palabras o conjunto de palabras incluidas en un tesauro con prohibición de uso y reenvío a uno o más descriptores reutilizables. Se denominan también palabras vacías o antidescriptores.

Términos o descriptores auxiliares: conjunto de descriptores particulares con sentido relativamente poco preciso, o bien listados alfabéticos, sin relaciones semánticas entre sí ni con descriptores pertenecientes a otros grupos.

2.3.2.2 RELACIONES SEMÁNTICAS

La norma ISO (1986) incluye las siguientes relaciones entre los elementos de un tesauro:

Equivalencia de conceptos o sinonimia: Representa la relación asimétrica entre un descriptor y un sinónimo (término menos representativo) que expresa un concepto único o conceptos próximos (Slype, 1991). La relación de sinonimia es una relación de sustitución entre descriptores.

- Jerarquía de conceptos: Esta relación jerárquica puede ser de varios tipos:

- Genérica: Establece una conexión entre una clase o categoría y sus miembros o especies.

- Todo-parte: El nombre de la parte implica en cualquier contexto el nombre del todo al que pertenece.

- Enumerativa: Conexión entre una categoría general de objetos o acontecimientos, expresados mediante un sustantivo común, y un caso individual de tal categoría.

- Polijerarquía: Algunos conceptos pueden pertenecer a más de una categoría al mismo tiempo.

- Asociación de conceptos: Proporciona términos relacionados con un descriptor determinado. La posibilidad de utilizar los términos relacionados con un descriptor en un tesauro permite enriquecer mucho los procesos de recuperación puesto que, al incluirlos, se consigue una ampliación del espacio de búsqueda.

  La asociación es la relación simétrica entre dos descriptores que designan conceptos que, aunque no ligados entre sí por una equivalencia semántica o jerárquica, son susceptibles de evocarse mutuamente, por asociación de ideas dentro de un tesauro monolingüe o de una versión lingüística de un tesauro multilingüe. Pueden existir descriptores relacionados permanente o circunstancialmente, dependiendo del área de conocimiento a la que se adscribe el tesauro.

- Equivalencia lingüística: Traducción del descriptor a otro idioma.

- Cuasirrelación de conceptos: Expresa la relación semántica, de pertenencia a grupo, entre descriptores y temas o facetas, los que pertenecen dichos descriptores.

2.3.2.3 CLASIFICACIÓN FACETADA

La teoría de clasificación facetada del documentalista indio Ranganathan (1967), es aplicable en un principio a un esquema de clasificación universal y utilizada posteriormente como método de clasificación en áreas parciales de conocimiento. La Colon Classification de Ranganathan es uno de las clasificaciones más influyentes. Las cinco facetas con las que trataba eran: la

personalidad (objeto o centro de interés), materia (material, instrumentos o métodos), energía (acciones y procesos), espacio y tiempo. Su objetivo era destinarlo principalmente a optimizar los procesos de búsqueda y recuperación de información.

La clasificación facetada es sintética. Las clases se forman seleccionando términos predefinidos a partir de unas listas facetadas (Maniez, 1993). Los esquemas facetados se pueden configurar en función de las necesidades del usuario, creciendo si los nuevos términos a clasificar no se encuentran en el esquema. La definición de este proceso, realizado manualmente (por el desconocimiento de la ubicación de la nueva clase en el esquema), impide la automatización de la creación de esquemas facetados (Velasco, 1998).

También, está la clasificación bibliográfica de Henry Bliss (Maniez, 1987), que ha sido utilizada para el análisis de dominios de software (Llorens, 1996). La clasificación de Blis es facetada, dividiéndose en producto, tipo, parte, material, propiedad, proceso, operación, agente, lugar y tiempo. Esta buscaba crear un consenso en la clasificación dentro de la comunidad científica. Proponía que era posible identificar y plasmar una estructura básica y permanente dentro de un área. Sin embargo, esta idea de permanencia era equivocada porque, hoy en día, se acepta sin discusión la evolución en el tiempo del conocimiento.

La dependencia de esta técnica del lenguaje hace que, al ser difícil caracterizar cada faceta, si no se selecciona un buen lenguaje controlado asociado a un conjunto de relaciones (principalmente sinonimias) con el resto de los términos que componen el vocabulario de un idioma, sea prácticamente imposible clasificar automáticamente por medio de facetas.

Dependiendo del contexto al que se va a aplicar el tesauro podemos considerar un tipo de descriptores u otro. En las bases de datos de prensa suelen tener varios campos, como se indica en la regla de las seis W (what, why, how, where, who, when). En cierto sentido a estos atributos se les puede considerar facetas. Para construir los encabezados de las noticias, por lo tanto, los descriptores deberán especificar si son de tipo geográfico (p.e., Baratz), onomásticos de personas o lista de etapas históricas y siglos. En realidad, de todas éstas, las únicas facetas universales son tiempo y lugar (Maniez, 1993).

2.3.2.4 EVALUACIÓN DE TESAUROS

Puede realizarse una evaluación superficial de un tesauro simplemente examinándolo. Así se podría ver sus principales características: su ámbito, relaciones, la ambigüedad de los términos, la riqueza de las notas de alcance, etc. Un experto podría realizar una evaluación más concienzuda comprobando si hay varios temas representados y la especificidad de los términos representados. Esta comprobación también puede realizarse tomando una muestra aleatoria de artículos y verificando si las palabras clave aparecen en el tesauro (Lancaster, 1991).

Igualmente se puede comprobar si el tesauro se ajusta a los estándares internacionales en convenciones de plural/singular, formas aceptadas de las palabras, entradas directas y demás materias de consistencia. Asimismo puede considerarse el aspecto estético de la tipografía y la presentación. Todos estos métodos implican una inspección directa y personal por parte del revisor. A principios de los años 70 surgen varios métodos de análisis estadísticos de los tesauros, que hacían factible la automatización de los métodos. En un entorno de generación de relaciones jerárquicas y asociativas estos métodos adquieren un papel prometedor. Este tipo de medida basada en recuentos, ha sido ampliamente desarrollada por el Departamento de Marcel Van Dijk (1976). A continuación se revisan algunos de estos indicadores:

- Kochen and Tagliacozzo (1968) evaluaron varios lenguajes controlados de acuerdo a su ratio de conectividad y a la medida de accesibilidad. La ratio de conectividad es la ratio de coocurrencias (p.ej.: términos relacionados, como mínimo a otro término, por términos genéricos, términos específicos o términos relacionados) con respecto al total de términos en el vocabulario. El grado de accesibilidad es la media de referencias recibidas por un descriptor en un vocabulario. Estas medidas indican la amplitud de la vinculación entre los términos del vocabulario (por ej., referencias cruzadas). Números mayores indicarían mayor operatividad del tesauro.

- Conectividad se define como:

  $(b - a)/b$,

  Donde a es el número de descriptores aislados dentro de un vocabulario (p.e. los que no tienen vinculación con ningún otro) y b es el número total de los descriptores del vocabulario.

Cuanto más se acerque a la unidad mejor será el tesauro. La medida de accesibilidad de Kochen and Tagliacozzo se ha convertido en una ratio de enriquecimiento. Se recomiendan valores entre 2 y 5; demasiadas referencias por descriptor (básicamente más de 5) serían más un obstáculo que una ayuda. Otras medidas recomendables serían:

- La ratio de equivalencia, la ratio entre descriptores y no-descriptores de un vocabulario, una medida de la riqueza del vocabulario de entrada. Los autores del Departamento de Van Dijk recomiendan que este valor supere al 1, lo que indicaría más términos de entrada que descriptores.

- La ratio de reciprocidad, indica la extensión de la reciprocidad existente en las relaciones entre términos genéricos, términos específicos y términos relacionados.

- Ratio de Definición, representada por la ecuación *(b – a)/b*, siendo *a* el número de descriptores que pueden resultar ambiguos por falta de notas de alcance, cualificadores o relaciones de jerarquía que los sitúen dentro del contexto y *b* es el número total de descriptores en el vocabulario.

- Flexibilidad, es la proporción de palabras que aparecen en descriptores compuestos en el vocabulario como descriptores o no-descriptores. Es recomendable un valor de 0'6 o mayor.

- Nivel de precoordinación, es la media de palabras por descriptor. Se recomienda que este valor se mantenga entre 1,5 y 2,0 para tesauros en inglés y francés y entre 1,1 y 1,2 para tesauros en alemán.

- El tamaño de los grupos de términos (p.ej.: los grupos que componen la lista de categorías de un tesauro). Se recomiendan entre 30 y 40 términos por grupo.

- El Grupo de Tecnologías de la Información de la Universidad Carlos III, ha desarrollado medidas de homogeneidad en las relaciones jerárquicas y asociativas de los términos de un dominio. El Coeficiente de especificidad (Velasco, 1998), se establece para cada documento del corpus. Este coeficiente se define como el número de términos existentes en el Tesauro que son más específicos, ya sea en un primer o posterior nivel,

que los descriptores que aparecen en el documento dividido por el número total de descriptores del documento.

$$C_{esp}=(\text{número de términos más específicos del Tesauro})/(\text{número descriptores del artículo})$$

El número de términos más específicos en el Tesauro se calcula mediante el sumatorio de todos los términos específicos que tiene cada uno de los descriptores del artículo.

Este coeficiente intenta medir homogeneidad de jerarquía. Se supone que si los documentos han sido correctamente seleccionados, el tipo de terminología que los componen será similar. De esta forma, existirá homogeneidad entre los documentos, de tal forma que el coeficiente de especificidad tomará un conjunto de valores cuya desviación típica no debe ser alta.

Se establece la siguiente fórmula para contrastar la desviación típica del coeficiente para todo el corpus documental:

$$\frac{\sigma}{C_{especificidad}} <= 0.25$$

Si este cociente toma un valor superior a 0.25 puede decirse que existe dispersión entre los tipos de descriptores (muy generales o muy específicos) que se encuentran juntos en un documento.

Puede crearse también un índice complementario a éste, que mida la generalidad de cada documento.

- Coeficiente de Lloréns (Velasco, 1998), Este coeficiente mide, para cada descriptor y documento pertenecientes al corpus del dominio, la calidad de su relación con la jerarquía del Tesauro. Se calcula en función de la aparición del descriptor en cada documento del corpus y del número de descriptores del Tesauro que aparecen referenciados en el documento.

  Este coeficiente se define mediante la siguiente fórmula:

  - $C_i = C_m - 2(N_d^2 - 2N_d + 1)$

Donde:

C_m es la suma del número de nexos que distancian a cada descriptor del documento con cada uno de los demás descriptores del documento y N_d es el número de descriptores referenciados en el documento.

Este coeficiente C_i toma un valor 0 si todos los descriptores pertenecientes a la jerarquía se encuentran referenciados en el documento y la jerarquía creada tiene sólo dos niveles con una única raíz. Para atenuar el valor en el caso, prácticamente seguro, de que la jerarquía que representa al Tesauro tenga más de dos niveles se divide este coeficiente por una función creciente que toma valores dependiendo del número de niveles de la jerarquía. Así, el coeficiente de Lloréns, para cada descriptor, queda definido como sigue:

$$C_{Lloréns} = C_i / F(\text{número de niveles})$$

F es una función que puede definirse de formas variada. Se ha utilizado en este trabajo la siguiente definición: $F(x) = x^2 / (x+1)$

El coeficiente da medida de la homogeneidad de la jerarquía (relación entre el número de descriptores y el número de niveles).

Si existe mucha dispersión entre los descriptores referenciados en el documento el coeficiente tomará valores altos, lo que se considera que indica poca calidad de la jerarquía. En el caso en el que los descriptores referenciados se encuentren muy próximos en la jerarquía el coeficiente tomará valores bajos, lo que indica mayor calidad.

2.3.2.5 CONSTRUCCIÓN DE TESAUROS

Se pueden definir una serie de pasos que nos indiquen una pauta a seguir para la construcción automática o manual de un tesauro (Frakes, 1992).

1. Definición del límite del tema a tratar, que en la construcción automática, coincide con el área de conocimiento contenida en nuestro corpus. La definición de los límites pasa por identificar el tema central y los periféricos ya que no todos los temas incluidos tienen la misma importancia. Así, el dominio quedará estructurado en divisiones o subáreas.

2. Una vez establecidos el dominio y sus subáreas deberán identificarse las características que se desean para el tesauro. Los tesauros manuales son más complejos estructuralmente que los automáticos.

3. Selección de términos para cada subárea. Deben usarse variedad de fuentes, desde índices hasta catálogos, así como algún tesauro importante ya existente. Identificado el lenguaje de partida cada término debe analizarse identificando: sinónimos, términos genéricos y específicos y a veces, definiciones y notas de alcance. Estos términos y sus relaciones deben organizarse en estructuras jerárquicas.

   En el caso de la generación automática, este paso, se puede realizar bien por estos medios de fusión de tesauros y catálogos, o bien, mediante filtros y clasificadores. En resumen, las opciones más difundidas son:

   - A partir de una colección de documentos.

     Se parte de un cuerpo de texto representativo. Se aplicarán técnicas estadísticas para identificar los términos relevantes y sus principales relaciones. Lo que se pretende al utilizar unos algoritmos estadísticos es la identificación de la estructura semántica de un tesauro. Reproduciendo así estructuras de sinonimia y jerarquía mediante relaciones significativas entre los términos. Sobre que clasificadores se pueden aplicar en este punto y como trabajan se hablará en la sección de este capítulo de clasificación.

   - A partir de la fusión de tesauros existentes.

     Este método ha sido estudiado con detenimiento por Forsyth y Rada (1986) en el ya descrito de fusión entre el MeSH y SNOMED.

   - Tesauro generado por usuarios.

     Se pretende utilizar todo el conocimiento de los usuarios. Se parte de la idea de que los usuarios de sistemas de recuperación de información conocen y utilizan muchos términos relacionados en sus estrategias de búsqueda mucho antes de que éstos se encuentren con un tesauro.

Esta sería la base de TEGEN- sistema de generación de tesauro diseñado por Guntzer (1988), que propone este sistema como una alternativa viable para la construcción automática de tesauros. El procedimiento consistiría en examinar los operadores booleanos entre términos de búsqueda y las rectificaciones a búsquedas infructuosas para generar así términos relacionados y sinonimias. Así, la experiencia de los usuarios se utiliza para solventar ambigüedades e incertidumbres en el tesauro.

4. Este proceso de organización del lenguaje puede mostrar fallos que llevarán a la inclusión de nuevos términos o mostrar la necesidad de incluir nuevos niveles en las jerarquías.

5. Finalizada la organización inicial el tesauro en su totalidad debe ser revisado para comprobar su consistencia. Para esto, son útiles los mecanismos expuestos en la sección precedente.

6. El mantenimiento del tesauro implementado sirve para asegurar su viabilidad y efectividad. Se ha de tener en cuenta que un tesauro es un sistema vivo cuyos términos reflejan conceptos nuevos que aparecen, se relacionan, evolucionan a otros nuevos y desaparecen (Llorens, 1998). El tesauro debe reflejar, a lo largo del tiempo, todos los cambios en la terminología del área.

   Cuando evoluciona un término se deben de reindizar los documentos antiguos de forma acorde a este término. Las actualizaciones en los sistemas no automáticos son tradicionalmente lentas y requieren de personal numeroso para las revisiones y modificaciones tanto del lenguaje como de las relaciones. Por tanto, típicamente un tesauro evoluciona lentamente. Este no es el caso de las clasificaciones automáticas donde sistemas como el índice de transformación de Callon (1995), permiten monitorizar la evolución de un tema. Por otro lado existen sistemas que nos orientan sobre la posición previsible de nuevos términos (Llorens, 1998; Uramoto, 1996).

2.4 INDIZACIÓN

Los sistemas actuales de clasificación automática de información se basan en la utilización de técnicas que permiten enumerar los conceptos sobre los que trata un documento y representarlos por medio de un lenguaje combinatorio. A este proceso se le denomina indización de información.

Según Codina (1994), el proceso de indización puede realizarse de varias maneras, la primera, mediante extracción directa de los términos del documento. La segunda, mediante asignación, este método, según Albrechten (1993), consiste en interpretar semánticamente el documento, aislar conceptos relevantes y traducirlos a descriptores mediante un tesauro, ya sea de manera manual o automática. Se puede añadir un tercer tipo, enfocado a las necesidades del usuario (Soergel, 1985) por el que se indizaría de acuerdo a las necesidades y vocabulario de los usuarios potenciales.

2.4.1 FACTORES QUE AFECTAN A LA CALIDAD DE LA INDIZACIÓN

Según Fidel (1994), los problemas que afectan a la indización humana son su coste (de aprendizaje y laborales), rigidez, lentitud y sus variaciones en la calidad. Las variaciones de calidad pueden ser debidas a varios factores desde la inconsistencia en la indización a los meros errores en la indización (Hlava, 1992).

Existen varios trabajos sobre los indicadores que afectan a la calidad de la indización, como las recopilaciones realizadas por Pinto (1994) o la de García-Marco (1995). Algunos de estos factores son:

2.4.1.1 INCONSISTENCIA

La inconsistencia en la indización es un problema derivado de la diferente asignación de descriptores para un mismo documento por parte de distintos indizadores (consistencia externa) o por el mismo indizador en dos momentos diferentes (consistencia interna). Un problema asociado con la evolución del lenguaje es la consistencia histórica. Las posibles razones por las que se produce inconsistencia, son:

- Evolución histórica de los descriptores y de los términos de los documentos. La aparición de un nuevo descriptor debe de implicar la revisión de los documentos afectados por la nueva incorporación (Codina, 1993).

- Diferente apreciación por parte de cada indizador de cuales son los conceptos más relevantes del documento

- Distintos descriptores para un mismo concepto. Los lenguajes documentales deben expresar siempre el mismo concepto con el mismo descriptor. La relación entre descriptor y concepto debe de ser siempre biunívoca, para no tener problemas de inconsistencia.

- Desconocimiento de los descriptores de la base de datos por parte del indizador.

- Distinta apreciación por el indizador de cuáles van a ser las necesidades y vocabulario de búsqueda de los usuarios de la base de datos.

2.4.1.2 EXHAUSTIVIDAD Y RELEVANCIA

La indización se debe regir por el principio de exhaustividad. Se define como el número de conceptos del documento representados. La exhaustividad esta directamente relacionada con el número de nociones que caracterizan al documento. Por lo tanto, el número de descriptores no debe de estar limitado de forma arbitraria, aunque en ocasiones se ha fijado el valor de entre 8 y 12 descriptores (Slype, 1991).

En una indización ideal si existen n temas relevantes en un documento deben de existir n descriptores. La relevancia se estima según el tema la tipología documental y su valor potencial para la recuperación. La relevancia es un indicador pragmático.

