UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE CIENCIAS

"APLICACION DE LAS REDES NEURONALES A LA MINERIA DE DATOS".

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

EDGAR VALENCIA ROMERO

FACULTAD DE CIENCIAS
UNAM

DIRECTOR DE TESIS: DR. HUMBERTO ANDRES CARRILLO CALVET
CO-DIRECTORA DE TESIS:
M. EN C. NIEVES MARTINEZ DE LA ESCALERA CASTELLS

2006

Índice General

Introducción

El vertiginoso desarrollo de la tecnología computacional en los últimos años, nos permite almacenar cantidades de datos que hace poco tiempo era impensable guardar. Los datos que se almacenan, incluso sobre un mismo tema, son tan diversos en lenguaje, forma, tamaño, etc., que sin herramientas apropiadas es prácticamente imposible analizarlos. Además, se tiene la certeza de que estos *almacenes* contienen *conocimiento* en espera de ser descubierto.

Las herramientas tradicionales para el análisis de datos como el Análisis Multivariado, Análisis de Componentes Principales etc., resultan limitadas para llevar a cabo este trabajo, cuando es necesario cubrir grandes volúmenes de información. Una nueva herramienta, idónea para realizar esta tarea es la que proveen las *Redes Neuronales Artificiales.* Las *Redes Neuronales Artificiales* son herramientas altamente eficientes para procesar grandes cantidades de datos (en órdenes de *terabytes*, *petabytes* o incluso *exabytes*) en escalas de tiempo razonables.

El *ViBlioSOM* es una metodología para la extracción de conocimiento en bases de datos científicas - tecnológicas. Esta metodología ha sido desarrollada en el Laboratorio de Dinámica no Lineal de la Facultad de Ciencias de la Universidad Nacional Autonoma de Mexico en colaboración con el Instituto Finlay de Cuba. Su aplicación requiere la utilización secuencial de varios sistemas comerciales de software. Algunos de estos sistemas de software se utilizan para el preprocesamiento de los datos y otros llevan a cabo el análisis inteligente de los datos con redes neuronales. *ViBlioSOM* usa la tecnología SOM *(Mapas Auto Organizados, Self Organizing Maps, SOM)* para la generación automática de mapas de conocimiento. Estos mapas auto-organizados permiten realizar la proyección de datos que habitan un espacio multidimensional a una cartografía bidimensional, de una forma no lineal. Esta técnica de visualización de información es muy útil para llevar a cabo una síntesis cognitiva de la estructura del conjunto de datos.

La tecnologia SOM se ha experimentado con resultados muy positivos por parte de analistas de información y otros especialistas (científicos de la información, gestores de proyectos, etc.) que necesitan apoyarse en la minería de datos y textos para lograr descubrir conocimiento en acervos voluminosos. La ventaja obtenida en estas aplicaciones con *ViBlioSOM*, en lo que se refiere a economía de esfuerzo humano es espectacular y su

utilización coadyuva al alcance y profundidad que los profesionales pueden añadir a sus investigaciones.

La presente tesis tiene como objetivo exponer los conceptos cibernéticos básicos en los que se basa la metodología *ViBlioSOM* y experimentar con ella haciendo los siguientes tres análisis bibliométricos basados en información científica, obtenida de una de las bases de datos de información biomédica de mayor envergadura, sofisticación y disponibilidad: MEDLINE.

En el Capítulo 1 se expone *El Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Database, KDD)*. Esta metodología tiene como objetivo extraer conocimiento de grandes cantidades de datos por medio de una serie de pasos sistematicos que van desde la limpieza de los datos hasta la visualización e implementación del conocimiento. La metodología ViBlioSOM tiene como columna vertebral al Descubrimiento de Conocimiento en Bases de Datos.

En el Capítulo 2 se exponen brevemente las bases teóricas que se han desarrollado en el campo de la Bibliométria y la Cienciométria. Dichas bases teóricas permiten realizar análisis cuantitativos de la literatura científica y técnica a varios niveles de especificidad.

En el Capítulo 3 se exponen las cualidades que posee MEDLINE para el análisis cuantitativo de la literatura biomédica. La base de datos MEDLINE ha sido generada, administrada y mantenida por la Biblioteca Nacional de Medicina de los Estados Unidos *(National Library of Medicine, NLM)* y es accesible gratuitamente en la siguiente dirección de Internet: http://www.ncbi.nlm.nih.gov/.

En el Capítulo 4 se exponen las bases teóricas de la Red Neuronal de Aprendizaje no Supervisado llamada *Mapas Auto – Organizantes (Self – Organizing Maps, SOM)*. Esta red neuronal es apta para el procesamiento masivo de grandes volúmenes de datos.

Y finalmente, en el Capítulo 5 se experimenta con la metodología ViBlioSOM realizando tres análisis bibliométricos. Estos tienen por objetivo observar el comportamiento de los términos MeSH pertenecientes a la subcategoría ciencias biologicas desde la perspectiva de la subcategoría de Matemáticas. Dicho comportamiento está reflejado en la forma en que algunos de las campos de las matemáticas (sistemas dinámicos, procesos estocásticos, estadística multivariada, etc.) son utilizados y en que forma en

la investigación en el área biomédica (genética, farmacología, enfermería clínica, etc.). Los análisis bibliométricos son:

- *Distribución de Términos:* los términos MeSH de la subcategoría de Ciencias Biológicas (Biological Sciences, vea apéndice C) se distribuyen en las regiones de los mapas auto – organizantes. Lo anterior se puede interpretar como sigue: el mapa clasifica los términos agrupando aquellos términos que tratan de la misma temática en la misma región.

- *Recuperacion de Información:* el ejemplo es similar a lo siguiente: cuando se busca información sobre algún tema, generalmente se acude a la biblioteca. En la biblioteca se escribe el término o la frase a buscar en la interfaz de búsqueda que proporciona la biblioteca. Entonces, los resultados se muestran en una nueva ventana. Estos resultados contienen citas de libros, revistas, etc. Y por último, seleccionamos aquellas citas que se creen más idóneos al tema. En el ejemplo se recuperán todas las citas que tratán temas relacionados con “Neural Networks (Computer)” y “Nonlinear Dynamics”.

- *Descubrimiento de Conocimiento:* se presentá un ejemplo en el que se analiza el resumen de algunos documentos. El análisis tiene como objetivo encontrar información de la cual se derive conocimiento.

En la presente tesis los análisis bibliométricos no tienen el formalismo deseado pues se pretende exponer los posibles usos de la metodología ViBlioSOM.

Capítulo I

El Descubrimiento de Conocimiento en Bases de Datos

Gracias a las nuevas *tecnologías de información y comunicación*[1] el volumen de las bases de datos en cualquier tema crece año con año. Estas tecnologías [1] han hecho que las sociedades del siglo XXI dispongan de una enorme rapidez en la generación y difusión de datos, gracias a que:

- Los usuarios potenciales son muchos.
- Se reducen persistentemente los costos de equipos y sistemas.
- Hay un aumento constante en la calidad y la capacidad de los equipos y sistemas.

Por ejemplo, algunas instituciones o empresas que utilizan estas tecnologías generan, en periodos breves, una gran cantidad de información [2], [3].

- El sistema SKICAT [Sky Image Cataloging and Analysis Tool] explora el cosmos y ha enviado más de 3 terabytes de imágenes de estrellas, planetas, galaxias, etc.
- La empresa Wal–Mart, realiza más de 20 millones de transacciones diarias y tiene una base de datos de 11 terabytes.
- La empresa Mobil Oil, busca almacenar más de 100 terabytes de datos de exploración petrolera.
- La base de datos Genbank, contiene más de 400 millones de secuencias de DNA.

La mayoría de los analistas de información, administradores de conocimiento, etc., consideran que en estas grandes cantidades de datos se esconde *conocimiento* en espera de ser descubierto. La extracción de este *conocimiento* requiere de metodologías de análisis de datos que incorporen técnicas de aprendizaje inteligente que vayan examinando los datos a través de procesos automatizados.

[1] Las tecnologías de información y comunicación comprenden todas las tecnologías basadas en computadora y comunicaciones por computadora, usadas para adquirir, almacenar, manipular y transmitir información a la gente.

En la tabla I.0.1 se ve el volumen que llegan a tener algunas colecciones de datos producidas por estas tecnologías [4].

VOLUMEN	COLECCIÓN
2 kilobytes	Aproximadamente una página completa de texto
20 megabytes	Una radiografía de cuerpo entero
10 gigabytes	Todas las sinfonías de Beethoven con excelente calidad de reproducción
100 gigabytes	100 imágenes de la tierra tomadas por satélite
50 terabytes	Todos los datos de una gran compañía
500 terabytes	Todas las bases de datos y bibliotecas científicas en Alemania
1 petabyte	Todos los datos de los viajes espaciales desde sus comienzos
10 petabytes	Toda la información en la Internet
20 petabytes	Casi tres años de televisión no interrumpida
2 exabytes	El volumen de datos digitales generados mundialmente en un año

Tabla I.0.1: El volumen de algunas colecciones de datos2

Hasta la fecha, el *Descubrimiento de Conocimiento en Bases de Datos*[3] ha sido la metodología más aceptada para el análisis de grandes cantidades de datos.