No todos los descriptores que se pueden aplicar a un corpus determinado tienen el mismo poder de discriminación. El poder de discriminación de un termino es una medida del poder que tiene éste para diferenciar documentos relevantes de los que no lo son. Normalmente, se utiliza la siguiente fórmula para poder evaluar la discriminación del término A (Chu, 1989):

A= (nº de documentos con el descriptor A)/(nº de documentos de la base de datos).

Una buena discriminación estaría en torno al 0.05, valores más bajos indican una discriminación más fina y superiores sería demasiado vasta. Bastante asociado a este concepto se encuentra el de IDF del que se hablará más adelante.

Por profundidad se entiende el nivel hasta el que se ha descendido en la representación del documento o en el corpus. En ocasiones se ha medido como una relación entre el número de palabras del texto y el número de palabras representadas. También se ha medido mediante conceptos jerarquizados. En función de la especificidad en la jerarquía se puede caracterizar el documento.

Estos indicadores, junto a otros como la exactitud, se sitúan en un nivel conceptual.

2.4.1.3 PERTINENCIA

Grado en que los términos de la descripción coinciden con las necesidades de los usuarios. Es un indicador en el ámbito pragmático: está directamente relacionado con la recuperación de documentos pertinentes.

2.4.1.4 EXACTITUD

Relación entre los conceptos indizados y los que están en el documento.

2.4.1.5 NATURALIDAD

Relación entre los términos que utilizan los usuarios y los términos de indización.

2.4.1.6 DENSIDAD

Porcentaje de palabras no vacías en el resumen.

2.4.1.7 LEGIBILIDAD

Puede ser física (en un nivel señalizante), pero también, como se utiliza en este estudio, en su acepción de comprensibilidad de un texto (en un nivel semántico).

2.4.1.8 ESPECIFICIDAD

El principio de especificidad establece que los descriptores se deben de situar en el mismo nivel de especificidad (o generalidad) que los conceptos del documento. En principio, el término debe de ser lo más específico posible (UNE 50-121-91). Un descriptor muy específico daría muy pocos documentos, uno muy general demasiados. Opera a nivel semántico.

2.4.1.9 POSTCOORDINACION

Grado de conexión entre los descriptores en el ámbito significante. Se mide mediante la ratio de precoordinación. La ratio de precoordinación es la división entre el número de descriptores y el número de palabras de nuestro vocabulario (Slype, 1991). Lo recomendado es entre 1,5 y 2. Los descriptores deberían de ser más precoordinados en el dominio central del corpus documental, de otro modo los unitérminos correspondientes al dominio serían utilizados casi siempre, haciendo inservible la indización. Para recuperar información un sistema precoordinado es más flexible, pero deja abierta la posibilidad de falsas combinaciones entre términos (Codina, 1993).

2.4.2 INDIZACIÓN AUTOMÁTICA

La indización automática de información es la operación que consiste en que un ordenador reconozca los términos que figuran dentro del título, del resumen, del texto completo (sí éste ha sido almacenado junto con la descripción documental) y a veces también dentro de la indización humana. Posteriormente el proceso emplea estos términos, bien tal cual, o bien después de transformarlos en otros términos, equivalentes o conceptualmente próximos, con el fin de convertirlos en elementos que se incorporan al archivo de búsqueda y quedan disponibles para recuperar el documento (Slype, 1991).

Como resultado de la ejecución del proceso de indización de un documento se consigue un conjunto de referencias (relación entre descriptores y documentos). Las referencias encontradas durante el proceso de indización de la información contenida en un documento se utilizan posteriormente para la realización de la búsqueda documental. Este proceso pretende localizar un conjunto de documentos que reflejan el resultado de una búsqueda sobre dichas referencias. El usuario debe crear una consulta, bien mediante una

sintaxis estructurada (por ejemplo, SQL) o bien mediante lenguaje natural, que se traduce a una expresión compuesta exclusivamente por descriptores del tesauro y un conjunto de operadores. Los operadores aceptados son aquellos que permiten implantar las relaciones sintácticas permitidas por el sistema: consultas booleanas, búsqueda por proximidad, búsqueda por contexto, etc.

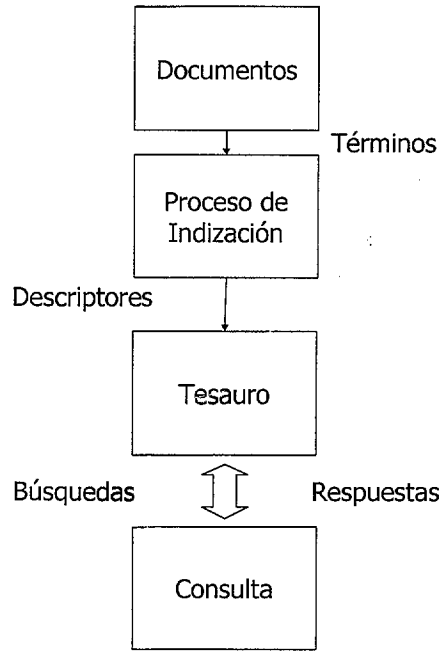


Figura 2: Proceso de indización

A los sistemas automáticos siempre se les ha achacado el operar tan solo en campos muy restringidos. Pero, sistemas de indización como AIPIA (dentro del programa de recuperación MORPHS) que empezó en la industria del plástico, hoy se utiliza desde la industria del aceite hasta en sistemas de información para el turismo (Jones, 1992). El sistema de información del API (American Petroleum Institute) se aplicó, desde los años 70, a empresas petroquímicas, hoy en día, una versión actualizada por Access Innovations, Inc. (MAI) es utilizada por gran número de industrias, pasando por la unión de consumidores norteamericana o el Getty Conservation Institute (Hlava, 1992). Si bien la adaptación de los términos de un campo al programa es un trabajo costoso, el programa en si es muy flexible.

Se han publicado gran cantidad de métodos de selección y valoración de descriptores extraídos de manera automática. Los primeros trabajos de Luhn (1957) recomendaban que las cadenas encontradas repetidamente en un documento eran las más

apropiadas para indizar y basar los pesos de los términos en este cálculo. Carrol y Roeloffs (1969) utilizaron las desviaciones estándar de los cálculos observados frente a los cálculos predictivos. Sparck Jones (1973) argumentó que las palabras que se encontraban en el documento que se estaba estudiando, y que por el contrario era difícil encontrarlas en otros documentos, eran palabras importantes, y así desarrolló la inversa de la frecuencia documental como el peso de cada término. Salton (1973) combinó la inversa de la frecuencia documental con la frecuencia en el documento tomando su producto como una medida de la importancia del término.

Resumiendo, ha habido diferentes métodos automáticos que observan las estadísticas de los términos sobre un texto y, basados en estas estadísticas, obtienen descriptores. Un completo estudio, para la comparación de funciones para el cálculo de los pesos de los términos, fue realizado por Ro (1988).

A pesar de la variedad de estadísticas en las valoraciones, parece haber un acuerdo sobre que un buen descriptor es, entre otras cosas, uno cuya frecuencia de ocurrencia en el documento bajo estudio no se puede predecir observando los restantes del corpus.

2.4.2.1 ANÁLISIS LÉXICO

El análisis léxico es un paso previo a la indización automática. La decisión de lo que constituye una palabra, aunque trivial en la indización manual, es importante en la indización automática. A continuación se exponen algunos de los puntos más importantes que se deben discutir antes de empezar a indizar (Frakes, 1992; Harman, 1994):

- Guiones, éstos pueden ser o bien una división de una palabra por salto de línea o bien por ser parte de una palabra (p.e. MS-DOS) o para querer recalcar la unión entre términos (p.e. State-of-the-Art).

- Números, no se utilizan para indizar salvo campos técnicos pero sólo en combinación con letras (p.e. vitamina B12).

- Mayúsculas, los descriptores normalmente están todos en minúsculas o en mayúsculas. Para identificar títulos o nombres propios es importante que sepa distinguir entre mayúsculas y minúsculas.

- Caracteres simples, frecuentes como iniciales de nombres o etiquetas de apartados en un texto, no se suelen tener en cuenta. Esto da problemas en los casos en que sí que tienen mayor semántica, por ejemplo, vitamina C o rayos X.

- Palabras que incluyen signos de puntuación, como palabras técnicas del tipo command.com u OS/2.

2.4.2.2 LISTAS DE PALABRAS VACIAS

Son términos muy frecuentes en cualquier documento de la colección y por tanto con muy poco poder de discriminación. Mediante su eliminación se consigue velocidad de procesamiento y ahorro de espacio en las bases de datos, sin perjudicar la efectividad del sistema. La lista de palabras vacías, las que aparecen en la mayoría de los documentos de la colección, varía con la cobertura documental de la base de datos que estemos indizando.

El número de palabras vacías en sistemas comerciales tiene una gran variación, desde las ocho que tiene ORBIT, las 418 de SMART o las 425 de la colección Brown.

2.4.2.3 TÉCNICAS PARA MEJORAR LA INDIZACIÓN AUTOMÁTICA. FILTRADO

El filtrado es un paso previo a la indización automática. Con esto se consiguen varios fines: el primero es reducir el tiempo dedicado a la indización, mediante la reducción del texto. Además, se utiliza para estandarizar los términos simplificando las posibles búsquedas que se realicen posteriormente. En el caso de querer realizar, posteriormente, una clasificación es mucho más necesario. Por ejemplo, mediante el método de palabras asociadas (Callon, 1995), el número de relaciones posibles entre los términos sin filtrar harían inmanejable cualquier resultado. Los métodos tradicionales pasan por el filtrado manual, normalización a mayúsculas (o minúsculas), supresión de los espacios en descriptores coordinados o la supresión automática de las palabras vacías (Frakes, 1992). Otros métodos que se han propuesto en la literatura son:

- IDF (Inverse Document Frecuency) (Cleveland, 1990; Harman, 1994). Basado en las ideas de Zipf y en la asignación de pesos.

- Transformación de los términos a una variante cuasi-homófona (Knuth, 1973). Este método se ha empleado en este estudio, mediante los algoritmos Soundex y Metaphone, para los títulos de las revistas y la lista de descriptores. El fin de estos algoritmos es evitar duplicaciones ocasionadas por distintas grafías, ampliando innecesariamente el vocabulario. Mediante un estudio fonológico se desarrollan un conjunto de reglas que permiten traducir a un código cada término (Lawrence, 1990).

- Algoritmos de normalización y stemmers.

- Extracción automática de palabras clave (p.e. n-grams)

2.4.2.3.1 FRECUENCIA DE LOS TÉRMINOS DE ZIPF

La mayoría de las técnicas de indización automática se basan en las frecuencias de aparición de palabras en el texto y en sus coocurrencias (Cleveland, 1990). En 1949, Zipf publicó un libro llamado 'Human Behaviour and the Principle of Least Efford' en el que exponía su idea de que en el lenguaje se busca siempre utilizar el menor número de palabras posible, sin pérdida de información. Zipf, ordenó las palabras por orden de frecuencia, observando la siguiente relación:

Rango x Frecuencia= constante

Donde el rango se corresponde con el número del orden, el término más frecuente es el que tiene rango uno.

Por ejemplo, en el "Ulysses" de Joyce hay:

16432	palabras utilizadas		una sola vez
4776	"	"	dos veces
2194	"	"	tres veces

Así, unas pocas palabras se repiten pocas veces y mientras van creciendo las palabras en número, aumenta el número de veces en que se repite en el texto, de manera que si se multiplica el número de orden de una palabra, según su frecuencia, por las veces que aparece ésta, resulta un número constante, por ejemplo, en este mismo libro:

la palabra nº

10 se usa	2653 veces	=> 10 * 2653	=26530
100	265	=> 100 * 265	=26500
1000	26	=> 1000 * 26	=26000

Para los términos con menor frecuencia, la relación era:

$$I_1 / I_n = (4n^2 - 1)/3,$$

donde I_1 es el número total de palabras con frecuencia igual a uno e I_n con la que tiene frecuencia n.

Booth, tras observaciones empíricas, modificó esta fórmula:

$$I_1 / I_n = n(n+1)/2$$

Goffman, postuló que las palabras de mayor frecuencia seguían las pautas marcadas por Zipf, teniendo una función predominantemente funcional, mientras que las de menor frecuencia reflejaban el estilo de autor. La zona intermedia, sería la que contendría la semántica del documento. Este punto de transición entre la zona intermedia y la de aparición igual a uno lo calculó con la siguiente función:

$$n = (-1 + \sqrt{1 + 8I_1})/2,$$

Para suprimir las palabras de mayor ocurrencia pero con baja semántica se propuso crear una lista de palabras vacías.

Según otros autores, como Willett, son las palabras con bajas frecuencias las que dan más información. Las de frecuencias superiores, y que no son vacías, están relacionadas con el campo en cuestión, por ejemplo, en química: experimento, laboratorio, preparación, ... Son palabras con poca información (Jones, 1992).

Para mejorar este método se han propuesto la utilización de tesauros, asignación de pesos por situación en el documento, stemmers, aparición de términos compuestos y medidas de concurrencia de términos.

2.4.2.3.2 ASIGNACIÓN DE PESOS A LA INDIZACIÓN

2.4.2.3.2.1 INVERTED DOCUMENT FREQUENCY

El Método IDF para ponderación de descriptores fue definido por Spark Jones el 1972, midiendo la frecuencia relativa de un término respecto al conjunto de documentos de la colección (Cleveland, 1990; Muñoz, 1994). El IDF es una medida de la eficacia de un descriptor en una búsqueda basada en el poder de resolución. El poder de resolución de una palabra indica la importancia de esa palabra como término clave. El método IDF nos indica cual es el poder de resolución de un término de búsqueda o de clasificación. Cuando un término aparece en la mayoría de los documentos de la colección su poder de resolución disminuye. Por ejemplo, en una base de datos de medicina el término 'medicamento' nos devolvería un número desproporcionado e inmanejable de documentos, mientras que el término 'indinavir' es de esperar que no estuviese presente en la mayor parte de los documentos de la base de datos. La expresión de este método, es:

$$a_{ij} = \log_2(N / n_j)$$

Donde:

n_j = número de documentos del corpus en los que aparece el término t_j

n = número total de documentos del corpus

y tomando la representación vectorial del documento D_i como (a_{i1}, a_{i2}, …, a_{in}) se toma

Con esta definición para los pesos se cumple lo anterior: si un término aparece en muchos documentos, n_j toma un valor alto, y el peso a_{ij} toma un valor pequeño, y al revés, cuantas más veces aparezca un término en un documento mayor será su peso en él.

2.4.2.3.2.2 FRECUENCIA EN EL DOCUMENTO

Mejora los resultados, sobre todo en combinación con el IDF. Compensa las frecuencias altas del término y la longitud del documento (Harman, 1994).

$$cfreq_{ij} = \log_2(freq_{ij} + 1) \Big/ \log_2 length_j$$

donde,

$freq_{ij}$, es la frecuencia de términos i en documento j

$length_j$, el número de términos únicos en el documento j

Si se calcula el peso combinado del término i en el documento j, basado en el producto de frecuencia del término e IDF:

$$d_{ij}=tf_{ij} \times \log(N/df_j)$$

Tfij, número de ocurrencias del término t_j en el documento

Dfj, numero de documentos en el corpus con el término

N, número total de documentos en el corpus

Aplicado a multitérminos se asignan mayores pesos, ya que a una palabra aislada tiene menor semántica. Si un término aparece en todos los documentos el resultado de la fórmula será menor, cuantas más veces aparezca el término mayor valor.

2.4.2.3.3 NORMALIZACIÓN DE TÉRMINOS

Son sistemas que permiten relacionar términos morfológicamente emparentados (Frakes, 1992). El objetivo es mejorar la efectividad en la recuperación y disminuir el tamaño de los ficheros.

El problema de los stemmers es que, o bien hace que se pierda toda la información sobre las variantes morfológicas o bien hace que haya que duplicar el tamaño del archivo al guardar por un lado la forma normalizada y por otro la forma sin normalizar. El stemming pueden dar lugar a dos errores: Intentar suprimir una secuencia demasiado larga, emparentando términos no relacionados. También, puede ocurrir lo contrario, cuando el stemmer actúa por debajo del número de caracteres correcto, produciendo que documentos pertinentes no sean recuperados.

La aproximación que realiza WordNet (del que ya se habló en la sección de lingüística documental), utiliza una lista de 15 reglas, que recogen las terminaciones más frecuentes en inglés y, en su caso, las terminaciones que normalizarían el término. Por ejemplo, -ing por nada o por -e; o s por nada, etc. Para los términos irregulares (verbos irregulares, plurales irregulares, etc.) tiene unos ficheros en los que se buscaría el término y se sustituiría por el término normalizado.

Según Frakes (1992) los distintos tipos de stemmers son:

- Diversidad de la sucesión, de Hafer, que calcula el número de caracteres opcionales que hay ente los términos de un corpus, y elige el término más probable entre los sustitutos.
- Eliminación de afijos (Harman, 1996):
- Eliminación de plurales (Harman, 1991)
- Lovins, se suprimen, iterativamente, un conjunto de cadenas de caracteres según un grupo de reglas. Tiene un total de 260 sufijos.
- Porter, tiene un conjunto de reglas que sólo operan en determinadas condiciones (a nivel de raíz, de sufijo y de aplicación de la regla). Sólo tiene 60 sufijos. Cuando se cumple la condición no sólo se suprime el sufijo, sino que se suele sustituir por otra cadena. Por lo que la forma normalizada es

algo mayor que la producida por Lovins pero menor que la de eliminación de plurales.

OKAPI: suprime los plurales, terminaciones en 'ed' e 'ing'.

- N-gramms

2.4.2.3.4 N-GRAMS

2.4.2.3.4.1

N-grams (Cohen, 1995). El algoritmo realiza un filtrado estadístico calculando las ocurrencias de grupos de caracteres (grams). El número de caracteres nos da el tamaño del gram. La fórmula por la que se evalúan los grams es (Merino, 1999):

$$y_i = \begin{cases} C_i \ln(C_i/S) + B_i \ln(B_i/R) - (C_i + B_i)\ln[(C_i + B_i)/(S+R)], & RC_i \geq SB_i \\ 0, & RC_i < SB_i \end{cases}$$

donde:

C_i sería el valor del n-gram en el documento.

Bi representa el valor del n-gram en el background

S es el valor para el conjunto de n-grams en el documento

R es el valor del conjunto de n-grams en el background

Este proceso de filtrado selecciona los términos cuyas ocurrencias son intermedias en la colección.

El método fue utilizado ya en 1951 para optimizar el tiempo de búsqueda (Shanon, 1951). El N-Grams precisa conocer cómo localizar las palabras, en concreto, sus delimitadores. La selección de los descriptores se basa únicamente en observaciones del corpus. En la práctica, aunque las listas no llegan a ser exhaustivas, se consiguen buenos resultados. Aunque se ha llamado la atención al hecho de que en este proceso de filtrado se seleccionan los términos cuyas ocurrencias tienen valores intermedios en la colección:

- Aparte de su sencillez, es un método muy flexible, algunos algoritmos se basan en la palabra y consideran la posición n-gram, otros no consideran los n-grams que existen entre las palabras. Y otros sólo cuentan los n-grams que aparecen y no sus ocurrencias.

- Por su concepción los n-grams hacen innecesaria la utilización de algoritmos de normalización de términos.

- El N-Grams es, en principio, independiente del lenguaje, de hecho, Cohen (1995) probó el algoritmo, sin ninguna modificación, para el inglés, español, finés, alemán y japonés.

Cohen (1995), expuso una larga serie de aplicaciones entre las que destacó:

- Relacionados con la ortografía
- Corrección código Morse
- Errores de mecanografía
- Detección y corrección de los errores de ortografía de OCR Las predicciones se llevan a cabo por las estadísticas de n-grams.
- Ayuda para mecanografía.
- Análisis Lingüístico
- Predicciones acerca de categorías de palabras inglesas.
  - Reconocimiento de palabras
  - N-grams de palabras (no de caracteres) han sido aplicados en Pietra (1992).

Reconocimiento de palabras por modelos de n-grams de fonemas ha sido estudiado en Yannakoudaris (1992).

Este método se basa en representar el texto mediante la enumeración de n-grams, o cadenas de caracteres de tamaño n. Fijando un entero n, los n-grams representan en un texto agrupaciones de caracteres adyacentes de longitud n. En otras palabras, existe una ventana de longitud n que se desliza a lo largo del texto, moviéndose un carácter cada vez; en cada posición de la ventana, la secuencia de caracteres dentro de ella compone un n-gram. El documento se representa entonces por un vector compuesto por las diferentes secuencias del texto y el número de veces que se observan (Merino, 1999).

2.4.2.3.4.2 PASOS PARA REALIZAR EL NGRAMS

- Conseguir un corpus de comparación (background). El background es un texto, que sirve para conocer cuales son las palabras vacías en el idioma que utilizamos. El background debe ser lo más diferente posible a los documentos que queremos

analizar. Por ejemplo, un cuento clásico en principio compartiría pocos descriptores con un documento de astrofísica. En principio, los términos más comunes serían las palabras vacías más típicas (pero incluyendo formas derivadas y flexionadas). Un background mal seleccionado implica que el filtro no funcionará bien.

- Se debe filtrar el documento, eliminando los caracteres extraños así como los signos de puntuación y los números. Generalmente, los signos de puntuación y los números se sustituyen por separadores rodeados por espacios (en posteriores etapas, n-grams que contienen separadores se ignorarán). Insertar espacios a los lados de un separador sirve para asegurar que palabras que aparezcan con signos de puntuación al lado tendrán el mismo tratamiento que las que no los tengan. Siguiendo esto, espacios consecutivos se reemplazan por un único espacio y las letras minúsculas se reemplazan por mayúsculas.

- Se forman las ocurrencias de los n-grams. Se considera un texto de ejemplo de longitud S caracteres, representado según los símbolos s_1, s_2 ,..., s_s. Fijando n como un entero positivo, se define el j-ésimo n-gram g_j como la subsecuencia de texto centrado en el carácter j-ésimo:

$$g_j = (s_{j-(n-1)/2}, s_{j-(n-1)/2+1}, \ldots, s_{j-(n-1)/2+n-1})$$

- Para cada n, se pueden tomar caracteres hacia arriba o hacia abajo. El valor de n, generalmente, suele estimarse entre 3 y 6, aunque este valor sí puede considerarse dependiente del lenguaje, aunque suelen dar mejores resultados los valores impares de n.

  En esta etapa del proceso se forman los n-grams, almacenando cada n-gram que se observa e incrementando apropiadamente sus ocurrencias. Si en el n-gram se encuentra un separador, no se tiene en consideración en los cálculos

- Para cada n-gram encontrado en el documento, su resultado se compara con el resultado en el background, obteniéndose una puntuación. Puntuaciones altas se corresponden a palabras innovadoras.

- Para cada carácter del documento, se calcula una puntuación basada en los n-gram que forma.

- Se calcula una puntuación umbral de carácter. La capacidad de establecer un umbral para cada valor de n-gram es crucial para la selección de descriptores sin la necesidad de eliminar las palabras separadoras. Esto se consigue mediante la comparación con el background.

- Cada palabra del documento que contiene un carácter que excede el umbral se extrae. Las palabras que se extraen contiguas forman frases (ver el siguiente apartado)

- Cada vez que una palabra o frase se selecciona, el término se introduce en una lista de términos y así se va acumulando el peso de cada término en la lista. El peso final es simplemente la suma de los pesos de sus diferentes instancias.

- Se incluyen en una lista todas las palabras y frases seleccionadas. Las puntuaciones de cada ocurrencia se suman para obtener un peso total para cada elemento de la lista.

- Finalmente se ordena la lista resultante por puntuación, obteniéndose así la lista de descriptores.

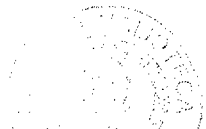
De una manera similar, pero variando el umbral, es posible definir palabras clave para un conjunto de documentos.

2.4.2.3.4.2.1 SELECCIÓN DE DESCRIPTORES SIGNIFICATIVOS

Tras haber hallado el peso de cada n-gram, a cada byte del texto (filtrado) se le debe asignar un valor. Las palabras que contienen letras con pesos altos serán seleccionadas para continuar examinándolas en etapas sucesivas. El método le da al j-ésimo carácter el valor del n-gram g_j, valor que afecta únicamente al carácter del centro del n-gram. Esto significa que los n-grams con un peso elevado le proporcionarán su peso únicamente a las palabras que contengan el centro del n-gram. Se da esta situación teniendo en cuenta que no es demasiado grande. Para valores de n grandes, se puede distribuir el peso del n-gram entre los caracteres de éste, quizá dando un mayor peso a aquellos caracteres que se encuentren más cerca del centro. La asignación de los pesos de los caracteres requiere una segunda pasada a lo largo del texto filtrado.

Cuando un n-gram se considera importante, la palabra que lo contenga también lo será. Así, se tomará como significativa aquella palabra que contenga, al menos, un carácter cuyo peso sea igual, o mayor que el umbral de caracteres. Si se reconoce una palabra como significativa, se seleccionará para ser tratada como descriptor.

Si un n-gram significativo abarca dos palabras, la combinación de esas dos palabras también se tomará como significativa. Dos palabras contiguas se concatenan para formar una frase, incluyendo el delimitador común, si las letras de los lados del delimitador contribuyen a formar un n-gram significativo. Esto se determina examinando los pesos de los caracteres que resultan de los dos n-grams que incluyen el separador y los dos bytes de cada lado del separador.



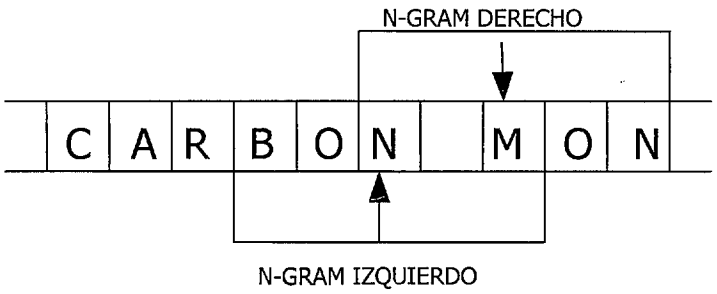


Figura 3: Figura N-grams entre términos

Este dibujo muestra los 5-grams derecho e izquierdo de un espacio que contiene el delimitador y al menos un carácter de cada uno de sus lados. Existen tres 5-grams semejantes, y sus pesos se centrarán en los caracteres N, espacio, y M. Así pues, para determinar si estas palabras pueden formar una frase, se deben examinar los pesos de los caracteres N, espacio y M. Si cualquiera de ellos tiene un peso igual, o mayor, que el umbral de caracteres τ, se selecciona la concatenación de los dos o más términos como descriptor.

La figura muestra los pesos de los caracteres de una porción de un texto ("carbon monoxide" con espacios a los lados). El dibujo también nos muestra su correspondiente umbral. Si el espacio entre "carbon" y "monoxide" tiene un peso mayor que el umbral, la frase se toma como un descriptor único. Realmente, si cualquiera de los pesos del espacio, N o M excede el umbral, la frase se toma en conjunto. En este ejemplo, los descriptores que se encontraron fueron, por orden, "monóxido de carbón", "arritmias", "arritmia", "ejercicio", "niveles", "nivel" y "rítmica".

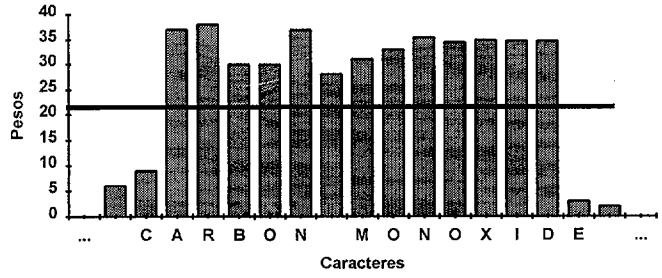


Figura 4: Ejemplo de pesos de n-grams

En este ejemplo el umbral está representado por la línea horizontal y se corresponde con dos desviaciones estándar. Cualquier peso generado por un n-gram que contenga un espacio se estudiará con mayor profundidad para saber si el blanco debería formar una frase. Para los 5-grams, los valores de N, espacio y M serán los examinados. En este caso los tres exceden el umbral.

2.5 EL ANÁLISIS DE DOMINIOS EN CIENCIOMETRÍA

Durante los últimos 15 años se ha utilizado el análisis de dominios para la reutilización de software, apareciendo nuevos desarrollos actualmente. Un dominio, en un contexto amplio, es una esfera o actividad de interés. El análisis de dominios es el proceso en el cual la información y el conocimiento para desarrollar determinado software es identificado y organizado para constituir, así, una representación del dominio. Esta representación del dominio puede ser reutilizada en futuros proyectos que compartan características comunes.

Prieto Díaz (1996), ensayó una representación que relacionara distintos componentes de software mediante clusters en una representación facetada. Llorens (1996), definió una estructura basada en la arquitectura de un tesauro de software. La construcción semiautomática del tesauro partía del análisis de un conjunto de documentos. Esta estructura se complementaba con la clasificación decimal de Dewey mediante una estructura denominada Árbol de Áreas Temáticas. El tesauro estaba basado en la normativa ISO2788 (Aitchinson, 1987) para tesauros monolingües.

Para conseguir este objetivo se necesita cubrir ciertas etapas entre las que están: la abstracción o generalización de los elementos del software analizado; la clasificación y catalogación de sus componentes; y el desarrollo de herramientas de búsqueda que nos permitan reutilizarlo. Cada dominio está limitado por una serie de fronteras que definen y delimitan su ámbito. El análisis de dominios consiste en (Velasco, 1998):

- Identificar las áreas de conocimiento (dominios) más relevantes y definir sus límites.
- Organizar y eliminar la ambigüedad del vocabulario en cada dominio del problema.
- Seleccionar datos representativos de cada dominio.
- Actuar como revisor de los modelos resultantes.

El realizar análisis de los distintos dominios es uno de los métodos tradicionales para gestionar la información, siendo la organización del conocimiento y las relaciones entre los agentes es uno de los principales objetivos de las ciencias de la información. La idea de "análisis de dominios" no es nueva dentro de las ciencias de la información, sino que han existido aproximaciones en el pasado y en el presente que implícitamente comparten la mayoría de sus puntos de vista básicos.

2.5.1 CLASIFICACIÓN DE TÉRMINOS

La clasificación consiste en organizar datos según sus semejanzas. A los agrupamientos que conseguimos por este medio se les denomina clusters o agregados, y la forma en que se obtienen es la generación de agregados, análisis de clusters o clustering. El análisis de clusters es un nombre genérico para una gran variedad de métodos matemáticos que pretenden encontrar que objetos en un conjunto de datos comparten características comunes.

La generación de clusters por métodos estadísticos es una de las herramientas más ampliamente utilizadas y que mayor éxito ha deparado en cienciometría, algunas de sus aplicaciones, son: la recuperación de información, semejanzas pregunta-documento, conocimiento de frentes de investigación, redes de citas, colegios invisibles, la clasificación de documentos (Garland, 1983) y la clasificación de términos para generación de tesauros, mapas de la ciencia o evolución temática. Es sobre esta última aplicación sobre la que nos centraremos a partir de ahora:

Tras haber identificado y seleccionado los descriptores. Se realiza la indización, cuyos resultados deben de ser integrados antes de comenzar a establecer relaciones. Es tras esta etapa cuando se requiere analizar como se relacionan los distintos descriptores. Existen diferentes modelos estadísticos que permiten reconocer las interrelaciones entre los descriptores. El conocer las relaciones entre descriptores no sólo es una herramienta que permite construir las estructuras propias de un tesauro, sino que da una idea de cual es la semántica implícita en los documentos analizados y nos permiten representar la dinámica de los temas (Callon, 1995). Para modelar la similitud entre los distintos descriptores los clasificadores operan sobre dos parámetros, que son, en qué documentos y con qué frecuencia aparece cada descriptor.

Las relaciones entre descriptores se pueden reconocer utilizando técnicas de clustering. Estas técnicas crean conjuntos de descriptores relacionados entre sí. Mediante el empleo de un algoritmo de clasificación u otro es posible recopilar diferente información. Según Velasco (1998), existen dos grande grupos de clasificadores dependiendo del tipo de estructura de tesauro que se genera:

Conceptos relacionados y sinónimos. Se suele emplear el método de palabras asociadas (Muller, 1997). El método de palabras

asociadas también es denominado en la bibliografía como concurrencia o coocurrencia de términos (por similitud con cocitación o coautoría) o cowording. En el siguiente apartado se expone detalladamente el método de Chen. Este método es básicamente un método de concurrencia de términos, complementado con los pesos del IDF y de la frecuencia del término. Esta técnica estadística genera para cada par de términos un coeficiente que mide el grado de relación entre los términos. Según el grado de relación de cada par de conceptos se pueden obtener dos tipos de relaciones: sinonimias y conceptos asociados. La distinción entre uno y otro se basa en los umbrales de asociación. Los umbrales varían dependiendo del número de asociaciones obtenidas.

Jerarquías. Por jerarquías entendemos los términos específicos y genéricos propios de los tesauros. Para construir las jerarquías se realiza una aproximación top-down. Se va realizando una aproximación desde el descriptor más general al más específico mediante la aplicación iterativa de un algoritmo. En el punto 1.5.3 se abordará este proceso con detenimiento. Entre las técnicas empleadas para conseguir este objetivo están: Comenzando por los descriptores filtrados, la jerarquía se forma desde el descriptor más general al más específico. Tras seleccionar una raíz, el proceso de agregación debe ser realizado con los restantes descriptores. Para conseguir el siguiente nivel de jerarquía se realiza la repetición de la extracción de los componentes principales del cluster. El proceso finaliza cuando se alcanza determinada condición, que es propia de cada método.

Algunas técnicas que se emplean para realizar este proceso son:

Clasificadores estadísticos: K-Means, axial K-Means, max-min e isodata.

Redes neuronales: Kohonen, ART-1, ART-2, Fuzzy ART

En Velasco (1998), se propone un modelo de integración de relaciones mediante la combinación de varios de estos clasificadores. Según la siguiente figura:

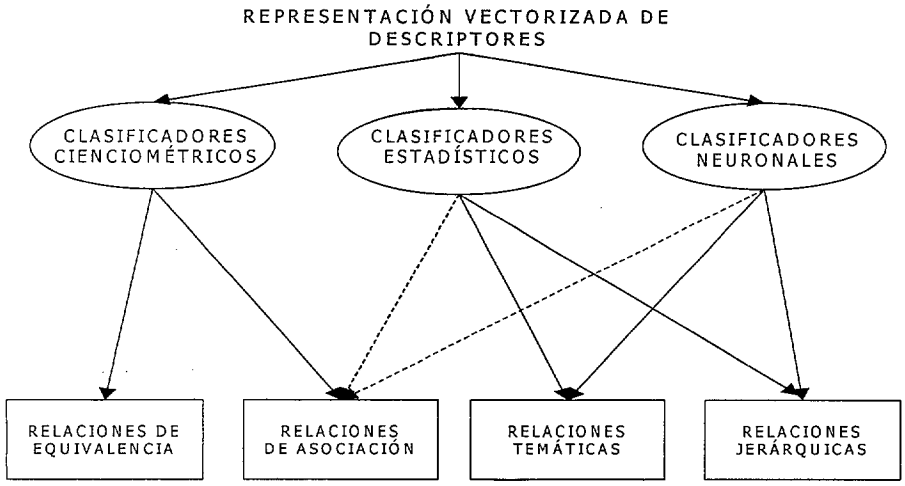


Figura 5: Proceso de generación de relaciones en un tesauro (tomado de Velasco, 1998)

2.5.2 MÉTODO DE PALABRAS ASOCIADAS. CHEN

2.5.2.1 MÉTODO DE PALABRAS ASOCIADAS

Muller (1997), explica este método de la siguiente manera, 'cette méthode considère les mots-clés comme des indicateurs de connaissance (contenu des documents indexés) et se base sur leur coocurrence pour mettre en évidence la structure de leurs relations (clusters). Un cluster est une classe de mots entre lesquels il existe des associations fortes'[3].

El método de palabras asociadas esta dentro de lo que Callon (1995) denominó como un indicador relacional de segunda generación, como ya se explicó más arriba. Durante años, los indicadores cienciométricos han ignorado el texto del cuerpo de los documentos. Tradicionalmente, se han utilizado a los clusters de citas, de autores y de referencias para realizar agrupaciones temáticas. De esta manera, se ignoraba a los documentos que no poseían, por ejemplo, referencias.

Según Callon, un texto puede ser identificado por las relaciones que se establecen entre las palabras. El método de palabras asociadas se basa en el computo de las apariciones conjuntas de palabras clave que definen el índice para los diferentes documentos de un archivo. El índice de equivalencia mide la intensidad de las relaciones de asociación entre dos términos dentro del conjunto de documentos de la colección. De esta manera, el índice de equivalencia (C) se define como:

$$C = C_{ij}^2 / C_i * C_j$$

[3] "el método [de palabras asociadas] considera las palabras clave como unos indicadores del conocimiento contenido en los documentos indizados, basándose en su coocurrencia se pone en evidencia la estructura de sus relaciones (mediante clusters). Un cluster es una clase de palabras entre las que existe una relación estrecha" (N.A.)