1.1 Reseña histórica

El descubrimiento de conocimiento se entiende como la evolución natural de las técnicas de análisis de datos [5]. Básicamente, los factores que estimularon dicha evolución son: en primer lugar, las mejoras tecnológicas de las bases de datos que hicieron posible la *recopilación masiva* de datos, a principios de la década de los ochenta. En segundo lugar, el conjunto de técnicas de *aprendizaje inteligente*[4] desarrolladas por la comunidad de inteligencia artificial a mediados de la década de los ochenta.

La figura I.0.1 muestra la evolución del descubrimiento de conocimiento. La primer era recibe el nombre de *era datos,* en la cual, los análisis de datos simplemente se realizaban sobre algún conjunto pequeño de datos utilizando técnicas estadísticas. La segunda era recibe el nombre de *era de bases de datos,* en la cual los análisis de datos empiezan a realizarse

[2] Vea el apéndice A.
[3] Knowledge Discovery in Databases, KDD.
[4] i.e. aprendizajes inductivos y aprendizajes abductivos.

sobre cantidades considerables de datos. La característica de esta era es la recopilación masiva de datos.

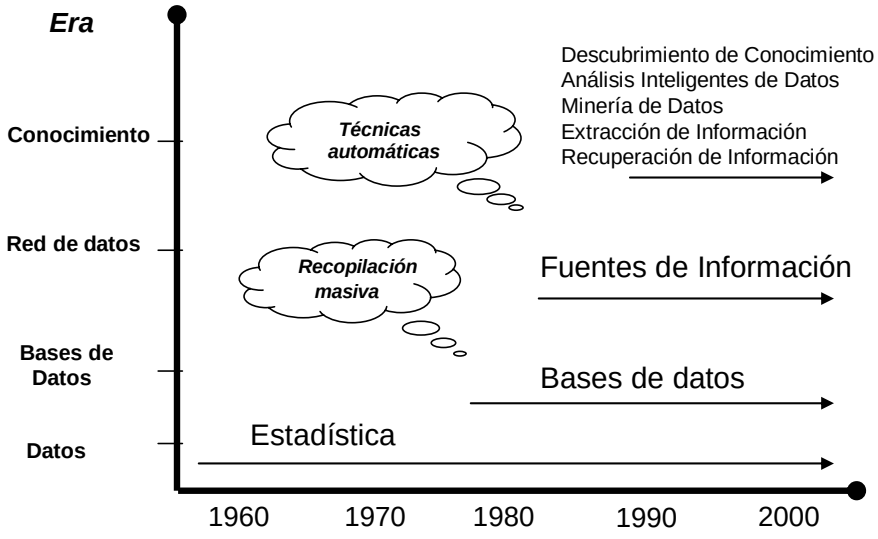


Figura I.0.1: Evolución del Descubrimiento de Conocimiento

A la par del surgimiento de Internet se desarrolla la *era de la red de datos.* En esta era se desarrollan toda clase de *sistemas*[5] que hacen posible tener acceso a una inmensidad de datos proveniente de distintas bases. A través de estos sistemas los análisis de datos disponen de diversas fuentes de datos. En esta era la Inteligencia Artificial desarrolla todas las *técnicas de aprendizaje automático* existentes hoy en día.

En la *era del conocimiento,* los análisis de datos se realizan sobre cantidades considerables de datos provenientes de diversas fuentes. Como consecuencia de esto, algunas de las *técnicas de análisis de información*[6] que se vienen utilizando para tal fin, empiezan a tener problemas con el procesamiento de estas grandes cantidades de datos. El descubrimiento de conocimiento en bases de datos se destaca por resolver el problema utilizando técnicas de aprendizaje inteligente que examinan las grandes

[5] Estos sistemas reciben el nombre de Fuentes de Información. Entre estos encontramos a Los Sistemas de Gestión de Bases de Datos (Database Management System, DBMS), a los Sistemas de Gestión de Bases de Datos Relacionales (Relational Database Management System, RDBMS), etc.

[6] La Recuperación de Información, La Extracción de Información, El Procesamiento del Lenguaje Natural, El Análisis Inteligente de Datos, La Minería de Datos, etc.

cantidades de datos a través de procesos automatizados con la finalidad de extraer conocimiento.

Los investigadores Fayyad y Piatetsky-Shapiro en sus artículos [6], [7], [8] y [9] sobre el descubrimiento de conocimiento lo definen como: *el proceso de extracción no trivial, para identificar patrones que sean válidos, novedosos, potencialmente útiles y entendibles, a partir de datos.*

La definición hace énfasis en un *proceso de extracción no trivial,* que obtiene *patrones* que bajo ciertas condiciones impuestas *(válido, novedoso, potencialmente útil y entendible)* por el usuario representan *conocimiento*.

En resumen, este proceso requiere utilizar técnicas de otros campos de investigación como la estadística, la inteligencia artificial, las ciencias de la computación, etc., para la extracción exitosa de conocimiento. Debido a las grandes cantidades de datos, el descubrimiento de conocimiento requiere de nuevas tecnologías de almacenamiento, acceso y manipulación de datos. El uso de algunas técnicas estadísticas (el muestreo, modelación de datos, evaluación de hipótesis, regresión, etc.) hace que el usuario tenga confianza en el conocimiento obtenido. El proceso debe implementar procesamientos automáticos similares a los implementados en inteligencia artificial, aprendizaje maquina, sistemas expertos, reconocimiento de patrones, etc., para obtener patrones. Estos procesamientos automáticos, deben estar implementados en *interfaces* que permitan a los usuarios interactuar con los datos o los patrones en todas las etapas del proceso. La figura I.0.2 muestra algunos campos que tienen relación con el descubrimiento de conocimiento en bases de datos.

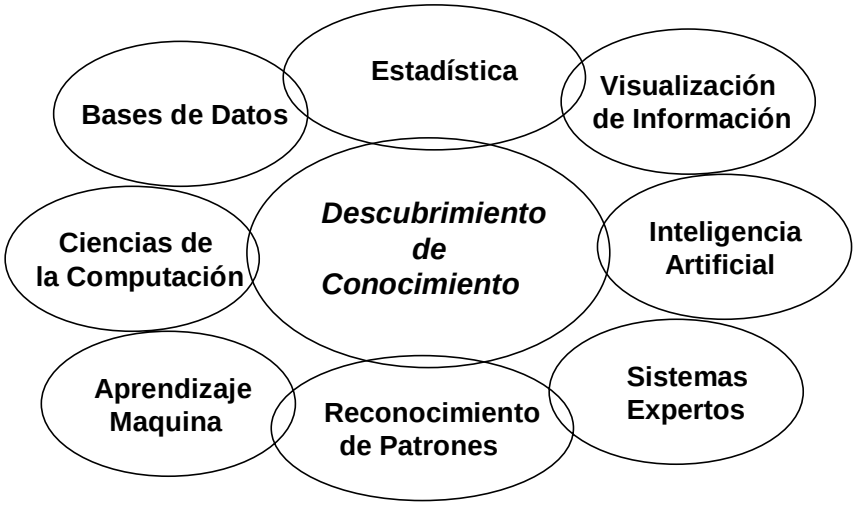


Figura I.0.2: Algunos campos relacionados con el descubrimiento de conocimiento.

No se debe pensar que el descubrimiento de conocimiento es una extensión de los campos siguientes debido a que:

- Las bases de datos no ofrecen técnicas que permitan transformar los *datos brutos*[7] en *conocimiento*. Por consiguiente, la utilización plena de los *datos brutos* depende del uso de técnicas de análisis automático e inteligente que es lo que ofrece el descubrimiento de conocimiento.

- Lo que distingue al descubrimiento de conocimiento de métodos propios de la estadística y de las llamadas técnicas de análisis de datos además de los métodos particulares y algoritmos usados, es el volumen del conjunto de datos que puede manejar, la complejidad de los datos y los resultados.

- Mientras que la Inteligencia Artificial, el Aprendizaje Maquina, los Sistemas Expertos, etc., se apoyan solamente en procesamientos automáticos para obtener conocimiento. El descubrimiento de conocimiento combina los procesamientos automáticos con la interacción humana para obtener conocimiento exacto, útil y entendible.

[7] i.e. los datos que se obtuvieron directamente de la fuente de información

- No se debe confundir al descubrimiento de conocimiento con los *buscadores*[8] de información disponibles en Internet. Los buscadores solamente se limitan a recuperar información sobre algún tema en particular, mientras que el descubrimiento de conocimiento obtiene *conocimiento*.