Donde:

C_j, es el número de veces que se ha utilizado el descriptor j para indizar un documento.

C_i, es el número de veces que se ha utilizado el descriptor i para indizar un documento.

C_{ij}, el número de apariciones conjuntas en los documentos de los descriptores i y j

Su valor será uno, cuando la presencia de una palabra lleva aparejada la presencia de la otra. Por otro lado, el valor cero, significaría que la presencia de una palabra en un documento excluye la presencia de la otra en la colección.

El cálculo de todas las combinaciones de términos dos a dos da una cantidad de datos inmanejable, por lo que las asociaciones se suelen representar mediante clusters de los términos con pesos más significativos, un ejemplo de esta aproximación es el programa SDOC (Polanco, 1995).

Por otra parte, este método ha sido utilizado por Courtial (1990) para caracterizar campos temáticos y estudiar su evolución por medio de estos clusters. Courtial (Courtial, 1990; Callon, 1995) desarrolló una serie de índices y gráficos que reflejaran cada tema. Estos índices son:

- Centralidad: medida de la intensidad de relaciones con otros. Es la media de las asociaciones entre los términos de un cluster con los términos de otro cluster (asociaciones externas).

- Densidad: intensidad de las relaciones entre las palabras que forman el cluster. Se mide por la media de las asociaciones entre las palabras claves que forman el cluster.

  Mediante la representación de los valores medios de la centralidad y densidad se puede reflejar la morfología de una red de investigación.

- Indice de transformación, es el que mide la similitud entre dos clusters. Normalmente se utiliza para estudiar cómo un cluster se transforma con el paso del tiempo.

2.5.2.2 METODO DE CHEN

El método de Chen (Chen, 1995) es una técnica estadística que genera para cada par de términos una medida del grado de relación. Chen, mediante coocurrencia de términos, realizó un estudio del uso de grupos de palabras que aparecen simultáneamente en varios documentos. Las palabras pueden pertenecer a un lenguaje controlado o a texto libre.

Mediante el método de Chen se pueden obtener los términos relacionados y sinónimos. El propósito es establecer un peso a la relación que existe entre dos descriptores. Para aplicar tal método se deben haber identificado los descriptores, y posteriormente se debe proceder a realizar el análisis de concurrencias para todos los documentos del corpus documental. Se calcula un peso para cada término basado en el modelo de espacio vectorial y en una función de semejanza asimétrica. Cuanto mayor sea el grado de relación entre dos términos mayor será la probabilidad de que los términos sean sinónimos.

Se definen como relacionados permanentes a los términos relacionados de una manera constante, independientemente de la selección documental que se realice en ese dominio. Los relacionados circunstanciales son los relacionados sólo en un dominio especifico debido a los documentos que integran el corpus (Llorens, 1996). Estas relaciones varían en función del grado de relación. En cualquier caso, la aplicación de un umbral para diferenciar estas relaciones es variable según varios índices.

Los pasos que propone Chen para realizar el análisis son:

Cálculo de las frecuencias de término y documental

Se calcula la frecuencia de término y la frecuencia documental para cada término en el documento. La frecuencia del termino, tf_{ij}, representa el número de ocurrencias del término j en el documento i. La frecuencia documental, df_j, representa el número de documentos de un conjunto de n documentos en los que se encuentra el término j.

Una frecuencia de término alta indica que un término está muy vinculado a un documento, mientras que una frecuencia documental alta indica que es un término demasiado general como para ser utilizado como descriptor.

A los términos identificados en los títulos se les asignan mayores pesos que a los términos identificados en el contenido de los documentos y mayor peso, también, a los términos identificados por filtrados que los identificados por indizadores automáticos. Esto se debe a que las palabras del título suelen ser más representativas.

Posteriormente se calcula el peso combinado del término j en el documento i, d_{ij}, basado en el producto del IDF y la frecuencia del término, como se muestra a continuación:

$$dij = tfij \times \log\left(\frac{N}{dfj} \times wj\right)$$

Donde N representa el número total de documentos en el corpus y w_j representa el número de palabras que forman el descriptor T_j. A los términos formados por múltiples palabras o multitérminos se les asignan pesos mayores que a los términos formados por palabras simples ya que los términos formados por varias palabras suelen aportar mayor contenido semántico.

Se genera una tabla de términos concurrentes basada en la función cluster asimétrica desarrollada por (Chen, 1992). El factor de peso que aparece en las siguientes ecuaciones es un perfeccionamiento del algoritmo cluster asimétrico. El coeficiente asimétrico y el factor de peso estimulan términos que son específicos.

$$PesoCluster\left(T_j, T_k\right) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times factordePeso\left(T_k\right)$$

$$PesoCluster\left(T_k, T_j\right) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ik}} \times factordePeso\left(T_j\right)$$

Estas dos ecuaciones indican la semejanza de pesos entre el término T_j y el término T_k (la primera ecuación) y entre el término T_k y el término T_j (la segunda ecuación). d_{ij} y d_{ik} se deben calcular sobre la base de la ecuación del paso anterior. d_{ijk} representa el peso combinado de los descriptores T_j y T_k en el documento i. d_{ijk} se calcula como sigue:

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right)$$

donde:

tf_{ijk} , número de ocurrencias de ambos términos, j y k, en el documento i (se elige el menor número de ocurrencias entre ambos)

df_{jk} representa el número de documentos (en un conjunto de N documentos) en los que las ocurrencias de los términos j y k coinciden.

w_j representa el número de palabras que forman el descriptor T_j. Para "penalizar" los términos generales (términos que aparecen muchas veces) en el análisis de concurrencias, se calculan los siguientes pesos:

$$T_k = \frac{\log\frac{N}{df_k}}{\log N}; \qquad \text{y} \qquad T_j = \frac{\log\frac{N}{df_j}}{\log N}$$

No se obtendrán relaciones de peso entre dos descriptores si T_j tiene una frecuencia documental igual al máximo de documentos, porque el divisor de la fórmula de Chen sería cero (Campo, 1998).

En Velasco (1998), se ha propuesto una división mediante los umbrales calculados a partir de la aplicación del método de Chen. Así, los distintos valores permiten diferenciar sinonimias y relacionados mediante los umbrales de asociación. Para considerar que dos descriptores son sinónimos debe existir alto grado de concordancia entre los coeficientes de relación de éstos con el resto de descriptores. Así, los vectores $\{a_1, a_2, a_3, \ldots, a_m\}$ y $\{b_1, b_2, b_3, \ldots, b_m\}$, que representan los coeficientes de Chen para los descriptores A y B con el resto de descriptores del dominio, deben ser similares para considerar que entre A y B existe una relación de equivalencia.

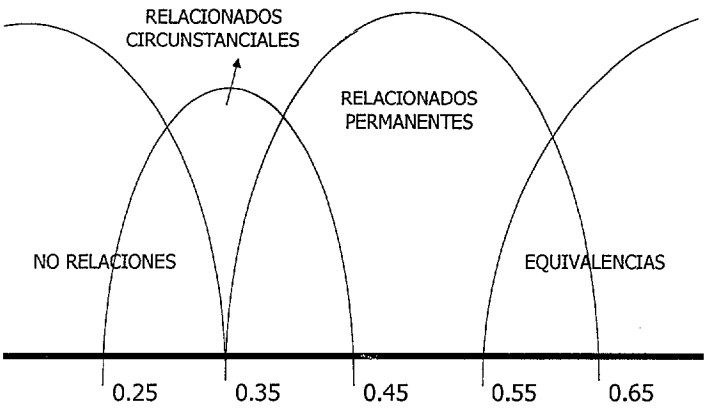


Figura 6: Umbrales de Chen para las distintas relaciones

Aquí se emplea el término sinónimo con el sentido de que dos palabras son substituibles entre sí, por ser utilizadas indistintamente para invocar un mismo concepto del documento. Las sinonimias se pueden detectar así, sobre todo en lenguas romances. En estas se tiende a emplear términos distintos para expresar conceptos semejantes con el fin de evitar la repetición abusiva de un término. En ese caso, para un concepto dado existe un reparto de ocurrencias entre los términos que lo representan de forma más o menos equitativa. Esto hace que los términos tengan una fuerte correlación. El problema surge cuando existen términos preferidos en la mayoría de los documentos y otros que se utilizan, de forma más infrecuente, en esos documentos, o de manera exclusiva, en otro conjunto de documentos. Esto se puede comprobar con los casos ya referidos entre cienciometría, informetria y bibliometría o en el recientemente abordado de palabras asociadas, coocurrencias, concurrencia de términos o cowording. En este caso la probabilidad de obtener grados de relación altos entre los términos que representan el mismo concepto disminuye.

2.5.3 ALGORITMOS ESTADÍSTICOS PARA JERARQUÍAS

Para construir las jerarquías se realiza una aproximación top-down. Esto es, desde el descriptor más general al más específico, mediante la aplicación reiterada de un algoritmo. El primer paso es clasificar la información en jerarquías. El siguiente es clasificar en clusters esa información (Llorens, 96). En un primer momento se actúa sobre la totalidad de términos procedentes del filtrado y de su posterior indización. Una vez se ha seleccionado el descriptor más representativo de una jerarquía se procede a aplicar el algoritmo a los restantes términos. Cada nivel de la jerarquía está representado por un término raíz, que representa al conjunto de descriptores agrupados en ese cluster y nivel. La selección del término raíz se puede realizar con tres criterios: mayor número de ocurrencias, aparición en un mayor conjunto de documentos y menor distancia al centroide. El proceso finaliza cuando se alcanza determinada condición, que es propia de cada método. Este tipo de algoritmos es heurístico y se suelen basar en la minimización de algún índice, por ejemplo, índices cuadráticos basados en distancias.

Algunas técnicas que se emplean para realizar este proceso son:

Clasificadores estadísticos: K-Means, axial K-Means, max-min e isodata.

Redes neruronales: Kohonen, ART-1, ART-2, Fuzzy ART

Los métodos de clasificación jerárquica siguen una serie de pasos predefinidos:

1. Obtención de datos en una matriz de variables (atributos de los objetos) por cada caso.

2. Calcular los valores mediante un coeficiente de similitud.

3. Producción de un método de clustering, y representación en un dendrograma o árbol que muestre la jerarquía.

Uno de los principales problemas para elegir un tipo u otro de jerarquía es que no se conocen a priori el número de 'óptimo' de clusters. Si bien es cierto que no siempre hay que suministrárselo directamente como parámetro de entrada (como en K-Means), también es cierto que, a menudo, se le indica indirectamente, por ejemplo mediante la distancia al centroide para crear un nuevo

cluster o cualquier otro parámetro. Los algoritmos de clustering pueden dividirse de varias formas:

- Directos (constructivos), que tienen aproximación heurística.
- Indirectos (optimizados), que utilizan una función criterio para optimizar la clasificación.

2.5.3.1 ALGORITMO K-VECINOS

Desarrollado por Mac Queen (1967). Este algoritmo es parte de la familia de algoritmos de clasificación de centros móviles. Los centros son recalculados a cada nueva entrada de datos (Lelu, 1993).

Este es uno de los algoritmos de clusters más populares que existen. Hay muchas variantes del algoritmo, y una de las más eficientes es el algoritmo convergente de las k-medias de Anderberg. Es un algoritmo iterativo más formalizado, frente a la excesiva heurística del max-min. Es un algoritmo rápido y eficaz, si la distancia que utiliza es adecuada para el problema considerado. Equivale al algoritmo de Kohonen cuando en éste último se hace que las neuronas no tengan neuronas vecinas.

Este algoritmo comienza por los centros de los conglomerados que se especifiquen (se pueden tomar esos valores de la solución de conglomeración jerárquica) o utiliza los primeros casos k del archivo de datos, donde k es el número de conglomerados solicitado como centros temporales. A medida que se procesan los casos siguientes, un caso sustituye a un centro si la distancia menor del caso al centro es mayor que la distancia entre los dos centros más próximos, se sustituye el centro que esté más próximo al caso y así sucesivamente. Después de seleccionar los centros iniciales de los conglomerados, se actualizan los centros de los conglomerados en un proceso iterativo. Todos los casos se agrupan en el conglomerado con el centro más próximo. La reasignación de casos a los centros de los conglomerados y el cálculo de nuevos centros continúan hasta que los centros dejen de cambiar o hasta que se alcance el número máximo de iteraciones.

Si se tiene una colección de N patrones de entrenamiento, representados por x_i. El centroide (centro de masas) viene representado por:

$$1/N\sum x_i$$

El modo de funcionamiento del algoritmo consiste en mover cada vector al cluster cuyo centroide esté más cercano, actualizando después los centroides de los clusters.

Al principio se tiene una partición inicial que agrupa a K clusters. Se cogen las muestras al azar como clusters individuales y se asignan los N-K restantes al centroide más cercano. Tras cada asignación se recalcula el centroide del cluster ganador. Se calcula la distancia del vector a cada centroide de los K clusters. Si el vector no estaba en el cluster ganador se asigna el vector a ese cluster y se actualizan los centroides. Se repite el proceso hasta conseguir convergencia. Es decir, cuando se deja de producir una reasignación a los patrones en los clusters (Domenech, 1998).

Este algoritmo busca minimizar un índice de rendimiento, basado en la suma de distancias euclídeas cuadráticas de todos los miembros de un cluster a su centroide.

El método exige conocer el número de clases, que es el número de clusters en los que se desea clasificar la muestra de vectores de la población. Se tiene una limitación cuando el número de clases no se conoce por adelantado. Una solución es dejar que el algoritmo determine el número de clusters utilizando parámetros definidos por el usuario. Es, en el fondo, la solución que adoptan muchas redes neuronales.

La convergencia depende también mucho del número de clases. Este algoritmo funciona bien cuando el número de clases es conocido o aproximado. El algoritmo Isodata, funciona mejor cuando no se sabe si el número de clases es el adecuado.

El método k-vecinos es un método iterativo, corrige sus propias asignaciones a base de volver a comprobar en subsiguientes iteraciones si la asignación de la muestra total es la óptima. Esto es una ventaja frente al método incremental, que se explicará más adelante, el cual sólo tiene la distancia limite para decidir si crear una nueva clase.

El método de k-vecinos o k-medias es un método que sólo toma un parámetro externo de funcionamiento: el número de clases o clusters en los que agrupar los términos. Para estimar este valor se ha propuesto el siguiente método (Velasco, 1998).

"Dependiendo del número de términos y del número de documentos, y después de numerosas pruebas, incluyéndose análisis de regresión, se establece que el número de clases pertenece al intervalo [c-2, c+2]donde c = (4 * número de términos + 2 * número de documentos + 2000) / 1000"

2.5.3.2 ALGORITMO K-VECINOS INCREMENTAL

Este algoritmo calcula los clusters de forma incremental. A partir de un patrón de entrada el algoritmo debe actualizar la representación de los clusters y devolver el índice del cluster al cual pertenece el patrón.

En el proceso incremental, un determinado número de atributos son puestos en común. Este algoritmo, como su nombre indica, calcula los clusters de forma incremental (Lelu 93). Pertenece a la familia de algoritmos de clasificación por centros móviles. Es una variante del algoritmo k-vecinos en su versión adaptativa, y del algoritmo de Forgy, en el caso iterativo. Este método está ligado a los modelos neuronales aplicando una ley de aprendizaje del tipo "winner takes all". En lugar de construir los clusters en función de la definición de centros de gravedad define las clases por medio de K semiejes, maximizando el criterio de inercia inter-ejes.

Dado un patrón de entrada, el algoritmo debe actualizar la representación de los clusters y devolver el índice del cluster actual al cual pertenece el patrón, sin necesitar tener presentes los demás patrones. De este modo puede tratarse una sucesión arbitrariamente grande de patrones en tiempo real. Los algoritmos de cluster incremental son muy atractivos para el tratamiento de patrones documentales, dado el gran espacio de almacenamiento que requieren dichos patrones.

Existen varios algoritmos genéricos sobre los que, imponiendo determinadas condiciones, se obtienen sucesivamente los algoritmos que rigen el funcionamiento de las arquitecturas ART 1, Fuzzy ART y la arquitectura híbrida Fuzzy ART para cluster difuso.

Es interesante hacer constar las siguientes observaciones generales:

En este algoritmo el número de clusters no está prefijado de antemano.

Los prototipos son móviles: los patrones de entrada pueden ser reclasificados, y los vectores prototipo cambiar en el tiempo.

Se trata de un metaalgoritmo: deben precisarse las definiciones de "más cercano", "demasiado lejos" y "mover más cerca".

El algoritmo de cluster euclídeo no converge necesariamente a un conjunto fijo de prototipos: los prototipos pueden variar infinitamente, sin converger en el tiempo. El número de clusters creados tampoco es necesariamente finito, y depende de las funciones utilizadas en el algoritmo.

Uno de los problemas que presenta este método es que, en algunos casos, no puede aplicarse de un modo óptimo iterativamente a las clases resultantes, al contrario que los otros clasificadores. A partir del cluster global inicial la primera ejecución del proceso presenta una clasificación en la que se obtienen directamente todas las áreas temáticas, para todos los niveles del árbol. Normalmente, el número de elementos por cluster es bajo y no aconseja volver a dividirlo.

La clasificación temática obtenida es de mayor calidad que en el resto de clasificadores pero tiene el problema de que no existen enlaces entre las distintas áreas tal como sí existían en los demás métodos. Para ello deben contrastarse los resultados globales sobre creación de jerarquías con los resultados temáticos obtenidos en este método.

Aunque menos que el k vecinos, al ir clasificando nivel por nivel, deja una gran cantidad de términos en una clase central. Lo que hace el método es clasificar aquellos términos cuya distancia es bastante grande respecto a la clase central, dejando en esta clase central el resto de los términos. El número de elementos que pertenece a cada cluster en una división específica no es homogéneo. De esa forma el árbol no queda balanceado, con clases dispersas con pocos elementos, y una o dos clases que acumulan la mayor parte de los descriptores. El cluster central es el que aporta, siempre, los niveles a la jerarquía (Velasco, 1998; Domenech, 1998).

2.5.3.3 REDES NEURONALES

Las redes neuronales se utilizan como herramientas o métodos para resolver problemas, fundamentalmente relacionados con el conocimiento humano, especialmente reconocimiento de patrones, reconocimiento de lenguaje hablado, reconocimiento de imágenes,

procesos de control adaptativo y en el estudio del comportamiento de ciertos problemas para los que los ordenadores tradicionales no están muy bien dotados. Estos tipos de problemas suelen estar relacionados con el procesamiento paralelo de un gran número de pequeños elementos relativamente sencillos (Velasco, 1998).