En cierta forma, el conocimiento se obtiene a través del análisis de los patrones generados por el proceso. ¿Pero que se debe entender por *patrón* en el campo del descubrimiento de conocimiento? Fayyad y Piatetsky-Shapiro proponen lo siguiente: Dado un conjunto de datos F, un lenguaje L y alguna medida de certeza C, definimos *patrón* como una proposición S en L que describe las relaciones entre un subconjunto F_S de F con certeza c, tal que S es más simple (en algún sentido) que la enumeración de todos los hechos en F_S.

Para concluir esta sección Fayyad y Piatetsky-Shapiro indican que el Descubrimiento de Conocimiento en Bases de Datos resalta lo siguiente: *El producto final del descubrimiento en análisis de datos es el conocimiento*[9].

1.2 El diseño del proceso KDD

El descubrimiento de conocimiento consiste en una serie de etapas sistemáticas que comúnmente reciben el nombre *proceso KDD*[10]. La forma genérica de esta serie consta de una etapa de preprocesamiento, una etapa de extracción de patrones y de una etapa de posprocesamiento. Los expertos en el descubrimiento de conocimiento [10] han identificado cuatro elementos básicos que intervienen en el diseño del *proceso KDD*:

A) La base de datos *D*.
B) La representación del conocimiento *L*.
C) Evaluación de los patrones *S*.
D) Las etapas *E*.

En lo que resta de esta sección, se analizarán brevemente las funciones que desempeñan estos elementos en el diseño del *proceso KDD*.

[8] Por ejemplo, *Google.*
[9] "The knowledge is the end product of a data driven discovery"
[10] Abreviación de "Knowledge Discovery in Databases Process".

A) La base de datos *D.*

Una vez que se han seleccionado las *fuentes de información*[11], se necesitan bases de datos para almacenar los conjuntos de datos que se extraigan de éstas. En el mercado existen una enorme variedad de *bases de datos*[12] para realizar esta tarea. Hay que tener en cuenta que cada tipo de base de datos ofrece capacidades distintas de almacenamiento, conectividad a otras bases de datos, precio, etc.

En la Figura I.0.3 se ve una forma de configurar nuestras bases de datos. El núcleo esta representado por el *Sistema de Gestión de Bases de Datos*[13] llamado Procesamiento de Transacciones en Línea.

Este extrae los datos ya normalizados de la base de datos y los distribuye en una unidad pequeña de almacenamiento llamada *Data Warehouses*[14]. Los usuarios tanto de la Aplicación A, como de la Aplicación B acceden a estos datos por medio del *lenguaje*[15] llamado OLAP.

[11] i.e. una fuente de información es una persona u objeto que provee datos

[12] Por ejemplo, las bases de datos relacionales, las bases de datos transaccionales, las bases de datos de objeto-orientado, las bases de datos de objeto-relacional, etc.

[13] El Sistemas de Gestión de Bases de Datos llamado Procesamiento de Transacciones en Línea (Online Transaction Processing) consiste de una colección de programas que permiten almacenar, modificar y extraer información de una base de datos.

[14] Los Data Warehouses son bases de datos estructuradas especialmente para la interrogación y el análisis. Estos contienen datos históricos de la empresa. Los Data Marts son colecciones pequeñas (Sub-base del Data Warehouses) debido a su función de almacenar datos sobre algún tema específico.

[15] i.e. Lenguajes de Interrogación (Query Languages). Estos lenguajes son utilizados para pedir información a la Base de Datos Central, al Data Warehouses, al Data Marts, etc. Entre los lenguajes más populares para realizar dicha tarea encontramos al SQL (Structured Query Language), y al OLAP (Online Analytical Processing).

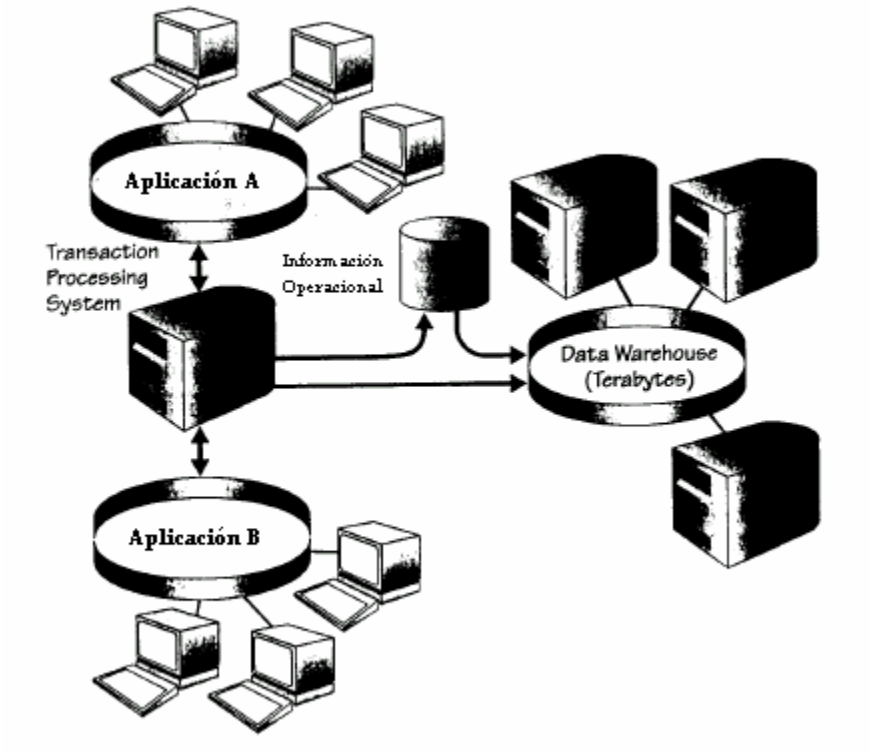


Figura I.0.3: Sistema de Gestión de Bases de Datos.

Fuente: Bigus J., "Data Mining with neural networks", Mc GrawHill, USA, 1996.

B) La representación del conocimiento *L*

El Aprendizaje Maquina ha desarrollado *lenguajes*[16] para representar al *conocimiento*. Estos *lenguajes* se clasifican en dos grandes grupos: *Lenguajes Abductivos y Lenguajes Inductivos*. Estos últimos son de especial interés debido a que los humanos lo emplean cotidianamente, pues sencillamente, entiende su medio simplificándolo en un modelo.

Los lenguajes inductivos [11] se clasifican, por lo general, en tres grandes grupos que son los siguientes:

- *Algoritmo de aprendizajes basados en casos (Instance-based learning algorithms)*: los conceptos se aprenden a través del almacenamiento

[16] No confundir con los lenguajes de interrogación o los lenguajes de programación.

de casos prototipos de los conceptos; y no se construyen representaciones abstractas.

- *Algoritmos de aproximación de función (Function approximation algorithms)*: incluyen métodos conexionistas y estadísticos. Estos algoritmos están más cercanos a las nociones matemáticas de aproximación e interpolación; los conceptos se representan como formulas matemáticas.

- *Algoritmos de aprendizaje simbólico (Symbolic learning algorithms)*: los conceptos se aprenden construyendo una simbología, la cual describe una clase de objeto. La simbología puede ser lógica proposicional (propositional logic) o lógica de primer orden (first-order logic).

Los árboles de decisión, la producción de reglas, las redes semánticas, esquemas, cuadros, son algunos ejemplos de lenguajes inductivos que sirven para la representación de la asociación, la clasificación, el conglomerado, etc.

Entre los algoritmos de aprendizaje utilizados para realizar estas tareas se encuentran el aprendizaje supervisado y el aprendizaje no supervisado, entre otros [12]. En la sección 4.2 se explica brevemente en que consiste el aprendizaje supervisado y el aprendizaje no supervisado.

Estos *lenguajes* determinan los conceptos que un algoritmo puede o no puede aprender. Ademas, algunas representaciones afectan la velocidad de aprendizaje, la legibilidad de la descripción del concepto, etc. En resumen, el uso incorrecto de estos *lenguajes* nos lleva a representar al conocimiento en forma equivocada.

C) Evaluación de los patrones *S*

La extracción de patrones usando técnicas de aprendizaje inteligente, que vayan examinando los datos a través de procesos automatizados *arroja* una enorme cantidad de patrones. La identificación de aquellos patrones que realmente proporcionen conocimiento es una tarea que una persona difícilmente lograría. En consecuencia, algunos investigadores propusieron las llamadas *funciones de evaluación de patrones* para resolver este problema.

Las *funciones de evaluación de patrones* son filtros, que solamente permiten el paso a aquellos patrones que cumplan las restricciones

impuestas por el usuario. Para ser más exactos, la *función de evaluación de patrón* es una función que *mapea* un conjunto de proposiciones expresadas en L a un conjunto de valores numéricos (usualmente) [13] y [14].

Por lo general, estas funciones se dividen en las siguientes clases: *objetivas y subjetivas*. Las *objetivas* solamente dependen de la estructura de los datos y de los patrones extraídos de éstos. Mientras que las *subjetivas* dependen de las necesidades específicas del usuario y de su conocimiento. La tabla I.0.2 muestra ejemplos de estas funciones.