El aprendizaje de una red neuronal está relacionado con los pesos de las conexiones entre sus nodos. Cuando se presenta un patrón a la red, ésta produce una respuesta. Si la respuesta o salida de la red no es la esperada, habrán de hacerse modificaciones para acercar la respuesta obtenida a la esperada. La señal que se recibe en la capa de neuronas de entrada cuando se le presenta el patrón se mueve a través de los enlaces o conexiones entre capas, hacia las neuronas de la capa de salida. Estos enlaces modulan la señal a su paso con los pesos que los caracterizan. Por lo tanto, si se quiere modificar la señal que llega al final a la capa de salida, habrá que actuar sobre dichos pesos.

Las reglas de aprendizaje especifican cómo se irán modificando los pesos de las conexiones a medida que se entrena la red para mejorar el rendimiento de la misma, es decir, que la salida se vaya aproximando cada vez más a la esperada.

Una de las aplicaciones principales que tienen las redes neuronales es su utilización en la clasificación de patrones. Un clasificador es un sistema que va a permitir determinar cuál de las M clases es la más representativa para un patrón de entrada no estático que contiene N elementos. La diferencia entre los dos sistemas (neuronal y tradicional) consiste en la forma de actuar de cada uno de ellos para llegar a la solución final.

El clasificador tradicional actúa en dos etapas; en la primera contabiliza el número de elementos del patrón de entrada que pertenecen a cada clase y en la segunda elige la clase que tiene el mayor número de elementos contabilizados en la etapa anterior. La primera etapa se alimenta con los N elementos que contiene el patrón de entrada y que van entrando al sistema secuencialmente. En él son descodificados y convertidos a un lenguaje interno que permite compararlo con el prototipo (patrón más representativo de cada clase) de cada una de las M clases para ver cual se encuentra más cerca. Posteriormente se codifican y se pasa a la segunda etapa. En la segunda etapa se vuelven a descodificar y se elige la clase que ostenta el máximo número de similitudes o coincidencias,

produciéndose como salida el símbolo que representa a la clase de máxima probabilidad.

El clasificador neuronal actúa también en dos etapas, contabilizándose en la primera el número de elementos que pertenecen a cada clase y en la segunda se selecciona el máximo. La primera etapa se alimenta con los N elementos del patrón de entrada en paralelo, produciéndose aquí la comparación del patrón de entrada con los prototipos de las distintas clases y pasando los resultados intermedios a la siguiente etapa en paralelo. En la segunda etapa se selecciona el máximo. Habrá salida para todas las clases, pero al acabar la clasificación sólo será apreciable la salida para la clase con mayor probabilidad, y el resto serán valores muy bajos o inapreciables. Se pueden utilizar las salidas como realimentación de la primera etapa adaptando los pesos iniciales según un determinado algoritmo de aprendizaje (principio de realimentación negativa).

2.5.3.3.1 MAPAS DE KOHONEN

Están basadas en las propiedades topológicas que presenta el cerebro humano. Ciertas redes neuronales pueden adaptar sus respuestas de tal forma que la posición de la célula que produce la respuesta pasa a ser específica de una determinada característica de la señal de entrada. Por tanto, la estructura topológica de la red absorbe a su vez aquélla que se produce entre las características de los datos, y el sistema no sólo es capaz de clasificar los estímulos sino que mostrará y conservará las relaciones existentes entre las diferentes clases obtenidas.

Existen varios modelos que tratan de explicar cómo se pueden producir las propiedades anteriores a partir de sistemas neuronales. Cada neurona está conectada con otras de su entorno de manera que produce una excitación en las más próximas y una inhibición en las más alejadas, produciendo una interacción lateral. Tanto la excitación como la inhibición lateral son gradualmente más débiles a medida que nos alejamos de la neurona en cuestión.

Este mecanismo hace que cuando un estímulo produce una reacción en una célula, las células de su entorno inmediato se vean influidas por dicha reacción, de modo positivo las más cercanas y de forma negativa las más alejadas.

A partir de estos estudios Kohonen diseñó un modelo, adaptado a estas características biológicas, llamado mapa de características de Kohonen. Consiste en una red neuronal de dos capas, una primera capa de entrada y una segunda llamada de competición.

Cada célula de la capa de entrada está conectada con cada una de las células de la capa de competición, mediante conexiones ponderadas. La capa de entrada tendrá la misma dimensión del estímulo, y será excitada por éste. El objetivo de la red será adaptar sus parámetros de manera que cada unidad esté especialmente sensibilizada a un dominio de la señal de entrada en orden regular. La comparación puede realizarse siguiendo varias medidas de distancia.

El proceso consiste en presentar un estímulo y propagarlo a través de la red, según una determinada función de comparación o distancia, y elegir como ganadora a la célula que produzca una menor señal en la capa de competición. Se pretende que la red responda de manera similar a estímulos parecidos (posteriormente se verá que así se consigue el efecto de clusterización) y para ello se aplica una regla de aprendizaje hebbiana, reforzando más aquellas unidades que hayan respondido en mayor grado al estímulo, de forma proporcional a éste.

Se utiliza el concepto de vecindario para definir la ordenación topológica de las células del sistema y es equivalente a las conexiones laterales del modelo de interacción lateral.

2.5.3.3.2 TEORÍA DE LA RESONANCIA ADAPTATIVA (ART)

Se trata de diseñar un sistema que reaccione de forma plástica a nuevos estímulos, pero de forma estable a aquellos que sean irrelevantes. Existe también la necesidad de poder conmutar entre estos dos estados plástico y estable, cuando sea necesario, para evitar la degradación de lo ya aprendido.

El sistema debe tener incorporada alguna estimación para decidir la relevancia o no de un estímulo al querer realizar esto de forma no supervisada.

ART-1

En este modelo se hace una distinción entre dos tipos de memoria:

Memoria a corto plazo: Es la capacidad biológica de recordar algo que acaba de ocurrir en un instante de tiempo corto, independientemente del número de veces que haya ocurrido. Se mide en función de los valores de activación de las neuronas.

Memoria a largo plazo: Es la capacidad de recordar cosas que ocurrieron en un instante de tiempo lejano, con tal que hayan ocurrido un número suficiente de veces. Es el equivalente a los pesos de las conexiones.

Para solucionar el problema de la estabilidad-plasticidad se introducen en el modelo dos subsistemas y un término de control.

Subsistema Atencional: Es una red de aprendizaje competitivo que se encarga de reaccionar ante estímulos nuevos y los aprende.

Subsistema Orientador: Se encarga de distinguir entre qué estímulos son relevantes para el sistema y cuáles no.

Gain Control: Es un sistema que lleva el control de los módulos que van a actuar y gobierna las señales entre ellos.

ART-2

El modelo ART-2 fue diseñado también por Grossberg y Carpenter. Es una ampliación hecha al modelo ART-1 para poder trabajar con patrones analógicos y no sólo con patrones binarios. Consta de una red de dos capas, utilizando aprendizaje competitivo y puede operar tanto en tiempo discreto como continuo.

2.6 ANÁLISIS DE DATOS MULTIVARIANTE

En las pasadas décadas en las ciencias sociales ha habido un aumento en el número de trabajos que aplicaban métodos estadísticos, en particular de los métodos de análisis multivariante (MVA). Estos métodos son apropiados cuando se establecen relaciones cuantitativas entre dos o más unidades de análisis. Los métodos MVA también están adquiriendo importancia creciente como una ayuda analítica en estudios bibliométricos a gran escala en ciencia y tecnología.

Las patentes y los artículos en revistas se consideran como los máximos canales de comunicación para presentar los hallazgos científicos y para la diseminación del conocimiento científico. Y su cómputo sirve como indicador de la producción científica (Tijseen, 1988). Esta transferencia de conocimiento puede ser mostrada en parte a través del "proceso de citación". El cómputo de las citas es un medio específico para mostrar el impacto de los documentos científicos sobre la comunidad científica. El objetivo general del análisis cuantitativo de datos es proporcionar una descripción de los datos reducida a pocos parámetros.

2.6.1 TIPOS DE ANÁLISIS MULTIVARIANTE

Muchos de estos tipos consisten en un grupo de métodos fuertemente relacionados o en modificaciones de cada método (Tenenhaus, 1985).

2.6.1.1 MODELO LINEAL GENERAL MULTIVARIANTE

El procedimiento MLG Multivariante proporciona análisis de regresión y análisis de varianza para variables dependientes múltiples por una o más covariables o variables de factor. Las variables de factor dividen la población en grupos. Utilizando este procedimiento de modelo lineal general, es posible contrastar hipótesis nulas sobre los efectos de las variables de factor en las medias de varias agrupaciones de una distribución conjunta de variables dependientes. Puede investigar las interacciones entre factores así como los efectos de los factores individuales. Además, se pueden incluir los efectos de las covariables y las interacciones de las covariables con los factores. Para el análisis de regresión, las variables (predictoras) independientes se especifican como covariables.

Se pueden comprobar los modelos equilibrados y desequilibrados. Un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos. En un modelo multivariado, las sumas de cuadrados debidas a los efectos en el modelo y las sumas de cuadrado error se encuentran en forma de matriz más que en la forma escalar encontrada en análisis univariado. Estas matrices se llaman matrices SCPC (sumas de cuadrados y productos cruzados). Si se especifica más de una variable dependiente, se proporciona el análisis multivariado de varianzas usando la traza de Pillai, la lambda de Wilks, la traza de Hotelling y el criterio de mayor raíz de Roy con el estadístico F aproximado así como el análisis univariado de varianza para cada variable dependiente. Además de contratar hipótesis, MLG Multivariante produce estimaciones de los parámetros.

Después de que una prueba F global haya mostrado significación, se puede utilizar una prueba post hoc para evaluar las diferencias entre medias específicas. Las medias estimadas marginales proporcionan estimaciones de valores medios pronosticados para las casillas del modelo y los gráficos de perfil (gráficos de interacción) de estas medias permiten visualizar fácilmente algunas de estas relaciones. Las pruebas de comparaciones múltiples post hoc se realizan de forma separada para cada variable dependiente.

Las variables dependientes deben ser cuantitativas. Los factores son categóricos. Las covariables son variables cuantitativas que están relacionadas con la variable dependiente. Para las variables dependientes, los datos son una muestra aleatoria de vectores de una población normal multivariada; en la población, las matrices varianza-covarianza para todas las casillas son las mismas. El análisis de varianza es robusto a las desviaciones de la normalidad, aunque los datos deben ser simétricos. Para comprobar supuestos, se pueden utilizar las pruebas de homogeneidad de varianzas (incluyendo la M de Box) y los gráficos de dispersión por nivel. También se pueden examinar residuos y gráficos de los residuos.

2.6.1.2 REGRESIÓN MÚLTIPLE

Una regresión múltiple está relacionada con el estudio de una variable dependiente sobre múltiples variables independientes. El objetivo será predecir o estimar el valor medio de la variable dependiente sobre la base de del conocimiento de los valores de las variables independientes. Los métodos de Regresión Múltiple asimismo, permite establecer la contribución relativa de cada una

de las variables independientes para predecir la variable dependiente.

Al tratar con varias variables debemos tratar que (Chatterjee, 1977):

Se pueda tener las variables dependientes en varios niveles predefinidos o que se puedan seleccionar previamente (no estocásticas). Esto es algo frecuente en química o en física pero en Ciencias Sociales es prácticamente imposible por lo que las conclusiones sólo serán ciertas para el conjunto de documentos con los que hemos trabajado.

Que las variables sean medidas sin error es también poco probable. Esto aumenta la varianza residual y distorsiona los coeficientes de correlación. En cualquier caso es complicado saber, sobre todo en Ciencias Sociales, la diferencia entre la varianzas del error aleatorio y la de los errores de medida.

Se debe aplicar siempre el principio de parsimonia, es decir emplear el menor número posible de variables. Hay varios motivos, para identificar las variables más importantes:

- Hacer más manejable y comprensible el modelo
- Acumular menos errores en las estimaciones, disminuyendo la varianza del valor esperado. De todas formas, en algunos casos, se pueden conservar variables poco importantes para profundizar en el conocimiento del modelo.

2.6.1.2.1 ANÁLISIS DE REGRESIÓN LINEAL SIMPLE

Si tenemos un conjunto coeficientes (b_0 ,b_1 ,b_2 , b_3 ,...,b_p) procedentes de los distintos datos de una variable de predicción **x** , y queremos relacionarlo con una variable dependiente (o de respuesta) **y,** podremos representar la relación mediante la Ecuación de regresión en forma matricial es:

$$\mathbf{y_i} = \beta_0 + \beta_1 \mathbf{x_{1i}} + \mathbf{u_i}$$

Donde:

β_0 y β_1 , son los parámetros del modelo de regresión:

u_i, es la Distorsión aleatoria, con media 0 y distribución normal y varianza similar(σ^2)

Así de acuerdo a esta ecuación, β_1 es el incremento en **y** correspondiente en un aumento de una unidad en **x**

Si se calculan β_0 y β_1, minimizando la suma de cuadrados de los residuos $S(\beta_0,\beta_1)=\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{1i})^2$. Los valores de la suma de cuadrados $S(\beta_0,\beta_1)$ son b_0 y b_1, se calculan como sigue:

$$b_1 = \frac{\sum (y_i - \overline{y})(x_{1i} - \overline{x_1})}{\sum (x_{1i} - \overline{x_1})^2} \text{ y } b_0 = \overline{y} - b_1\overline{x_1}$$

Donde: $\overline{y} = \sum y_1 / n$ y $\overline{x_1} = \sum x_{1i} / n$ la estimación de la varianza de b_0 y b_1 se hace mediante: $s^2 = \frac{\sum (y_i - b_0 - b_1 x_{1i})^2}{n-2}$

Para una observación determinada el valor teórico es:

$$\hat{y} = b_0 + b_1 x_{1i}$$

Habrá una diferencia con el valor real que definimos. Este valor es el residuo y equivale a $e_i = y_i - \hat{y}_i$

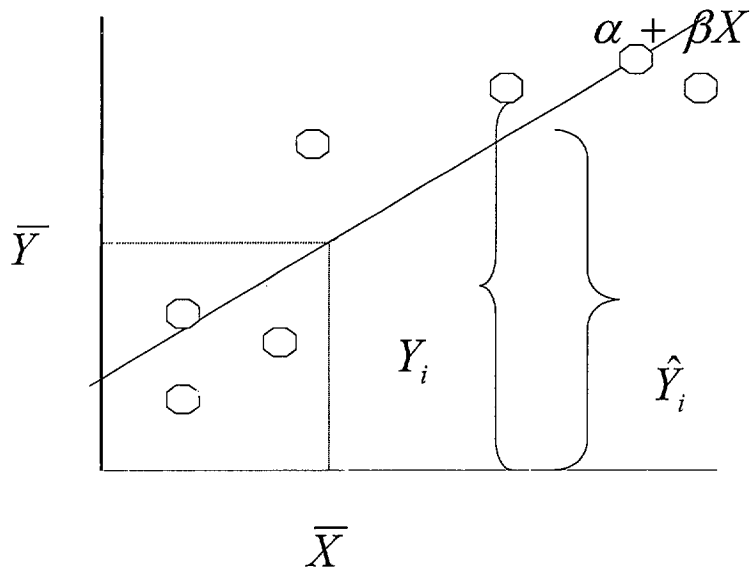


Figura 7: Diferencia entre el valor teórico y el real

O el mismo valor estandarizado,

$$e_{is} = e_i / s$$

El valor de e_{is} nos sirve para evaluar el modelo.

Los residuos estandarizados e_{is} tienen que tener de media cero (con una variación entre ± 2). Es importante ver la distribución de los residuos en un *normal probability plot* para comprobar que no siguen ninguna pauta. En estos diagramas las e_{is} se enfrentan: a valores de $\hat{y}$, a valores de x_1 o al orden en que ocurrieron las observaciones.

El estadístico t se utiliza para ver la diferencia entre el b_1 (o el b_0) de la hipótesis nula y un valor elegido por el investigador (β_1^0). B_1 (o b_0) tienen distribución normal con media β_1 (o β_0). La t tiene n-2 grados de libertad.

$$t = \frac{(b_1 - \beta_1^0)}{e.s.(b_1)}$$

, lo mismo resulta para b_0, donde e.s. es el error estándar, normalmente se toma como hipótesis nula el que β_1^0 sea igual a 0, o lo que es lo mismo, que el parámetro x_1 no tenga incidencia en el valor de la y. Por eso la t de student, para esa variable, tiene que salir significativa para rechazar la hipótesis de que $\beta_1^0 = 0$

El valor esperado se calcula mediante

$$\hat{y}_0 = b_0 + b_1 x_1^0$$

Al que hay que calcular el valor estándar y la varianza.

El Indice de ajuste mide como se ajusta el modelo que hemos desarrollado a los datos reales. Se utiliza el coeficiente de correlación, R (varía entre 1 y –1) o el más usado R^2 (varía entre 0 y 1) este valor representa la cantidad de variabilidad de y que se puede explicar con las x_s que hemos elegido, si es igual a uno el modelo explica toda la variación y si esta próximo a cero faltarían los principales factores que dan lugar a esa variación.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2},$$

Tiene asociada una t de student : $t = \frac{|R|\sqrt{n-2}}{\sqrt{1-R^2}}$

Que tiene también n-2 grados de libertad y que compara esta t con las tes ya tabuladas.

En ocasiones, la relación entre la variables dependiente y la independiente no será lineal. Lo cual es uno de los puntos de partida para aplicar el modelo. Si esto no es así, y existe una relación, habrá que transformarlas. La necesidad de una transformación no se sabe a priori sino que se deduce del ajuste de los residuos en el modelo lineal. En cualquier caso no siempre son linealizables. Algunos ejemplos de transformaciones se exponen en la tabla.