NOMBRE	REPRESENTACIÓN	FUNDACIÓN	CLASE
Rule-Interest Function de Piatetsky-Shapiro	Probabilística	Regla única	Objetiva
J-Measure de Smyth-Goodman	Probabilística	Regla única	Objetiva
Itemset Measures de Agrawal-Srikant	Probabilística	Regla única	Objetiva
Rule Templates de Klemettinen.	Sintáctico	Regla única	Subjetiva
Projected Savings de Matheus-Piatetsky-Shapiro	Utilitaria	Regla única	Subjetiva
Interestingness de Silbershatz-Tuzhilin	Probabilística	Conjunto de reglas	Subjetiva

Tabla I.0.2: Algunas funciones para la evaluación de patrones.

Los expertos en el descubrimiento de conocimiento se muestran conforme en:

- Con estas funciones, el usuario puede obtener *patrones interesantes*, que no son necesariamente interesantes para otro usuario.
- Se pueden generar una gran cantidad de *patrones interesantes objetivamente*, pero de poco interés al usuario.
- Los patrones pueden ser *interesantes objetivamente* y *subjetivamente* a la vez; *interesantes objetivamente* pero no *subjetivamente*; y viceversa.

En el siguiente ejemplo se explica brevemente en que consiste la Función Regla–Interés *(Rule-Interest Function)* propuesta por Piatetsky – Shapiro. Se usa para cuantificar la correlación entre los atributos de una regla simple de clasificación. Una regla simple de clasificación se refiere a una implicación lógica $X \Rightarrow Y$ en la que a un atributo (lado derecho) le

corresponde otro atributo (lado izquierdo). La Función Regla–Interés se define como:

$$RI = \left|X \cap Y\right| - \frac{\left|X\right| \cdot \left|Y\right|}{N}$$

En donde, N es el número total de atributos. $|X|$ y $|Y|$ son el número de atributos que satisfacen la condición X y Y respectivamente. Mientras que $|X \cap Y|$ es el número de atributos que satisfacen $X \Rightarrow Y$. Mientras que $|X||Y|/N$ es el número de atributos esperados si X y Y hubiesen sido independientes (es decir, no asociadas).

Cuando $RI = 0$ entonces, X y Y son estadísticas independientes y la regla no se considera interesante. Cuando $RI > 0$, $(RI < 0)$, entonces X es positivamente (negativamente) correlacionada a Y. La *significancia* de la correlación entre X y Y puede determinarse usando la prueba de la *Chi-cuadrada* para una tabla de contingencia *2 x 2*.

Como se aprecia, las funciones de evaluación son bastante complejas y de aplicación específica. En consecuencia, los expertos en la extracción de conocimiento, recomiendan tener en mente medidas más simples como las siguientes: *validez, novedoso, potencialmente útil* y *comprensibilidad.*

La *validez* se refiere a que los patrones deberán seguir siendo válidos, en nuevos conjuntos de datos o en conjuntos de prueba, bajo algún grado de certeza. *Novedoso* se refiere a que los patrones, tal vez, no sean los típicos patrones que siempre maneja el usuario. *Potencialmente útil* se refiere a que los patrones deben permitir obtener cierto beneficio al usuario. Y por ultimo, la *comprensibilidad* se refiere a que los patrones deben ser fácilmente entendibles al usuario, quizá no inmediatamente sino después de algún posprocesamiento. Por supuesto que estas medidas alternativas dependen del contexto de aplicación.

D) Las etapas *E*

Como se ha mencionado anteriormente, la extracción de conocimiento consta de una serie de etapas sistemáticas, que comúnmente reciben el nombre de *Proceso KDD*. La figura I.0.4 muestra la interacción que existe en esta serie de etapas sistemáticas. Algunas de las características de esta interacción son:

- Cada etapa cuenta con una serie de operaciones.

- El proceso es *cíclico*, es decir, la información fluye de una etapa (o de una operación) a la siguiente etapa (operación) e inversamente a etapas previas (operaciones previas), hasta obtener los resultados deseados.

- La configuración del proceso KDD puede llegar a ser compleja dependiendo del contexto de la aplicación.

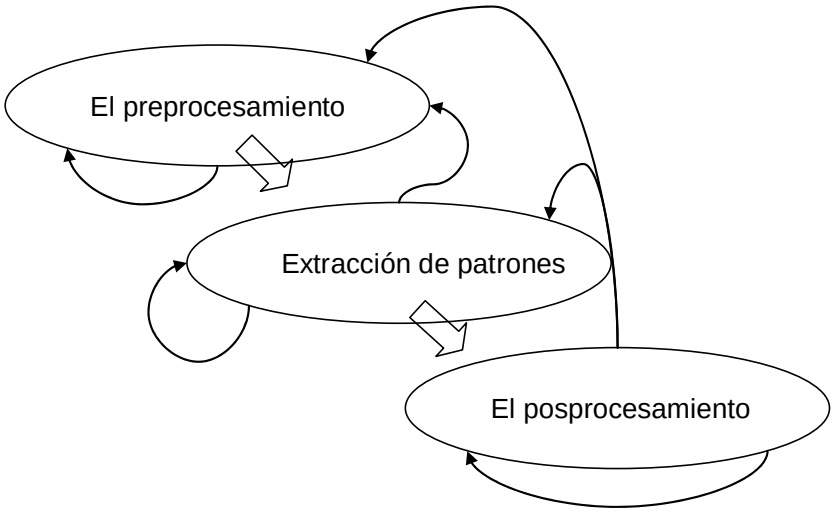


Figura I.0.4: El proceso KDD

En la sección 1.5 se presentan el CRISP-DM, el SEMMA y el ViBlioSOM como ejemplos de procesos KDD. Cabe destacar que el ViBlioSOM es una metodología para extraer conocimiento de bases de datos científicas-tecnológicas. A continuación se detalla brevemente cada etapa del proceso KDD.

1. La etapa de preprocesamiento

El preprocesamiento[17] consiste en la transformación de los *datos brutos* en *datos confiables* [15] y [16]. Entre las operaciones más comunes que se

[17] Realizar el preprocesamiento brindará un cierto grado de confiabilidad al conocimiento que se obtiene. De todas las etapas, esta consume hasta un 50% del tiempo

pueden realizar para este propósito están: *la limpieza, la integración, la selección, la transformación y la reducción de los datos.*

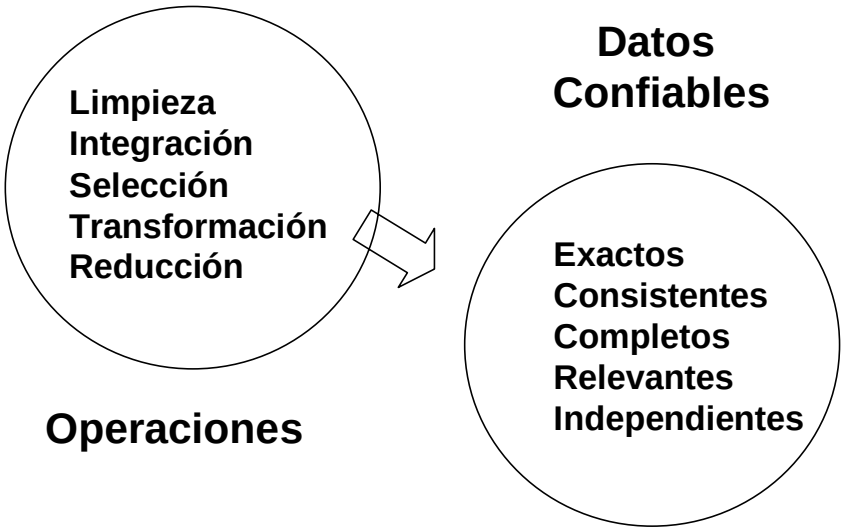


Figura I.0.5: Las operaciones del preprocesamiento.

En resumen, cada operación consiste en:

- *Limpieza*: el conjunto de datos obtenido de la fuente de información, por lo general, contendrá: *ruido*, esto es: errores, datos perdidos, datos irrelevantes, etc. Es necesario, eliminar estos desperfectos con el fin de conseguir un conjunto de datos lo más estandarizado o normalizado posible.

- *Integración*: es común obtener tantos conjuntos de datos como fuentes de información consultadas. Lidiar con muchos conjuntos de datos es inconveniente, por ello, debemos integrarlos en un sólo conjunto de datos.

- *Selección*: para obtener conocimiento confiable debemos seleccionar el subconjunto de datos más relevante para la extracción de patrones.

- *Transformación*: algunos algoritmos que se emplean en la extracción de patrones requieren representaciones apropiadas de los subconjuntos de datos relevantes.

- *Reducción*: algunas veces es posible reducir a formas más compactas al conjunto de datos o al subconjunto de datos más relevante.