Función	Transformación	Forma Lineal
$y = ax^{\beta}$	$y' = \log y$ $x' = \log x$	$y' = \log\alpha + \beta x'$
$y = ae^{\beta x}$	$y' = \ln y$	$y' = \ln\alpha + \beta x$
$y = \alpha + \beta \log x$	$x' = \log x$	$y = \alpha + \beta x'$
$y = x/ax - \beta$	$y' = 1/y$	$y' = \alpha - \beta x'$
	$x' = 1/x$	
$y = e^{\alpha+\beta x}/1 + e^{\alpha+\beta x}$	$y' = \ln(y/1-y)$	$y' = \alpha + \beta x$

Tabla 3. Transformaciones lineales

Las variables informétricas a menudo necesitan transformarse para poder ser lineales, ejemplos de estas transformaciones, son (Ferreiro, 1993):

- Referencias acumuladas por países (semilogarítmica)
- Ley de Zipf, número de palabras frente al rango de las frecuencias de esas palabras acumuladas (potencial)
- Ley de Bradford, Brookes dividió el modelo en dos partes: potencial y logarítmica.
- Ley de Lotka, potencial

La homoscedasticidad es otro de los requisitos necesarios en la teoría de mínimos cuadrados. Representa una varianza del error constante a lo largo de todas las observaciones. En caso de que la varianza aumente o disminuya con las distintas observaciones (heterocedasticidad) habrá que hacer transformaciones para eliminarla. Se puede detectar por el análisis visual de los residuos: si divergen o aumentan. En casos de heterocedasticidad, el error estándar y los intervalos de confianza

$$(t \pm (\alpha/2 * E.S.))$$

Aumentan, disminuyendo la calidad de los tests. Puede haber varios motivos. Por ejemplo, cuando la distribución de la *y* no sea normal sino que obedezca a una distribución de Poisson, en cuyo caso tendremos que trabajar con $\sqrt{y}$. La distribución de Poisson aparece en sucesos raros (por ejemplo, accidentes aéreos).También cuando la *y* sea Binomial en vez de normal: la transformación sería $\text{sen}^{-1}\sqrt{y}$ (en radianes o grados), si la función es binomial negativa sería:

$$\lambda^{-1}\,\text{senh}^{-1}(\lambda\sqrt{y})$$

Para quitar la heterocedasticidad debida al aumento de la desviación estándar de los residuos según aumenta la x, también se utiliza la función

$$\frac{y}{x} = a' + \frac{b'}{x},$$

Con y/x como variable dependiente y b´/x como independiente.

2.6.1.2.2 REGRESIÓN MÚLTIPLE

La regresión lineal estima los coeficientes de la ecuación lineal con una o más variables independientes que predicen mejor el valor de la variable dependiente. Por ejemplo, se puede intentar predecir el número total de ventas anuales de un comerciante (la variable dependiente) a partir de diferentes variables tales como la edad, la educación y los años de experiencia. O si, por ejemplo, se quiere saber el peso que tiene el número de capítulos, el número de términos científicos y el número de fórmulas sobre el valor de impacto como dependiente.

En los datos del análisis de regresión, las variables dependientes e independientes deben ser cuantitativas. Las variables categóricas tales como religión, mayoría de edad o lugar de residencia necesitan recodificarse en variables binarias (dummy o auxiliares) o en cualquier otro tipo de variables de contraste.

Los supuestos que deben cumplir son:

Para cada valor de la variable independiente, la distribución de la variable dependiente debe ser normal.

La varianza de distribución de la variable dependiente debería ser constante para todos los valores de la variable independiente.

La relación entre la variable dependiente y cada variable independiente debe ser lineal.

Las observaciones deben ser independientes.

Existe un modelo general (FM), análoga a la regresión simple:

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + u_i$$

O con notación de matriz: $\mathbf{Y}=\mathbf{X}\beta+\mathbf{u}$ (con $x_{0i}=1$).

Donde:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & . & . & x_{p1} \\ 1 & x_{12} & . & . & x_{p2} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{1n} & . & . & x_{pn} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ u_n \end{bmatrix} \text{ y } \beta = \begin{bmatrix} \beta_o \\ \beta_1 \\ . \\ . \\ \beta_p \end{bmatrix}$$

Como antes tendremos un valor teórico $\hat{y}$ (para matrices sería $\hat{y} =$ Xb), que tendrá un error frente al real igual a

$$e_i = y_i - \hat{y}_i$$ (como matriz sería e=y-Xb).

Es necesario que u_i y el vector b=(b_0, b_1, ..., b_p), tengan distribución normal e independientes entre sí. El vector b tendrá como media al vector β y como varianza - covarianza $\sigma^2 C$. La varianza de u es $\sigma^2 I_n$.

Como en el caso anterior tendremos que comparar gráficamente las variables, conjuntamente, frente a los residuos estandarizados mediante los *normal probability plots* .

Para deducir los coeficientes recurrimos, otra vez, a los mínimos cuadrados, obteniendo así los b_0, b_1, ..., b_p.:

$S(\beta_0, \beta_1, \beta_2, ..., \beta_p) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_p x_{pi})^2 =$ SSE.

Puesto en forma de matriz sería $S(\beta)=u'u=(Y-X\beta)'(Y-X\beta)$ y $b=(X'X)^{-1}X'Y$.

B_0, b_1, ..., b_p son distribuciones normales con media en β_0, β_1, β_2, ..., β_p, y con varianza en $\sigma^2 c_{ii}$ y con covarianza $\sigma^2 c_{ij}$ (c_{ii} y c_{ij} son elementos de la matriz C).

$\sigma^2 \sim s^2 = SSE/(n-p-1)$; o en notación de matriz:

$$s^2 = (\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})/(n-p-1)$$

Existe una t de Student con n-p-1 grados de libertad, que se corresponde a:

$$t = \frac{(b_1 - \beta_1^0)}{e.s.(s\sqrt{c_{ii}})}$$

El intervalo de confianza es $b_i \pm t\,(n\text{-}p\text{-}1, \alpha/2)\, s\sqrt{c_{ii}}$

Con $\beta_1^0 = 0$ (es lo normal), sólo las variables que tengan una t significativa tendrán importancia en el modelo.

Para evaluar los supuestos, es importante examinar los residuos. Un histograma de los residuos nos puede indicar las violaciones contra la normalidad. También es importante representar $(\hat{y}_i, \hat{e}_i)$ (p.e. los residuos frente a los valores pronosticados)

Para probar hipótesis en las que algunos o todos los coeficientes sean iguales a cero o para eliminar variables con la menor perdida de información posible, trabajamos con modelos reducidos (RM) para confrontar los resultados con el modelo general (FM). Para hacer esta comparación hayamos el estadístico F.

Si $SSE(FM) = \sum (y_i - \hat{y}_i)^2$ y $SSE(RM) = \sum (y_i - \hat{y}_i^*)$, entonces $F = \dfrac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$, se mira el resultado en la tabla de la distribución de la F.

La relación entre la t de Student para los coeficientes individuales y la F cuando probamos $\beta_1^0 = 0$, es, normalmente, que cuando

ninguna t es significativa normalmente F no será significativa y la hipótesis nula será aceptada.

Existe el problema de la multicolinealidad, debida a la no-independencia de las variables 'independientes'. Las consecuencias son: Imprecisión de las estimas (con mayores errores y correlacionados, varianzas muy grandes); o que los errores a la hora de suprimir variables.

La tolerancia, es un estadístico utilizado para determinar la cuantía de la multicolinealidad. La tolerancia de una variable es la proporción de su varianza no explicada por las otras variables independientes de la ecuación. Una variable con una tolerancia muy baja contribuye con poca información a un modelo (es colineal), y puede causar problemas de cálculo. Se calcula como

$$(1-R^2)$$

Para una variable independiente cuando es pronosticada por las otras variables independientes ya incluidas en el análisis. Su recíproco es el Factor de inflación de la varianza (FIV), cuando el factor de inflación de la varianza crece, también lo hace la varianza del coeficiente de regresión, haciendo que el estimador sea inestable. Los valores de VIF grandes son un indicador de la existencia de multicolinealidad. Para ver la colinealidad también se puede utilizar el análisis factorial.

El estadístico de Durbin-Watson (Chatterjee, 1977) se emplea para testar la autocorrelación, es decir, que los residuos no estén autocorrelacionados. Uno de los supuestos del análisis de regresión es que los residuos de las observaciones consecutivas no están correlacionados. Si esto es cierto, el valor esperado del estadístico Durbin-Watson es 2. Los valores menores que 2 indican autocorrelación positiva, un problema muy común en los datos de las series temporales. Los valores mayores que 2 indican autocorrelación negativa.

Sí $d = \sum_{2}^{n}(e_t - e_{t-1})^2 \Big/ \sum_{2}^{n} e_t^2$ y $r = \sum_{2}^{n} e_t e_{t-1} \Big/ \sum_{2}^{n} e_t^2$ entonces d=2(1-r)

Para seleccionar variables se puede proceder de dos maneras (Flury, 1988):

Hacia delante: introduce las variables en el modelo una a uno, siempre y cuando se superen los criterios de entrada. Se comienza por la variable que tiene el F más elevado, en pasos siguientes se van añadiendo la variable que maximiza la F que ya tenemos.

Hacia atrás: Comienza con todas las variables y las va eliminando una a una según el valor criterio de salida. El valor de la F se aplica en los métodos de selección de variables hacia adelante, hacia atrás o por pasos. Las variables se pueden introducir o eliminar del modelo dependiendo de la significación (probabilidad) del valor F o del valor F en sí.

Por pasos: es una mezcla de los dos anteriores. Se examinan las variables del bloque en cada paso para su entrada o eliminación.

2.6.1.2.2.1 APLICACIONES DEL ANÁLISIS MULTIVARIANTE

Las aplicaciones de éste procedimiento se pueden dividir en tres categorías (Egghe, 1990):

Predicción, estimando los valores de la variable dependiente para cualquier combinación de las variables independientes.

Descripción de un modelo de datos, para conocer qué combinación y en qué proporción de variables contribuyen a los valores de la dependiente. El resultado se utiliza para predecir el valor de la variable de una manera más sencilla.

Estimación de los parámetros, para conocer la magnitud que puede tener una variable independiente para conseguir determinado valor de la dependiente.

Se ha utilizado por Bennion y Karschmroon (1984) para medir la utilidad de las revistas científicas con varias estadísticas bibliométricas como variables independientes.

Koening (1983) los utilizó para encontrar indicadores bibliométricos de la actuación investigadora de las mayores compañías farmacéuticas.

2.6.1.3 ANÁLISIS DISCRIMINANTE

El Análisis discriminante es útil para las situaciones en las que se desea construir un modelo de pronóstico de pertenencia al grupo basándose en las características observadas para cada caso. El

procedimiento genera una función discriminante (o, para más de dos grupos, un conjunto de funciones discriminantes) basándose en las combinaciones lineales de las variables predictoras que proporcionan la mejor discriminación entre los grupos. Las funciones se generan a partir de una muestra de casos para los que se conoce la pertenencia al grupo; entonces se pueden aplicar las funciones a nuevos casos con medidas para las variables predictoras pero se desconoce la pertenencia al grupo.

La variable de agrupación debe tener un número limitado de categorías. Las variables independientes que son nominales se deben recodificar como variables "dummy" o de contraste.

Los supuestos que se deben de cumplir son:

Los casos deben ser independientes.

Las variables predictoras deben tener una distribución normal multivariada y las matrices de varianza-covarianza intra-grupos deben ser iguales en todos los grupos.

Se asume que la pertenencia al grupo es exclusiva (es decir, ningún caso pertenece a más de un grupo) y exhaustiva de modo colectivo (es decir, todos los casos son miembros de un grupo).

El procedimiento es más efectivo cuando la pertenencia al grupo es una variable verdaderamente categórica; si la pertenencia al grupo se basa en los valores de una variable continua, se debe considerar el uso de la regresión lineal para aprovechar la información más completa ofrecida por la variable continua.

2.6.1.4 ANÁLISIS LOGISTICO

La regresión logística resulta útil para casos en los que desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de variables predictoras. Es similar a un modelo de regresión lineal pero se ajusta a modelos donde la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar razones de las ventajas de cada variable independiente del modelo. La regresión logística se aplica a un rango más amplio de situaciones de investigación que a análisis discriminante.

La variable dependiente debe ser dicotómica. Las variables independientes pueden ser de nivel de intervalo o categóricas; si

son categóricas, deben ser variables dummy o de indicador codificado.

La regresión logística no se basa en supuestos distribucionales en el mismo sentido en que lo hace el análisis discriminante. Sin embargo, su solución debe ser más estable si los predictores tienen una distribución normal multivariada. Adicionalmente, al igual que con otras formas de regresión, la multicolinealidad entre los predictores puede llevar a estimaciones sesgadas y a errores típicos excesivos. El procedimiento es más eficaz cuando la pertenencia a grupos es una variable categórica auténtica.

2.6.1.5 ANÁLISIS PROBIT

Es un tipo de análisis logístico. Este procedimiento mide la relación entre la fuerza de un estímulo y la proporción de casos que muestran una cierta respuesta al estímulo. Es útil para situaciones con resultados dicotómicos que pueden estar influidos o causados por niveles de alguna o algunas variables independientes y se adapta muy bien a los datos experimentales. Este procedimiento permite estimar la fuerza de un estímulo requerido para inducir una determinada proporción de respuestas, como la dosis efectiva de la mediana.

Para cada valor de la variable independiente (o cada combinación de valores para múltiples variables independientes), la variable de respuesta debe ser una frecuencia del número de casos con esos valores que muestran la respuesta de interés y la variable de total observado debe ser una frecuencia del número total de casos con aquellos valores para la variable independiente. La variable de factor debe ser categórica, codificada como enteros.

Las observaciones deben ser independientes. Si se tienen un gran número de valores para las variables independientes relativas al número de observaciones, como es probable que suceda en un estudio observacional, puede que los estadísticos de chi-cuadrado y de bondad de ajuste no sean válidos.

2.6.1.6 ANÁLISIS FACTORIAL

El Análisis Factorial intenta identificar variables subyacentes, o factores, que expliquen la configuración de correlaciones dentro de un conjunto de variables observadas. El análisis factorial se utiliza con frecuencia en la reducción de datos, identificando un pequeño

número de factores que explique la mayoría de la varianza observada en un número mayor de variables manifiestas. El análisis factorial también puede utilizarse para generar hipótesis relacionadas con los mecanismos causales o para inspeccionar variables para análisis subsiguientes (por ejemplo, identificar la colinealidad previa a un análisis de regresión lineal).

Las variables deberían ser cuantitativas a nivel de intervalo o de razón. Los datos categóricos (como la religión o el país de origen) no son adecuados para el análisis factorial. Los datos deberían tener una distribución normal bivariada para cada pareja de variables, y las observaciones deberían ser independientes. El modelo de análisis factorial especifica que las variables se determinan por factores comunes (los factores estimados por el modelo) y factores únicos (que no se superponen entre variables observadas); las estimaciones calculadas se basan en el supuesto de que ningún factor único esté correlacionado con los demás ni con los factores comunes. Los autovalores representan la varianza asociada a un factor determinado. La suma de los autovalores no puede exceder el número de variables del análisis (cuando se está analizando una matriz de correlaciones) o la suma de las varianzas de todas las variables (cuando se está analizando una matriz de covarianza). Es útil para el desarrollo y prueba de teorías sobre la estructura de un set de variables independientes. Nadel (1981) lo utilizó para establecer la estructura subyacente de las matrices de citación y cocitación entre artículos de un estudio bibliométrico.

El análisis de las componentes principales puede considerarse como una importante variante del análisis factorial, pero también se considera como un método de análisis multivariante en sí mismo. El procedimiento Análisis de componentes principales permite reducir un conjunto original de variables a un conjunto menor de variables independientes que representen la mayor parte de la información contenida en el conjunto original de variables, al explicar la mayor parte posible de la variación del conjunto original de variables. Se ha utilizado en informetría para estudiar las relaciones entre disciplinas y colegios invisibles (Egghe, 1990)

2.6.1.7 ANÁLISIS DE CLUSTERS

Es un método de reducción de datos, que consiste en varios tipos de métodos de clustering. Todos ellos tienen el objetivo común de identificar grupos de elementos parecidos y diferenciarlos de otros grupos.

La construcción de subgrupos homogéneos se basa generalmente en el parecido o no de los perfiles de las variables. Métodos no jerárquicos de análisis de cluster utilizan centroides como representativos de los subgrupos. Estos pasos de conglomeración se muestran en un diagrama de témpanos o dendrograma opcional (árbol de conglomerados).

El análisis de conglomerados es un procedimiento multivariado para detectar agrupaciones en los datos. En los procedimientos de k-medias y jerárquicos, los conglomerados pueden ser grupos de casos. El procedimiento jerárquico también se puede utilizar para formar grupos de variables en lugar de grupos de casos. La conglomeración es una buena técnica a utilizar cuando se sospeche que los datos no son homogéneos y si se desea ver si existen distintos grupos o si se desea clasificar los datos en grupos. En otras palabras, puede empezar sin tener previos conocimientos acerca del grupo de pertenencia.

La clasificación también puede ser un objetivo para el análisis discriminante. Sin embargo, en este procedimiento se empieza con casos en grupos conocidos y el análisis encuentra combinaciones lineales de las variables que mejor caracterizan las diferencias entre los grupos (estas funciones pueden utilizarse para clasificar nuevos casos). Las variables pueden introducirse en la función por pasos-- de esta manera, se identifica un subconjunto de variables que maximiza las diferencias de grupo.

2.6.2 FUTURO DEL ANÁLISIS MULTIVARIANTE EN BIBLIOMETRIA

Las bases de datos han alcanzado una enorme importancia en los estudios bibliométricos Los métodos de análisis multivariante muestran representaciones espaciales de la estructura de los datos como el análisis de cluster que han contribuído a la bibliometría por su ayuda para la elaboración de mapas de la ciencia (Tijseen, 1988).

Con todo aún quedan problemas sin resolver con respecto a la robustez no sólo de los métodos análisis multivariante (Ferreiro, 1993) sino también en relación con aspectos del análisis bibliométrico, en particular con las citas.

Aunque Brookes (1984), parece haber resuelto las objeciones de Haitun sobre las distribuciones Zipfcianas de las ciencias sociales

en contraposición a las Gaussianas de la estadística clásica. Otros problemas parecen surgir a partir de los métodos de palabras asociadas (Leydesdorff, 1996; Campanario, 1994) debido principalmente a que dos factores. Por un lado, la obtención de mapas de la ciencia por estos métodos trabajan con un elevado número de valores perdidos. Por otro lado, los algoritmos propuestos por Courtial (1995), aunque funcionan bien a nivel de documentos individuales, a nivel de set , las coocurrencias no clasifican correctamente debido a la propia dinámica de la ciencia.

3 OBJETIVOS

3.1 OBJETIVOS

El objetivo de esta tesis es desarrollar un método para el estudio de las relaciones entre el análisis del discurso y las herramientas de análisis de la información documental. Por herramientas de análisis de la información documental se entiende los procedimientos que permiten caracterizar y organizar la información. En concreto, indicadores bibliométricos, lingüísticos y algoritmos de clasificación.

En todo el proceso subyace la idea de que las variaciones contextuales tienen una influencia directa sobre cualquier herramienta de análisis textual. Al ignorar esta influencia se adulteran no sólo los resultados de estas herramientas, sino también las herramientas de recuperación de la información, estrechamente emparentadas con aquellas. En resumen, se pretende constatar las variaciones en los diferentes medios.

En caso de existir variaciones significativas en las distribuciones entre contextos, implicaría que difícilmente se puede pretender crear sistemas automáticos eficientes que no tomen en consideración estas variaciones. En concreto, se quiere poner de manifiesto la necesidad de analizar su influencia sobre algoritmos de clasificación.

Más específicamente, se pretende valorar estas variaciones contextuales desde el punto de vista de análisis de género. Se quiere así obtener información sobre los valores de los indicadores cienciométricos, distribución de palabras, variables lingüísticas, dependiendo del área discursiva y de la estructura del género.

Se consideró significativo evaluar los siguientes discursos, agrupamientos y estructuras.

- Estructura del documento.
- Estilo del documento en función de los receptores potenciales (comunidad científica o público en general).
- Temática.
- Publicación.
- Tamaño del documento.

El objetivo que se pretendía con la elección de estas dimensiones era poder delimitar de la manera más realista posible el motivo de las variaciones. No hay que olvidar que las características cualitativas y cuantitativas de un documento, y aún más de un género, responden a una realidad multidimensional (Buchholz, 1995).

Los temas que se seleccionaron en patologías clínicas fueron tres:

- SIDA/HIV: por ser un tema con una gran presencia en cualquier medio y con un interés alto durante los últimos años.

- Síndrome de Creutzfeldt-Jakob, relacionado con la Encelopatía Bovina Espongiforme. El tema estaba adquiriendo una importancia creciente en el año 1996

- Hepatitis, con una larga trayectoria en investigación en la comunidad médica. Su interés en prensa y divulgación es menor.

Con esta selección se pretendía poder comparar diversos grados de evolución en la temática, y ver como evolucionaban. Algunas de las publicaciones seleccionadas tenían factor de impacto, con el fin de poder estudiar relaciones entre el valor de este factor y las características discursivas e informétricas.

Los programas creados para conseguir las anteriores metas, exploran el contexto. El objetivo es capturar información lingüística y del formato del texto, para poder estructurar un documento (Lazarinis, 1998), dividiéndolo en distintos apartados como conclusiones, métodos,.... Como se ha comprobado en observaciones cognitivas los indizadores profesionales utilizan marcadores textuales, estructurales y semánticos (Smith, 1994). Como también, tiene base cognitiva el considerar, que si un autor quiere recalcar determinada parte del documento empleara diferentes formatos de texto (Berri, 1996).

Aparte de algunos indicadores más tradiciones con los que se pretendía ver sus relaciones con el resto de las variables, se desarrollaron algunos indicadores para poder evaluar comportamientos tales como:

- La utilización abusiva de pronombres en determinado género, que disminuyeran los recuentos de descriptores implicando una peor representación del dominio.

- La influencia de la legibilidad del texto en la determinación del dominio.

- Impacto del empleo elevado de siglas. Las siglas son de hecho sinónimos de descriptores multitérminos.

- El porcentaje de citas asociadas a negaciones en una frase y el número de negaciones en un documento. Se pretendía calibrar su influencia, bien por negar los trabajos anteriores, o bien por haber obtenido resultados negativos.

- Varios parámetros morfológicos y vocabularios especializados se definieron, por el peso que previsiblemente; según la literatura, en su utilización como discriminante de los distintos discursos.

- Determinar el número de términos indicativos de la novedad, número de términos inclasificados. La intención es ver como estos influyen en el índice de transformación de un tema (Callon, 1995).

El empleo de la estadística sobre estos datos se realizó con los fines siguientes:

- Realizar un análisis de las componentes principales que nos indicase la procedencia de la varianza en cada discurso y estructura.

- Realizar un análisis discriminante que nos permita diferenciar de manera automática estos discursos para así mejorar las herramientas de clasificación y recuperación.

- Hacer un análisis de regresión multivariante que nos indique los pesos de las variables dependientes en relaciones consideradas de interés.

- Crear relaciones asociativas y jerárquicas para comparar así el vocabulario del MeSH y las obtenidas por el filtrado de términos. También, mediante estos estadísticos de clasificación se pretende ver cual si existe un vocabulario característico de estos géneros y estructuras. Este vocabulario, según Haas (1996), sería útil para la recuperación de determinada tipología documental.

4 ENTORNO DE TRABAJO

4.1 ARQUITECTURAS CLIENTE-SERVIDOR

A través de los años, los sistemas han ido evolucionando. Primeramente aparecieron los "*Sistemas de proceso centralizado*" en los que toda la capacidad de trabajo procedía del Host y el concepto de *personalización* se veía bastante complicado.

Posteriormente surgieron los "*Sistemas de proceso personalizado*" en los que cada máquina individual ya tiene capacidad de trabajo, pero está orientada a un sólo usuario y eso hace bastante complicada la compartición de información entre usuarios.

Más tarde, al darse cuenta de la necesidad de tal compartición de recursos, surgen los "*Sistemas de proceso personalizado con compartición de recursos*", aquí las estaciones poseían, al igual que en el caso anterior, todos los componentes necesarios para el trabajo sobre ellas de forma autónoma, ya que tenían su propia CPU y memoria. Sin embargo, comienzan a formarse ciertas arquitecturas para la compartición de recursos externos (Discos, Impresoras, ...), con lo que aparecen conceptos como *servidor de ficheros* o *servidor de impresión*, aunque el usuario final no tiene porqué darse cuenta de ello.

Esto que aparentemente resolvía los problemas de compartición, generaba nuevos problemas ya que la totalidad de los datos eran computados en las pequeñas máquinas locales, lo que hacía que el tráfico por la red aumentase de forma espectacular.

Tras estudiar estos problemas, se llegó a la resolución de la utilización de los "*Sistemas de proceso compartido (arquitecturas cliente-servidor)*".

Estos nuevos sistemas solucionan los problemas de compartición de recursos y permiten que las aplicaciones de usuario final (cliente o receptor de información) se descarguen del proceso de generación y elaboración de la información, ya que existe otro programa (servidor) que se encarga de procesar toda la información para sólo mandar al cliente los datos que éste tiene que presentar.

Este último cambio conlleva los siguientes puntos:

- La existencia de un software específico (el sistema gestor de base de datos).

- La necesidad de utilizar un protocolo o lenguaje mediante el cual se entiendan cliente y servidor (SQL), es decir, un protocolo de comunicaciones a través de la red, aunque ya era necesario en el punto anterior.

Con estos cambios, el tiempo de proceso de la información en el cliente desciende, por lo que puede dedicarlo a unos mejores sistemas de presentación de información.

La *tecnología cliente-servidor* es una forma de *procesamiento distribuido* en el que las actividades de los ordenadores están compartidas entre ordenadores conectados en red que cooperan (esto supone una gran ventaja de los sistemas de bases de datos que soportan múltiples usuarios). Una aplicación está dividida funcionalmente en dos o más programas que se ejecutan en distintos ordenadores y que se comunican entre ellos mediante el paso de mensajes a través de la red.

Los *programas cliente* se ejecutan en los PCs del usuario. Los programas del servidor lo hacen en los ordenadores más potentes. El cliente envía peticiones al programa servidor que está escuchando. Este programa las recibe, las procesa y envía los resultados de regreso al cliente. El servidor sólo envía los subconjuntos de la base de datos que se ha solicitado. La mayoría de los servidores de bases de datos utilizan comandos SQL porque es un lenguaje conveniente para especificar subconjuntos lógicos de datos.

Normalmente, una red de área local contiene más clientes que servidores. Los ordenadores cliente no pueden compartir sus recursos ni utilizar los de otros clientes. A través de arquitecturas cliente-servidor, el usuario obtiene acceso a las capacidades que no se podían conseguir en el caso anterior. En su lugar, estas aptitudes están disponibles desde un ordenador servidor. Una aplicación cliente-servidor ofrece estas capacidades tanto en uno como en otro.

Los servidores son los ordenadores más potentes en las redes de área local. Un servidor es, normalmente, multitarea de forma que puede atender a múltiples clientes simultáneamente (interactúan con varios clientes en un proceso que se denomina *tiempo compartido*). También el acceso simultáneo de varios clientes a la misma base de datos (en este caso el Sistema Gestor de la Base de Datos es el encargado de ejercer el *control de concurrencia*, no

permitiendo así que dos clientes distintos modifiquen el mismo dato).

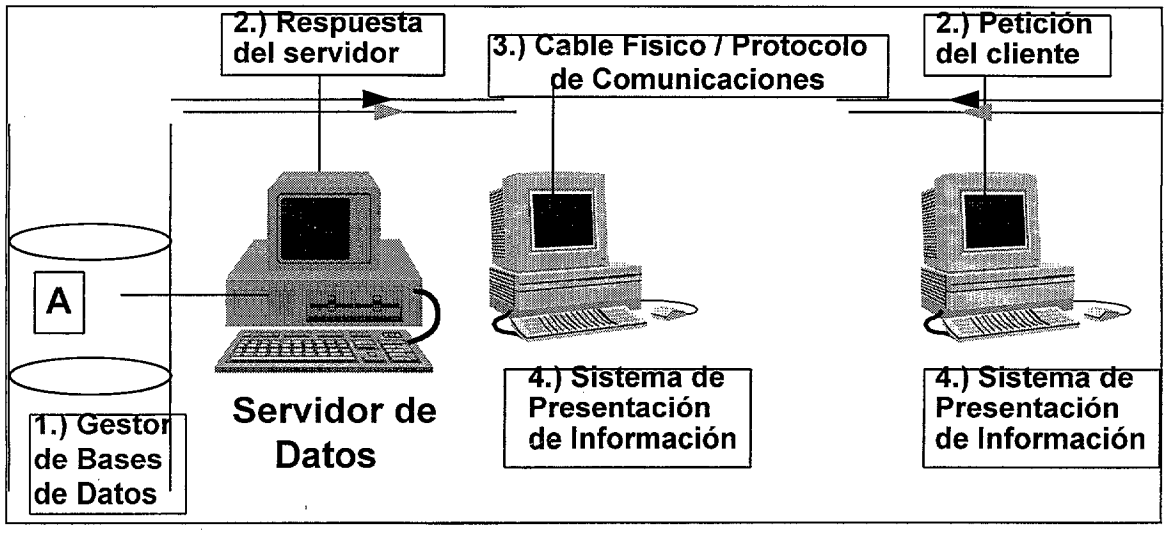


Figura 8: Arquitectura cliente-servidor

Todas las aplicaciones realizadas para desarrollar este trabajo han estado basadas en las arquitecturas cliente-servidor

4.1.1 SISTEMAS GESTORES DE BASES DE DATOS

Un Sistema Gestor de Bases de Datos Relacional es un conjunto grande y complejo de programas responsable de todos los aspectos de la creación, acceso y actualización de la base de datos, por lo tanto, es una aplicación servidora.

Además, proporciona una serie de cualidades:

- El emplazamiento de la aplicación cliente es totalmente independiente del lugar donde se gestiona la información
- La transparencia en la localización de la base de datos permite moverla sin afectar a las aplicaciones cliente.
- Una instalación puede crecer añadiendo más aplicaciones cliente o usando servidores más potentes.
- Una base de datos puede distribuirse y acomodarse tanto a Software como a Hardware.

4.1.1.1 SISTEMAS GESTORES DE BASES DE DATOS RELACIONALES

Una base de datos relacional reduce los datos a su nivel más básico de simplicidad: tablas de dos dimensiones que contienen columnas y filas de valores. La modelización de datos es flexible. Una colección de tablas puede representar complejas relaciones de datos. Las relaciones entre dos tablas están definidas por una columna contenida en ambas estructuras. Esto provoca que los enlaces entre las tablas sean más fáciles de ver y entender.

En un sistema relacional, el orden que ocupan las columnas y las filas no afecta a su significado.

4.1.1.1.1 COMPARACIÓN CON OTROS MODELOS

Los modelos *jerárquicos* y *en red* son procedimentales y realizan búsquedas de un registro cada vez. Para encontrar un registro, se tiene que navegar hasta encontrarlo y dar múltiples comandos procedurales que le indiquen al sistema el camino que debe seguir, paso a paso. En cambio, un sistema relacional proporciona navegación automática. No se ha de saber cómo están representados internamente los datos para obtenerlos. Esto facilita el acceso a los mismos.

En los modelos jerárquicos y en red, se deben utilizar punteros para asociar un registro a otro. Cuando se define la base de datos, se debe tomar la decisión de sí almacenar los datos como punteros o como valores. Estos sistemas utilizan relaciones basadas en punteros predefinidos. Por otro lado, en una base de datos relacional, se puede utilizar cualquier valor para asociar una tabla a otra y se definen las relaciones entre valores cuando se introduce una consulta, no cuando se crean las tablas. Se tiene una mayor flexibilidad.

4.1.1.1.2 VENTAJAS DE LAS BASES DE DATOS RELACIONALES

Los datos se pueden tratar y combinar fácilmente dentro de nuevas relaciones.

Las bases de datos relacionales reducen el almacenamiento de datos redundantes, lo que supone una segunda ventaja, ya que se reducen los requerimientos de espacio de disco.

Los datos son fáciles de actualizar: al existir menos instancias de datos, se reduce la probabilidad de errores de actualización.

4.1.1.1.3 ADMINISTRACIÓN DE LAS BD RELACIONALES

Un Administrador de Bases de Datos (DBA) controla el funcionamiento general de una base de datos y asegura su facilidad de funcionamiento y su operatividad; así mismo, es el responsable del diseño, planificación, instalación, configuración, seguridad, administración, mantenimiento y funcionamiento de un SGBD y de la red que soporta.

En una base de datos existen dos niveles de autoridad que el DBA puede utilizar:

Administrador del sistema (SYSADM): es el de mayor jerarquía en la gestión de la base de datos. Puede asignar uno o más usuarios con el nivel de DBA.

DBA: tiene todos los privilegios sobre todos los objetos de la base de datos y puede cambiar o revocar los niveles de prioridad de acceso de los usuarios.

Operaciones relacionales

El álgebra relacional, utilizado por los sistemas gestores de bases de datos relacionales, se basa en ocho operaciones.

- **Selección:** extracción de todas las filas de una tabla que cumplan una determinada condición.

- **Proyección:** Obtención de una o varias columnas de la tabla y especificación de su orden.

- **Unión**

- **Producto:** El resultado será la suma de las columnas de las dos tablas y tantas filas como el número de tuplas de la primera tabla multiplicado por el de la segunda.

- **Intersección:** el resultado serán las filas que pertenezcan tanto a una tabla como a la otra.

- **Diferencia:** Operación que da como resultado aquellas tuplas que pertenezcan a una tabla pero no a la otra.

●**Join:** obtención de columnas de diferentes tablas cuando los valores de una columna común son iguales. Esta importante operación distingue los sistemas relacionales de los que no lo son.

●**Cociente:** El resultado serán las tuplas de la primera tabla que al *completarse* con las de la segunda permiten obtener la primera.

La combinación de todas estas operaciones es lo que da plena funcionalidad a un sistema gestor de bases de datos relacionales.

4.1.2 LENGUAJE SQL

SQL es un conjunto completo de comandos que permiten el acceso a *bases de datos relacionales.* Es un lenguaje de consulta estructurado utilizado para administrar los datos. SQL fue desarrollado por el *American National Standards Institute* (ANSI) y por el *International Organization for Standardization* (ISO) como una interfaz estándar para un SGBDR. Características:

SQL se creó para utilizarse, o bien interactivamente a través de un programa interfaz, o bien incluido en lenguajes de programación (COBOL, C o SQLWindows).

Es la interfaz estándar para muchas bases de datos relacionales.

SQL es un lenguaje no procedimental: cuando se utiliza SQL, se especifica qué se quiere hacer, no cómo hacerlo. No es necesario describir el método de acceso para poder obtener los datos.

Tiene una estructura de comandos simple para la definición, acceso y gestión de los datos. Los comandos SQL permiten recuperar, modificar, añadir, borrar, definir y gestionar tanto datos como estructuras conceptuales de datos.

Está orientado al conjunto: se puede ejecutar un comando sobre un grupo de filas de datos o sobre una sola fila.

SQL tiene varias capas en las que se va incrementando la complejidad y capacidad.

Los objetos en SQL se organizan en:

- *Bases de Datos* (conjunto de objetos SQL): Contiene una o más tablas.

- *Tablas*: Contiene un número específico de columnas y un número desordenado de filas.

- *Índices*: Conjunto desordenado de punteros a los datos de una tabla; se almacenan en un lugar separado de la tabla. Cada índice está basado en los valores de los datos de una o más columnas de la tabla.

- *Vistas* (View): Forma alternativa de representar los datos que existen en una o más tablas; puede incluir todas o algunas de las columnas de una o más tablas base

4.1.2.1 SERVIDOR SQLBASE.

SQLBase (de Centura Corporation) es un sistema gestor de bases de datos relacionales (SGBDR) con una implementación completa de SQL que es compatible con el DB2 de IBM. SQLBase también proporciona extensiones útiles para DB2. El SQL implantado en el SGBDR SQLBase cumple al 100% las especificaciones del *ANSI SQL*.

SQLBase se puede instalar localmente para un único usuario en su propia máquina, o bien, se puede instalar en un servidor remoto, al que accedan múltiples usuarios, a través de una red de área local.

4.1.2.2 SQLTALK.

SQLTalk es un interfaz interactivo de usuario para SQLBase que permite introducir comandos SQL para realizar operaciones de administración de una base de datos relacional, tales como :

- Definir la estructura de una base de datos.

- Añadir, eliminar y cambiar datos en una base de datos.

- Consultar una base de datos.

- Controlar la seguridad y el acceso a la base de datos.

- Generar informes.

- Comprobar las sentencias SQL antes de embeberlas (introducirlas) en un programa de aplicación.

Este producto es una herramienta para el administrador de datos, para los desarrolladores de aplicaciones de bases de datos y para

usuarios finales que necesiten trabajar con múltiples bases de datos.

SQLTalk también se puede utilizar como una interfaz para otras bases de datos como DB2, ORACLE, INFORMIX, INGRES, etc. SQLTalk es un tipo de aplicación cliente. La base de datos (SQLBase, DB2, ORACLE, etc.) es el servidor.

Gupta Technologies tiene dos productos denominados SQLTalk:

SQLTalk Carácter utiliza una interfaz de línea de comandos que funciona con el sistema operativo DOS.

SQLTalk Windows utiliza una interfaz gráfica basado en ventanas y menús desplegables que funciona bajo el entorno Microsoft Windows. Esta interfaz se utiliza para SQLBase. Si la base de datos es DB2, es necesario utilizar, además, un gateway denominado *SQLGateway.*

Las dos formas de SQLTalk realizan, esencialmente, las mismas operaciones. Sus diferencias residen, principalmente, en que son dos entornos distintos.

Antes de ejecutar SQLTalk, el servidor de bases de datos (SQLBase) debe estar funcionando. Además, se debe efectuar una conexión a cada base de datos antes de realizar operaciones sobre ellas.

4.1.2.3 TIPOS DE COMANDOS SQL

Comandos de definición de datos: crean, modifican o destruyen cualquiera de los objetos de la base de datos tales como tablas, vistas o índices.