Finalmente, algunas características de los conjuntos de datos confiables, son las siguientes: *exactos*, cuando dan una representación fiel del dominio; *consistentes*, cuando todas sus partes tienen sentido dentro del contexto, *completos* es cuando todos sus atributos tienen valores distintos de cero (no son nulos); *relevantes* cuando llevan información de interés sobre el problema; *independientes* cuando la información que ofrecen no es redundante.

Las operaciones[18] requieren de herramientas matemáticas sofisticadas como: El Análisis de Fourier, La Teoría de Ondículas, Las Curvas Principales, El Método de Monte Carlo, etc. Además, de Estadística Paramétrica, Estadística No Paramétrica, Modelado de Datos, Diseño de Experimentos, Regresiones Lineales y Regresiones No Lineales, Regresiones Logísticas, Correlaciones Canónicas, Métodos Bayesianos, Métodos Multivariantes, etc.

El Análisis Exploratorio de Datos[19] se ha constituido como la herramienta más utilizada en esta etapa [17]. La finalidad del Análisis Exploratorio de Datos es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas. Con este propósito, el Análisis Exploratorio de Datos proporciona las llamadas *Técnicas de Preprocesamiento de Datos*.

[18] Dependiendo del tipo de dato, ya sea, nominal, continuo, intervalos, categórico, etc., seleccionamos la herramienta matemática.
[19] Exploratory Data Analysis, EDA

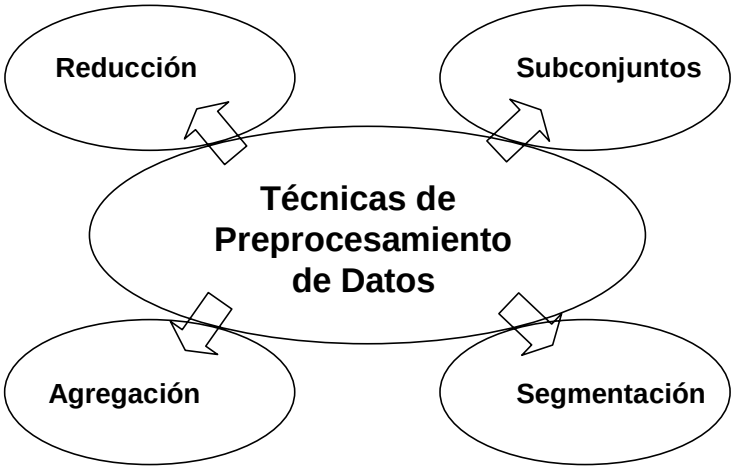


Figura I.0.6: Técnicas de Preprocesamiento de Datos.

- *Las técnicas para la reducción de la dimensión* permiten pasar de un conjunto de dimensión d a un conjunto de dimensión k con $k < d$. Entre estas técnicas encontramos el análisis de componentes principales, análisis de factor, escalado multidimensional, etc.

- *Las técnicas de subconjuntos* permiten pasar de conjuntos con d datos a subconjuntos con k datos ($k < d$). Entre estas técnicas encontramos, el muestreo, el muestreo tipo Jackknife y el muestreo tipo Bootstrap.

- Con *las técnicas de segmentación* se pasa de conjuntos de datos a conjuntos de subconjuntos de datos. Entre estas técnicas encontramos, la segmentación basada en valores de atributos o rangos de atributos.

- Con *las técnicas de agregación* se pasa de conjuntos de datos a conjuntos de valores agregados. La agregación consiste en sumar, contar, minimizar, maximizar, etc., ya sean los valores de los atributos o las propiedades topológicas. Una vez conseguida la agregación, la podemos visualizar por medio de graficas de histogramas, de pastel, de barras, de líneas, de burbujas, etc.

En la figura siguiente se muestran algunas de estas técnicas: (a) Gráfica de Pastel, (b) Gráfica de Histograma, (c) Gráfica de Burbujas, (d) Red Bayesiana, (e) Aproximación por las Funciones de Haar, (f) Gráfica de Líneas, (g) Dendrograma, (h) Caras de Chernoff, (i) Componentes Principales, (j) Escalado Multidimensional de las Capitales Europeas, (k) Curva Principal (Línea Continua) y Regresiones no Lineales (Líneas no Continuas).

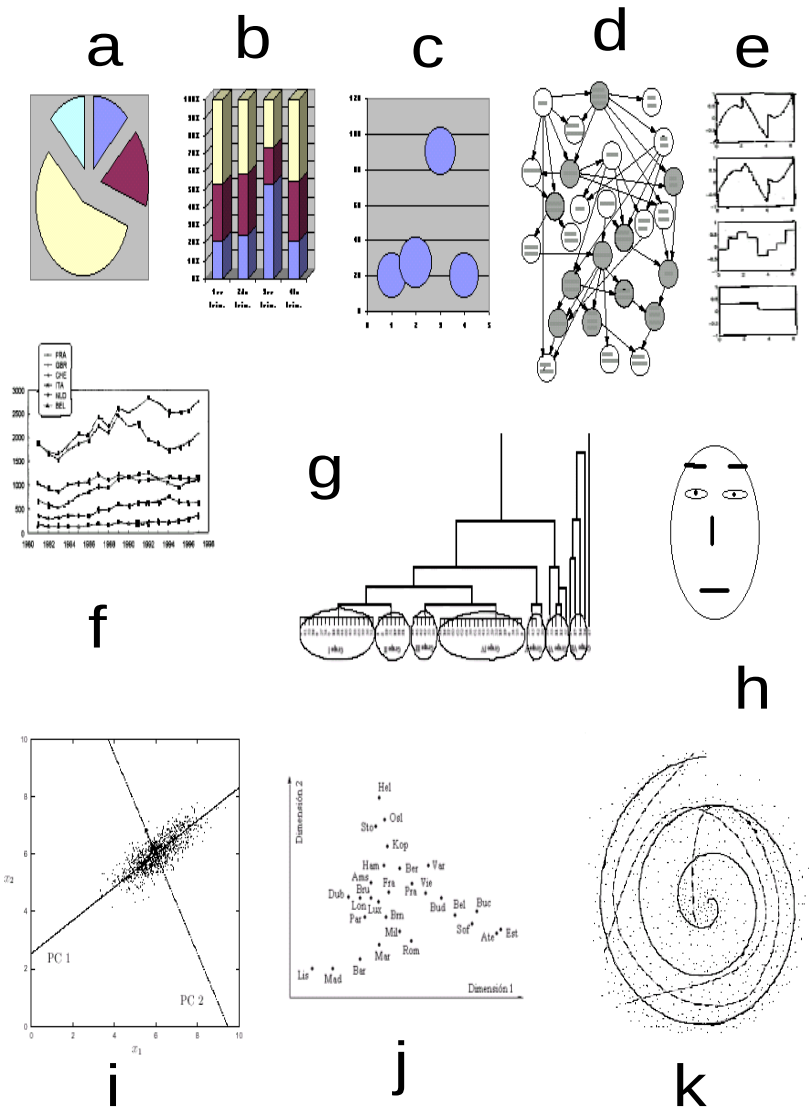


Figura I.0.7: Visualizaciones de algunas técnicas EDA.

2. La etapa de extracción de patrones

La extracción de patrones[20] es considerada por los expertos en el descubrimiento de conocimiento, como el motor del proceso KDD. La etapa se caracteriza por el uso de algoritmos (ver Representación del Conocimiento) para la extracción de patrones, estructuras, regularidades o singularidades en grandes y crecientes conjuntos de datos. Todo lo anterior puede estar representado en forma de reglas de asociación, reglas de clasificación, conglomerados, etc. La tabla I.0.3 resume las tareas más comunes que realizan dichos algoritmos [18], [19], [20]. Los fundamentos matemáticos de los algoritmos no serán tratados.

TAREA	ALGORITMO
Asociación	Estadística, Teoría de Conjuntos.
Clasificación	Árboles de Decisión, Redes Neuronales.
Conglomerado (Cluster)	Redes Neuronales, Estadística
Modelado	Regresiones Lineales y No Lineales, Ajustes de Curvas, Redes Neuronales.
Series de Tiempo	Estadística, Modelos ARMA, Metodos Box-Jenkins, Redes Neuronales
Secuencias de Patrones	Estadística, Teoría de conjuntos

Tabla I.0.3: Las tareas típicas de la Minería de Datos.

En resumen, la *asociación* consiste en determinar relaciones de similaridad o correlación entre atributos de los datos; la *clasificación* trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto predefinido de clases); el *conglomerado* trata de hacer corresponder cada caso a una clase, con la peculiaridad de que las clases se obtienen directamente de los datos de entrada utilizando medidas de similaridad; el *modelado* consiste en encontrar un modelo que describa dependencias significativas entre las variables. En *series de tiempo* se analiza la existencia de series, para detectar ciertas regularidades en los datos como son periodicidades, tendencias y desviaciones. En *secuencias de patrones* se intenta modelar la evolución temporal de alguna variable, con fines descriptivos o predictivos.

Como se ha mencionado anteriormente, el ViBlioSOM es una metodología para extraer conocimiento de bases de datos científicas-tecnológicas. Para ello, utiliza el algoritmo de los Mapas Auto-Organizados

[20] A esta etapa también se le conoce como Minería de Datos.

de Teuvo Kohonen. El capitulo 4 destaca las cualidades de dicho algoritmo en el procesamiento y visualización de datos multidimensionales.

3. La etapa de posprocesamiento

Los patrones extraídos por la minería de datos, deben ser presentados al usuario en forma apropiada para su análisis, mediante el uso de *formas visuales*[21]. Una forma visual hará que el usuario utilice el aparato sensitivo primario humano, que es la visión, tanto como todo el poder de procesamiento de la mente humana, para hacer que estados complejos del comportamiento de los datos sean comprensibles durante el análisis [21], [22].

El campo de la visualización de información, ofrece toda una gama de técnicas visuales para conseguir el *discernimiento*[22]. Las principales ventajas del *discernimiento* son el descubrimiento, la elaboración de decisiones y la posibilidad de explicar el comportamiento de los datos. El objetivo de la visualización de información, es transformar la información original en información más significativa, a partir de la cual el usuario puede ganar en comprensión.

Algunas técnicas visuales [17] que pueden ser implementadas en el *Sistema KDD*[23], son las técnicas geométricas, las técnicas basadas en íconos, las técnicas orientadas en píxeles y las técnicas jerárquicas, entre otras. En las dos figuras siguientes, se muestran formas visuales sobre una gran cantidad de páginas de Internet.

[21] En el campo de la Visualización de Información a las graficas se le llama formas visuales. Debido a sus características de interactividad con el usuario. Se pueden reconstruir, realizar acercamientos a partes específicas de la forma visual, manipular la información que esta representada en la forma visual, etc.
[22] Distinción
[23] Vea sección 1.4

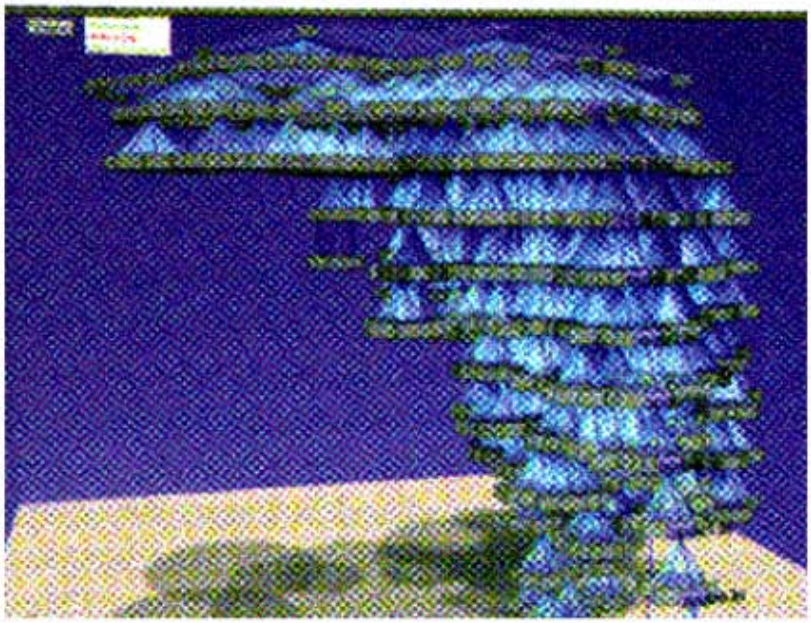


Figura I.0.8: Visualización de 20,000 páginas de Internet ordenadas en un *Cone Tree*.

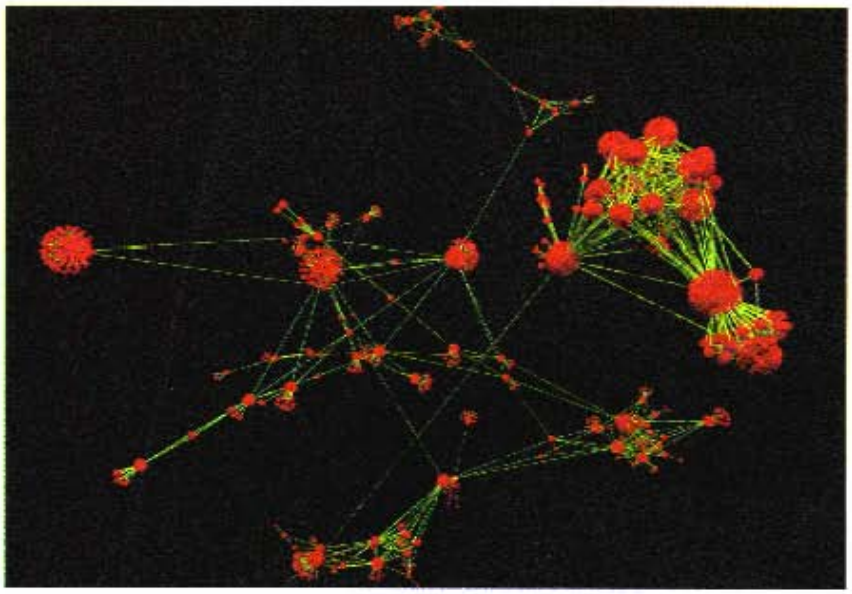


Figura I.0.9: Visualización de una gran cantidad de páginas de Internet interconectadas entre si.

Fuente de ambas figuras: Keim A. Kriegel H. “Visualization Techniques for Mining Large Databases: A Comparison”. In Transactions on Knowledge and Data Engineering TKDE'96), Special Issue on Data Mining, Vol. 8, No. 6, 1996, pp. 923-938.

1.3 Tareas del descubrimiento de conocimiento

Los profesionales de los más diversos campos del conocimiento utilizan y desarrollan todo tipo de Procesos y Sistemas KDD. En resumen, tales desarrollos brindan apoyo en la *Toma de Decisiones*. [23], [24] y [25]

CAMPO	TAREA
Comercio/Marketing	-Identificar patrones de compra de los clientes. -Buscar asociaciones entre clientes y características demográficas.
Banca	-Detectar patrones de uso fraudulento de tarjetas de crédito. -Identificar clientes leales. -Encontrar correlaciones entre indicadores financieros.
Seguros y Salud Privada	-Predecir qué clientes compran nuevas pólizas. -Identificar patrones de comportamiento para clientes con riesgo.
Transportes	-Determinar la planificación de la distribución entre tiendas. -Analizar patrones de carga.
Medicina	-Identificación de terapias médicas satisfactorias para diferentes enfermedades. -Asociación de síntomas y clasificación diferencial de patologías. -Estudio de factores (genéticos, precedentes, hábitos, alimenticios, etc.) de riesgo-salud en distintas patologías.
Procesos Industriales	-Predicción de fallos -Estimación de composiciones óptimas en mezclas. -Simulación costes/beneficios según niveles de calidad
Análisis y evaluación de la Información Científica y tecnológica.	-Sistemas de Información Científica. -Sistemas de gestión del conocimiento. -Sistemas de inteligencia empresarial. -Vigilancia Científico – Tecnológica. -Observatorios de ciencia y tecnología.

Tabla I.0.4: Algunas tareas del descubrimiento de conocimiento.

Por ejemplo, en el Comercio y en el Marketing el descubrimiento del conocimiento se utiliza para identificar las preferencias de compras de los clientes: *... se identifican a los clientes que compran sólo novedades, a los que compran todo tipo de mercancía, a los que compran sólo si hay ofertas, a los que compran esporádicamente o a los clientes visita. De esta forma, la empresa, decide que tipo de mercancía convertirá, por ejemplo, a los clientes que compran sólo novedades en clientes que compran todo tipo de mercancía.*

1.4 El Sistema KDD

El *sistema KDD*[24] permite a los usuarios interactuar con los datos en todas las etapas del proceso KDD. Los *Sistemas KDD* deben basarse en los siguientes principios: simplicidad, autonomía, confiabilidad, reusabilidad, disponibilidad y seguridad.

Además, el *Sistema KDD* no debe imponer conocimiento a los usuarios, sino que debe guiarlos a través del proceso KDD para que ellos elaboren sus hipótesis, ideas, etc. Entre las características deseables [26], [27], [28] del *Sistema KDD*, se encuentran:

- *Manipulación de grandes volúmenes de datos.* Se requiere de grandes cantidades de datos que proporcionen información suficiente para obtener un conocimiento adicional.

- El Sistema KDD tiene que ser capaz de procesar enormes volúmenes de datos, es decir, debe ser *escalable*. Esto significa que el tiempo requerido para extraer conocimiento es directamente proporcional al volumen de datos. Ver figura I.0.10.

- *Eficiencia.* Debido al volumen de datos que fluye entre cada una de las etapas del proceso KDD, es esencial, la *eficiencia*. Un proceso KDD mal planificado implicará que su respectivo *Sistema KDD* tenga problemas con la *Complejidad Computacional*[25].