Comandos de manipulación de datos: actualizan, añaden o borran los datos contenidos en las tablas.

Comandos de petición de datos: permiten construir peticiones combinando cualquiera de los operadores que se han visto anteriormente.

Comandos de control de transacciones: aseguran la integridad de los datos cuando se producen cambios.

Comandos de control de usuarios: controlan el acceso y la seguridad de tales usuario.

Toda la información necesaria para la investigación que aquí se presenta ha estado almacenada en SGBDR, en concreto en SQLBase Server de Centura Corporation.

4.1.3 LENGUAJES DE CUARTA GENERACIÓN (4GL)

4.1.3.1 INTRODUCCIÓN

Los lenguajes de programación han ido evolucionando de forma que cada vez poseen un mayor nivel de abstracción. Los lenguajes de cuarta generación (4GL) se caracterizan, principalmente, porque son no procedimentales: los programas establecen explícitamente las propiedades que necesita tener el resultado pero no establece cómo debe obtenerse; no importa la manera de producirlo siempre que posea las propiedades requeridas. Los lenguajes no procedimentales o lógicos se caracterizan porque las instrucciones que lo componen no tienen un orden de ejecución prefijado.

A diferencia de éstos, en los lenguajes procedimentales o imperativos, las instrucciones se ejecutan secuencialmente en un orden preestablecido que sólo depende de los valores de los datos a los que se aplica. Con estos lenguajes, un programa establece explícitamente cómo se producirá el resultado deseado y define que las propiedades que se esperan de él son aquellas que se obtengan siguiendo el procedimiento específico.

Los lenguajes de cuarta generación que se emplean en la actualidad son capaces de soportar muchos de los principios de la ingeniería del software, dicho soporte se realiza basándose principalmente en la integración de metodologías de Ingeniería de Software Asistido por Ordenador (CASE), la adaptación de las técnicas de programación orientada a objetos, la incorporación de las arquitecturas cliente/servidor, la asimilación de las nuevas técnicas de trabajo en grupo (*groupware*), etc.

De igual manera que las herramientas de cuarta generación especifican las características de una aplicación a elevado nivel y realizan la generación de código automático basándose en estas especificaciones, las herramientas CASE permiten al ingeniero de software mejorar su entorno de trabajo e incrementar su productividad. Así pues el principal objetivo que se persigue con las

herramientas CASE es reducir las fases de codificación y mantenimiento haciendo hincapié en las primeras fases del ciclo de vida con el fin de solucionar los errores en dichas etapas.

Ventajas principales	
3GL	4GL
Estandarización	Flexibilidad
Actualizaciones conjuntas	Nuevas aplicaciones
Volumen de código	Conversión de código
Rendimiento de la ejecución	Mayor productividad

Tabla 4. Ventajas frente a los lenguajes de tercera generación

El lenguaje más utilizado para desarrollar aplicaciones en las empresas suele ser un lenguaje 3GL (60%). Los 4GL (40%) son ampliamente utilizados por dichas empresas para el desarrollo de las aplicaciones restantes. El uso masivo de los 3GL es debido al dominio de IBM (lenguaje RPG) en el mercado de los miniordenadores y la preponderancia del COBOL en los mainframes.

Conocimientos necesarios para trabajar con los modernos 4GL

Para conseguir una mayor calidad en las aplicaciones y obtener una mayor productividad en el desarrollo de software, el conocimiento de los 4GL debe ser amplio y se debe dominar las técnicas de análisis y diseño de software que se explican a continuación.

Los 4GL requieren *Técnicas Estructuradas* para lo cual utilizan herramientas CASE en el análisis y diseño. Por lo tanto, es importante que el usuario de un 4GL haga un mayor énfasis en el análisis. El desarrollo de aplicaciones cliente-servidor y el soporte de principios de orientación de objetos son otras características de este entorno 4GL.

Para realizar el *Análisis de Datos*, los modernos 4GL implementan una arquitectura que les permiten una portabilidad en las aplicaciones y un ahorro de codificación procedural. Los nuevos 4GL diseñan inicialmente un modelo de datos (o diagrama entidad-relación) y luego implementan las funciones específicas correspondientes a la interfaz externa. Para realizar el modelo de

datos, se ha de dominar las reglas de normalización y los métodos de la modelización lógica de datos.

Los datos son la estructura subyacente bajo la cual operan todos los procesos. Para realizar el *Análisis de Procesos* se ha de construir una lista de las interfaces externas que debe tratar y proporcionar el sistema. También se ha de realizar un modelo de flujos de datos que enlace con el modelo de datos de manera que cada flujo de datos sea una entidad del diagrama entidad-relación.

Para validar el diseño de procesos y comprobar si los modelos de procesos y de datos son completos y consistentes, se utiliza la técnica de *diagrama de historia de entidad.* En esta técnica se consigue que todos los procesos que tienen una determinada entidad como flujo de entrada o salida se reúnan de manera que se pueda examinar los caminos que sigue esa entidad.

La nueva generación de herramientas 4GL ofrece las ventajas de una alta calidad del diseño de las aplicaciones y una mayor productividad en el desarrollo al integrar herramientas CASE que soportan las actividades de análisis de datos y de procesos con el propio lenguaje de cuarta generación.

La elección entre un 3GL y un 4GL no se centra en si uno es mejor que el otro, sino en si es mejor hacer las cosas como siempre o cambiar a un planteamiento diferente en el que los programadores son analistas que deben utilizar herramientas CASE, dominar las técnicas estructuradas, conceptos de diseño de interfaz gráfica, conceptos de arquitectura cliente-servidor, conceptos de orientación a objetos y principios de diseño. A cambio se consigue una mayor productividad, mayor facilidad de mantenimiento y mejor apariencia de la aplicación.

4.1.3.2 CARACTERÍSTICAS AVANZADAS DE LOS 4GL

Productividad / Eficacia: Los 4GL intentan superar la problemática que supone la disminución de prestaciones por aumento de productividad. En un intento por mejorar este cociente (Productividad / Eficacia), surgen los modernos 4GL, en los que se consigue igualar la eficacia de los 3GL y la productividad de los 4GL.

Integración de Metodologías CASE: Una buena integración entre las CASE y el 4GL no sólo puede ser buena en las fases de diseño

(CASE) y desarrollo (4GL) del sistema, sino también en la fase de mantenimiento.

Programación Orientada a Objetos (OOP): En la programación orientada a objetos, la ejecución de un programa se desencadena por medio de un mensaje que alguien (usuario, programa u otro objeto) envía a un objeto determinado. Está considerada como una de las mejores armas para reducir costes de desarrollo. La potencia de los 4GL se basa en la posibilidad de generar código automáticamente. Las nuevas técnicas OOP realizan librerías de objetos que complementan cada vez más el funcionamiento de 4GL.

Es posible programar los 4GL contra, prácticamente, todos los Sistemas Gestores de Bases de Datos Relacionales (SGBDR), independientemente de la plataforma hardware en la que se encuentren, a través de las técnicas de programación OOP de los 4GL y las librerías de objetos correspondientes.

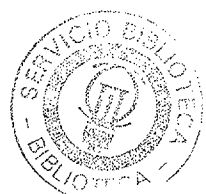
Trabajo en Grupo: Los 4GL actuales soportan muchos de los principios de la Ingeniería del Software y, debido a que ésta es una disciplina de trabajo en equipo, el proceso de desarrollo se estructura de tal forma que la tarea de programación pasa a ser una tarea de diseño, requiriendo a un personal con mayor capacidad de abstracción y de análisis.

Cada usuario posee un nivel de autoridad (privilegios) dependiendo de la función que realice (administrador del sistema de desarrollo, administrador de las bases de datos, desarrollador, etc.).

Multiplataforma a nivel cliente y servidor: Los modernos 4GL pretenden cubrir al máximo posible las necesidades de portabilidad que posea el usuario final.

Para poder trabajar transparentemente en cualquier plataforma cliente es necesario realizar una gran labor de portabilidad de los 4GL entre entornos. No existe ningún desarrollador de 4GL que posea una posición dominante en todos los entornos.

En cuanto al nivel servidor, los más importantes 4GL del mercado ofrecen conectividad suficiente con los más populares servidores relacionales del mercado (como puede ser SQLBase).



Mejora de la integración con GUI: La integración usuario-GUI (Interfaz Gráfico de Usuario) se encuentra perfectamente definida en los 4GL.

Cobertura de los avances en la tecnología de BD Relacionales: Los modernos 4GL se adaptan y evolucionan con el fin de implementar las herramientas necesarias para programar contra los modernos SGBDR que es la pieza clave de la estructura de un 4GL.

Los avances de los SGBDR, proporcionados por herramientas 4GL, son:

Orientación a la transacción: Cualquier proceso que se desarrolle contra una base de datos relacional realiza más de una operación SQL y, en muchas ocasiones, afecta a más de una base de datos. Todas las operaciones necesarias para completar un proceso forman una transacción. El trabajar a nivel transacción supone una enorme mejora en la seguridad e integridad de la información, puesto que ante un problema en la ejecución de cualquier parte de dicha transacción, los sistemas pueden recuperarla por completo. Además, las transacciones operan en un ámbito superior al de una conexión.

Triggers: Métodos para permitir distribuir la información entre distintos servidores. Permiten ejecutar instrucciones SQL contra información distribuida en más de una base de datos estableciendo una comunicación entre los servidores.

Consolas de control y optimización a nivel cliente: Utilización de una interfaz gráfico de usuario, fácil de manejar y capaz de controlar la ejecución de las sentencias SQL, los usuarios conectados en cada momento, etc.

Paquetes de acceso a las Bases de Datos diseñados para el usuario final: Los nuevos 4GL proporcionan conexiones directas a la base de datos desde el propio entorno mediante herramientas que no necesitan ninguna programación.

4.1.4 TÉCNICA DE PROGRAMACIÓN OLE

OLE (**O**bject **L**inking and **E**mbedding) es una técnica que nos permite ver y manipular diferentes tipos de información (por ejemplo, hojas de cálculo, texto, gráficos,...) en un mismo documento (cualquier tipo de objeto creado por una aplicación servidora), estableciendo una conexión dinámica entre las

aplicaciones creadoras de la información y la aplicación que contiene el documento.

Progresión: Portapapeles - DDE - OLE

Desde siempre ha existido la necesidad de integrar información procedente de otras aplicaciones, por este motivo los métodos han ido variando según la necesidad iba en aumento.

1º Portapapeles : Se selecciona la información deseada en el documento origen, se transfiere al portapapeles mediante el comando **Copy** y se inserta en el documento destino con el comando Paste.

Problemas: Los datos transferidos al portapapeles son estáticos, no se pueden modificar y la aplicación destino debe entender el formato de los datos transferidos.

2º DDE: Complementa al portapapeles incluyendo la capacidad de crear un enlace entre la aplicación destino y la aplicación origen, resolviendo de esta forma el problema de la actualización de los datos.

3º OLE: Aparece debido a los problemas no resueltos con DDE (compatibilidad de los formatos). Emplea además del portapapeles y enlaces DDE, los comandos *Copy, Paste* y *Paste Link*. De forma que los datos se siguen copiando de la aplicación cliente e insertándolos en la aplicación servidora a través del portapapeles, con la diferencia de que OLE define un conjunto de reglas que las aplicaciones destino pueden utilizar para decidir como va a ser la información insertada: *un objeto o datos convertidos*. En OLE la información se integra inteligentemente en la aplicación resolviendo de esta forma la compatibilidad de los formatos y la actualización de los datos debido a que en todo momento los datos (objetos) tienen referencia de la aplicación que los generó.

4.1.4.1 CARACTERÍSTICAS DE OLE

La programación está dirigida hacia el objeto, en vez de hacia la aplicación.

OLE permite a las aplicaciones desarrollar otras funciones para las cuales no fueron creadas. Un procesador de textos que implemente OLE puede incluir en sus documentos ilustraciones sin tener herramientas de ilustración. Un ejemplo de una relación de este tipo

sería la que se establece entre Microsoft Word y la aplicación de ilustraciones Microsoft Draw.

Los datos se actualizan automáticamente. El formato de los mismos puede alterarse ya que la aplicación destino no conoce el formato de los datos insertados.

Los ficheros pueden ser más compactos debido a que OLE permite manipular los objetos contenidos en los ficheros sin tener almacenado el contenido de dichos objetos. En este caso se están manipulando "Objetos Enlazados".

Un documento puede ser impreso o transmitido sin utilizar la aplicación que creó dicho documento.

Si un fichero contiene "Objetos Enlazados", éstos pueden ser actualizados dinámicamente.

La tecnología OLE ha sido utilizada para la interacción de los sistemas con el procesador de textos Microsoft Word. De esta forma las variables informétricas se pueden extraer de forma automática.

4.1.5 ENTORNO DE DESARROLLO DE CENTURA BUILDER

4.1.5.1 INTRODUCCIÓN

Centura es un sistema de desarrollo de aplicaciones diseñado, principalmente, para aplicaciones sobre bases de datos SQL.

Sus principales características son:

Funciona bajo los sistemas operativos Windows 95 y NT

Proporciona a los programadores un lenguaje de programación completamente funcional. Este lenguaje se llama SAL (Scalable Application Language) e incluye cientos de funciones construidas que facilitan y aceleran el desarrollo de aplicaciones.

Proporciona a los usuarios finales, mediante aplicaciones gráficas, facilidad de uso, una interfaz consistente y un sistema de ayuda extenso. Todo esto contribuye a que el usuario final aprenda fácilmente a utilizar la aplicación.

Etapas de una Aplicación en Centura.

Fundamentalmente podemos dividir las fases de creación de una aplicación con Centura en dos:

Etapa de Desarrollo (*Development Environment*): En esta etapa, se utiliza la ventana de Diseño para definir la interfaz de usuario. Centura genera automáticamente los elementos de la aplicación para cada objeto que se crea. Después, se deben identificar los mensajes, codificar las acciones y definir el flujo lógico de la aplicación. Existe una ventana de diálogo de ayuda que muestra los elementos o las acciones disponibles en cada nivel.

Etapa de Ejecución (*Executable Environment*): Después de completar la etapa de desarrollo de la aplicación, se crea un fichero con extensión .EXE que contiene el código compilado y listo para la ejecución.

Principales componentes de Centura

Ventana de diseño: es un medio de programación visual con una paleta de herramientas para diseñar pantallas de la aplicación. Se puede simplemente apuntar y pinchar para colocar un objeto cualquiera en una FormWindow y después arrastra y soltar dicho objeto en cualquier posición, así como cambiar su tamaño. Esta ventana también posee un *Customizer* para cambiar los atributos de los objetos.

Código de Aplicación: es un esquema plegable que proporciona una visión global de la aplicación. Por otro lado, cuando se añade cualquier objeto en la ventana de diseño, esa inserción se ve reflejada directa e inmediatamente en el código.

Lenguaje de Aplicación: a pesar de que gracias a los QuickObjects se pueden generar aplicaciones sin tener que introducir ni una sola línea de código, por lo genera, las aplicaciones cliente/servidor necesitan que se introduzca gran cantidad de código. SAL es un amplio conjunto de funciones divididas en categorías (funciones para cadenas, números, ficheros, listas, ...) que proporcionan cualquier funcionalidad que se pueda necesitar.

QuickObjects: son poderosos componentes predefinidos que se pueden introducir y configurar en cualquier aplicación para la resolución sencilla de algunos problemas concretos. Los QuickObjects están definidos en una librería de clases Centura de manera que se puedan utilizar en aplicaciones de este entorno. Debido a que los QuickObject están construidos sobre formas

estándar de programación orientada a objetos de Centura, se pueden utilizar junto con otras clases definidas por el usuario y con código procedimental SAL, además de poder ser retocadas y ampliadas.

Outline Options Bar: es la herramienta más útil para los programadores en este entorno, ya que consta de una lista en la que, dependiendo del lugar del código donde se encuentre el punto de inserción, mostrará todas las opciones válidas para ese lugar, y con simplemente pincharla, esa opción irá a nuestro código.

Depurador: herramienta que proporciona múltiples puntos de ruptura, ejecución paso a paso, visualización de variables y funciones, ...

Report Builder: herramienta utilizada para la creación de informes.

Objetos en Centura Builder

Los objetos en Centura Builder están formados por cualquier elemento que se pueda pinchar, arrastrar, cambiar de tamaño, etc. Son los siguientes: background text, list box, check box, multiline text field, combo box, picture, data field, push button, dialog box, radio button, form window, scroll bar, frame, table window (de nivel superior -top-level- o inferior -child), group box, table window column, line. Pueden ser de dos tipos:

Objeto de nivel superior (*Top-Level Object*): Cualquier *form window*, *dialog box*, MDI Window o *table window*, con excepción de las child table windows. Una aplicación contiene, al menos, uno de ellos. Todos estos objetos tienen una sección de Contenidos (*Contents*) en la que añaden objetos de nivel inferior (*child objects*). Cada uno de estos objetos de nivel superior es una plantilla o especificación de estructuras (*Template*).

Objetos de nivel inferior (*Child Objects*): Estos objetos son creados y destruidos junto con sus padres. No se pueden añadir a todo tipo de objetos de nivel superior. Puede ser también padres de otros objetos hijo (de nivel inferior).

Eventos y Mensajes

Un Mensaje de Centura Builder (**SAM**: ***S**calable **A**pplication **M**essage*) es enviado a una aplicación o a un objeto cuando ocurre

un evento, y proporciona información (contenida en los parámetros *lParam* y *wParam*) sobre este evento.

Centura envía mensajes a todos los objetos (incluida la sección Acciones de la Aplicación -*Application Actions*) excepto a los background text, group boxes, lines y frames (ya que éstos no pueden recibirlos). Así mismo, no todos los tipos de mensajes son enviados a todos los objetos.

También existen mensajes definidos por la aplicación (por el usuario), que son los que la propia aplicación envía a sus objetos o a otras aplicaciones.

4.1.5.2 FUNCIONES DE CENTURA BUILDER

El lenguaje de Centura Builder (***SAL: S**calable **A**pplication **L**anguage*) es un lenguaje procedimental utilizado para escribir las acciones (procedimientos) que se quiere que ejecute la aplicación cuando ocurra un evento.

Existen tres tipos de funciones en Centura:

Funciones del Sistema (*System Functions*): Son llamadas a funciones predefinidas suministradas por SQLWindows. Están identificadas mediante los prefijos Sal y Sql.

Funciones Internas (*Internal Functions*): Son llamadas a funciones definidas por el programador escritas en lenguaje SAL.

Funciones Externas (*External Functions*): Están escritas en lenguaje C o Ensamblador y están incluidas en las librerías de enlace dinámico (DLLs) de Windows (o OS/2).