- *Utilizar alguna forma de aprendizaje.* Una de las premisas mayores del descubrimiento de conocimiento es que el conocimiento se descubre utilizando técnicas de aprendizaje inteligente que van examinando los datos a través de procesos automatizados.

[24] Refiriéndose al sistema computacional de hardware y al software que corre en una computadora o computadoras

[25] Vea el apéndice A

- *Interactividad e iteración.* Debido a esta característica, el *Sistema KDD* puede ser interpretado como *El Descubrimiento de Computadora asistida por un Humano*[26] o simplemente como *El Descubrimiento Humano asistido por una Computadora*[27].

En la siguiente figura se ilustra este concepto de escalabilidad. En el eje de las X se muestran las unidades de almacenamiento y en el eje de las Y se muestra el tiempo. Por ejemplo, supongamos que nuestra base de datos esta en el orden de un Terabyte (T) entonces al sistema le podría llevar 2 horas procesarla o quizás 2 días. Ahora supongamos que nuestra base de datos esta en el orden de un Petabyte (P) esta vez al sistema le podría tomar 3 horas, 3 días o quizás 3 años procesarla.

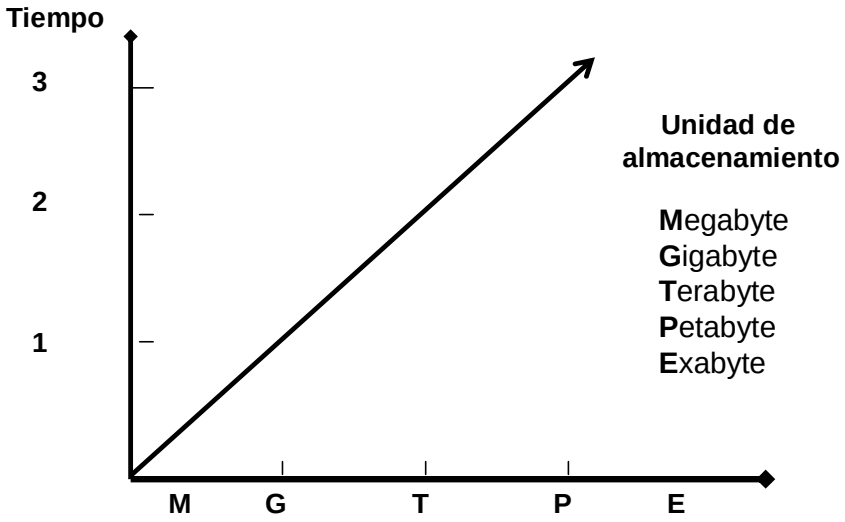


Figura I.0.10: Escalabilidad del Sistema.

Los expertos en el descubrimiento de conocimiento recomiendan que antes de diseñar algún proceso y sistema KDD se tomen en cuenta los siguientes criterios:

[26] Human-Assisted Computer Discovery
[27] Computer-Assisted Human Discovery

Criterios prácticos: analizar si existe potencialmente un impacto significativo; si no hay métodos alternativos; si existe apoyo del cliente para su desarrollo; si no existen problemas de legalidad o violación a información privilegiada; etc.; *Criterios técnicos*: consiste en cerciorarse si existen suficientes datos; atributos relevantes; poco *ruido* en los datos; conocimiento del dominio[28], etc. Además, algunos de los problemas que pueden surgir durante el desarrollo del *Sistema KDD* son: entrenamiento insuficiente de los métodos autónomos; herramientas de soporte inadecuadas; abundancia de patrones; cambios rápidos de los datos en el tiempo; datos complejos (espaciales, imágenes, texto, audio, video, etc.).

1.5 Ejemplos de Procesos y Sistemas KDD

En el mercado, existen una gran variedad de sistemas KDD, cuyo objetivo es dar soporte a cada etapa del proceso KDD. Entre los sistemas KDD más utilizados estan: Clementine, SAS Enterprise Miner, GeoMiner, DBMiner, MLC++, IBM Intelligent Miner, IDIS, MineSet, Kensington Discovery Edition, DataEngine, NGO NeuroGenetic Optimizar, Visipoint, Enterprise Miner, etc.

- *Clementine.* Es una rutina implementada en el software SPSS. *Clementine* se basa en el proceso KDD llamado CRISP–DM. Utiliza una gama enorme de algoritmos para la extracción de patrones en forma automática e inteligente.

- *MineSet.* Desarrollado por Silicon Graphics Inc. Proporciona múltiples algoritmos de minería de datos para la clasificación y asociación. La característica distintiva de *MineSet* es que proporciona una variedad de herramientas de visualización, en todas las etapas del proceso KDD.

- *IBM Intelligent Miner.* Entre los algoritmos de minería de datos que proporciona están las reglas de asociación, clasificación, modelado, análisis de secuencias de patrones, aglomerados, etc. La característica distintiva de *IBM Intelligent Miner* es que, cuando se integra al sistema de bases de datos DB2 desarrollado por IBM, proporciona una excelente escalabilidad de los algoritmos de la minería de datos.

[28] Domain Knowledge

En lo que resta de esta sección se expone en forma breve las metodologías CRISP-DM, SEMMA y ViBlioSOM. Las dos primeras son utilizadas por importantes empresas, mientras que la tercera es utilizada en el Laboratorio de Dinámica No Lineal de la Facultad de Ciencias de la Universidad Nacional Autónoma de México para el análisis y evaluación de información proveniente de bases de datos científicas - tecnológicas.

1.5.1 CRISP - DM

La metodología CRISP-DM[29] [29] fue propuesta en el año de 1999 por un grupo importante de empresas europeas, entre ellas, Daimler Chrysler AG (Alemania), NCR Systems Engineering Copenhague (Dinamarca), SPSS (Inglaterra) y OHRA (Holanda).

La metodología CRISP-DM, consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. A nivel más general, el proceso está organizado en seis fases (Figura I.0.11), estando cada fase a su vez estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, por ejemplo, si en el segundo nivel se tiene la tarea general “limpieza de datos”, en el tercer nivel se dicen las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, “limpieza de datos numéricos”, o “limpieza de datos categóricos”. El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de minería de datos específico.

[29] **Cr**oss–**I**ndustry **S**tandard **P**rocess Model for **D**ata **M**ining

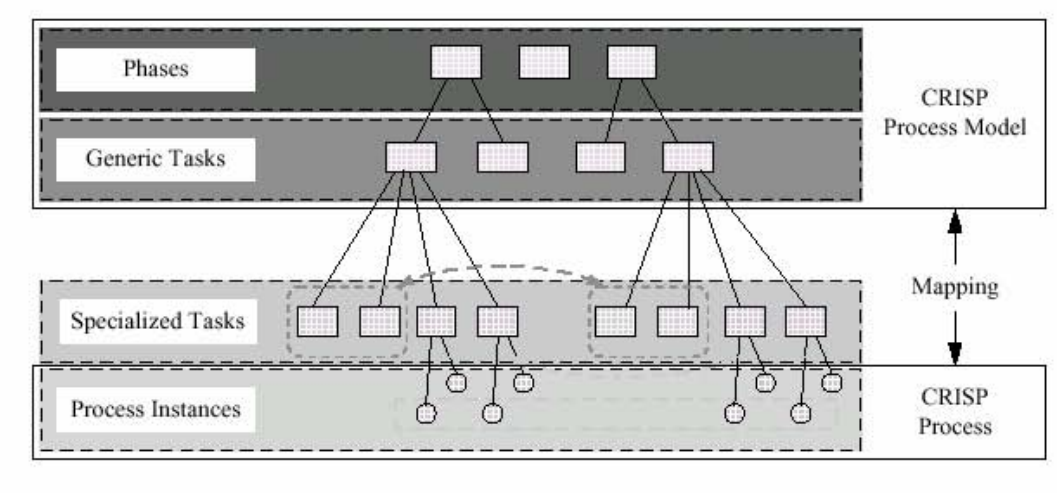


Figura I.0.11: Niveles de abstracción de la metodología CRISP – DM.

La metodología CRISP-DM proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de minería de datos: el modelo de referencia y la guía del usuario. El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de minería de datos en general. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de minería de datos específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase. La metodología CRISP-DM estructura el ciclo de vida de un proyecto de minería de datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

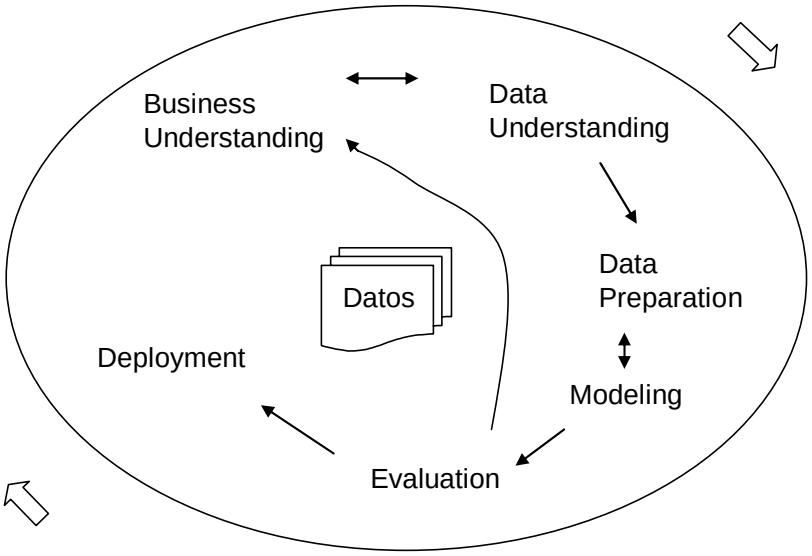


Figura I.0.12: Las fases del proceso de modelado CRISP – DM.

El ciclo de vida del proyecto de minería de datos consiste de seis fases. La figura de arriba muestra las fases del proceso de minería de datos. La secuencia de las fases no es estricta. El movimiento entre fases anteriores y posteriores es siempre necesario. Este movimiento se debe a los resultados de cada fase. A continuación se describe brevemente en que consiste cada fase:

- *Entendimiento del negocio (Business Understanding).* La fase inicial se enfoca en entender los objetivos del proyecto y requerimientos desde una perspectiva de negocios y entonces convertir este conocimiento en un problema de minería de datos y diseñar un plan preliminar para lograr los objetivos.

- *Entendimiento de los datos (Data understanding).* La fase de entender los datos empieza con la colección de datos y proseguir con actividades para conseguir familiarizarse con los datos, para identificar los problemas de calidad de los datos, descubrir visualizaciones de los datos, o detectar subconjuntos interesantes para formar hipótesis.

- *Preparación de los datos (Data preparation).* La fase de preparación de datos cubre todas las actividades para construir el conjunto de datos final (datos que alimentarán a las herramientas de modelado). Las

tareas de preparación de datos son probablemente realizadas muchas veces, tal vez no en orden. Las tareas incluyen: tablas, registros, selección de atributos, transformación y limpieza de datos para las herramientas de modelado.

- *Modelado (Modeling).* En esta fase, varias técnicas de modelado son seleccionadas y aplicadas con sus parámetros calibrados en valores óptimos. Típicamente, existen muchas técnicas aplicables al mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto frecuentemente es necesario regresar a la fase de la preparación de datos.

- *Evaluación (Evaluation).* Se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo.

- *Implementación (Deployment).* Normalmente los proyectos de minería de datos no terminan en la implantación del modelo, sino que se deben documentar y presentar los resultados de manera comprensible con el objeto de lograr un incremento del conocimiento. Además en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

1.5.2 SEMMA

SAS Institute [30] desarrolló la metodología SEMMA[30] para ser utilizada en el mundo de los negocios. SEMMA se define como el proceso de selección, exploración y modelado de grandes conjuntos de datos para descubrir patrones desconocidos en las áreas de negocios.

La figura I.0.13 muestra la dinámica de la metodología SEMMA. En resumen, la dinámica se inicia con la extracción de la *población muestral* sobre

[30] El nombre de esta metodología es el acrónimo correspondiente a las cinco fases básicas del proceso: **S**ample (Muestreo), **E**xplore (Exploración), **M**odify (Manipulación), **M**odel (Modelado) y por último **A**sses (Valoración)

la que se va a aplicar el análisis. Una vez determinada una muestra o conjunto de muestras representativas de la población en estudio, la metodología SEMMA indica que se debe proceder a una exploración de la información disponible. Con el fin de simplificar en lo posible el problema y optimizar la eficiencia del modelo.

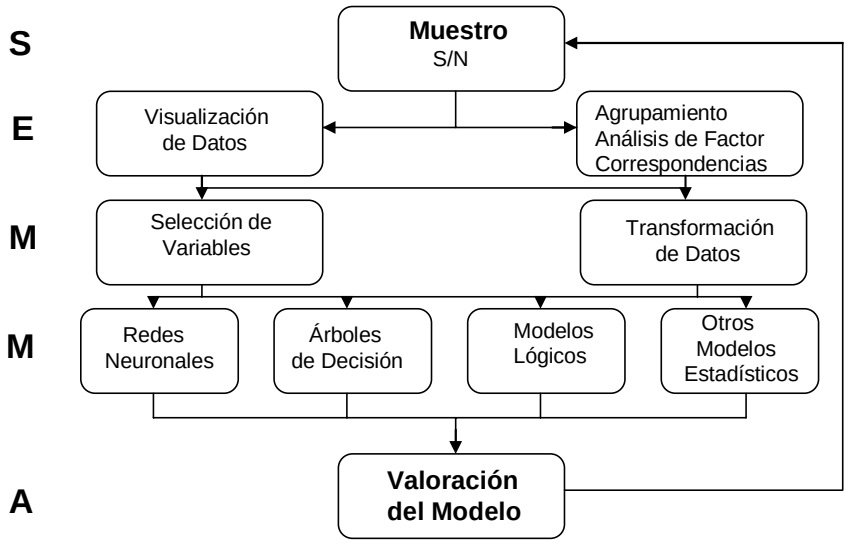


Figura I.0.13: Dinámica de la metodología SEMMA.

La tercera fase de la metodología consiste en la manipulación de los datos, con base a la exploración realizada, de forma que se definan y tengan el formato adecuado de los datos que serán introducidos en el modelo. Una vez que se han definido las entradas del modelo, con el formato adecuado para la aplicación de la técnica de modelado, se procede al análisis y modelado de los datos. El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. Finalmente, la última fase del proceso, consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos, contrastando con otros métodos estadísticos o con nuevas *poblaciones muestrales*.

Estas dos metodologías están implementadas en programas comerciales. La metodología CRISP–DM se implementó en *Clementine*, rutina

que forma parte de SPSS, mientras que SEMMA se implementó en *SAS Enterprise Miner*, rutina que forma parte de Statistica 6.0

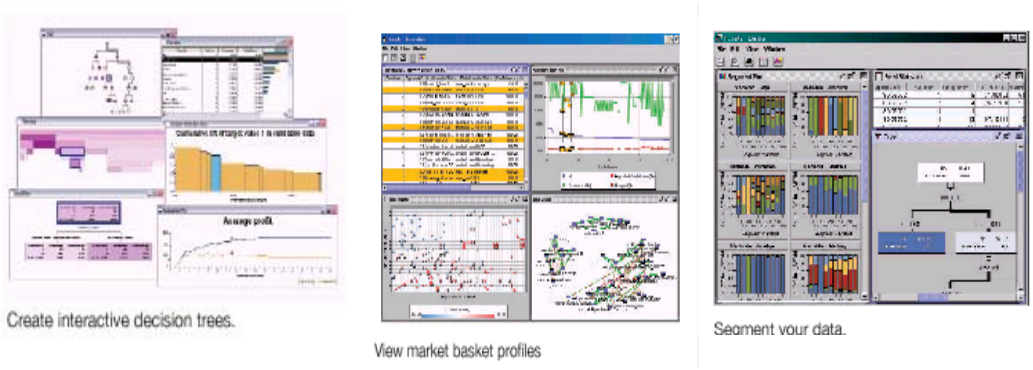


Figura I.0.14: Interfaces de los sistemas CRISP–DM y SEMMA.

1.5.3 Metodología ViBlioSOM para la Minería de Artículos Científicos

El ViBlioSOM[31] es una metodología propuesta para el análisis y evaluación de la Información Científica y Tecnológica. Se ha bautizado ViBlioSOM al incluirse el algoritmo de los Mapas Auto-Organizados para la visualización de grandes volúmenes de datos multidimensionales.

Esta metodología ha causado un gran impacto entre los analistas de información porque permite obtener *conocimiento* cuando se realizan tareas de información científica, gestión del conocimiento, inteligencia empresarial, vigilancia científico–tecnológica, entre otras [25].

El ViBlioSOM se utiliza en el Laboratorio de Dinámica No Lineal del Departamento de Matemáticas de la Facultad de Ciencias de la Universidad Nacional Autónoma de México para el análisis y evaluación de información proveniente de bases de datos científico-tecnológicas.

Esta metodología se ha concebido como un proceso que integre las etapas generales del descubrimiento de conocimiento[32] para realizar en

[31] Acrónimo de **Vi**sualización-Bi**blio**métrica-**S**elf-**O**rganizing **M**aps
[32] i.e. preprocesamiento, extracción de patrones y posprocesamiento. Ver sección 1.2

forma automática e inteligente[33] los análisis y evaluaciones de la información científica y tecnológica.

A partir de estas etapas generales se proponen las cuatro etapas específicas del ViBlioSOM.

- Adquisición y Selección de Documentos.
- Preprocesamiento.
- Minería de Datos y Textos a partir de Indicadores Bibliométricos.
- Visualización e Interpretación de los Resultados.

Con estas cuatro etapas se pretende que la metodología ViBlioSOM sea iterativa e interactiva con los analistas de información científica y tecnológica.

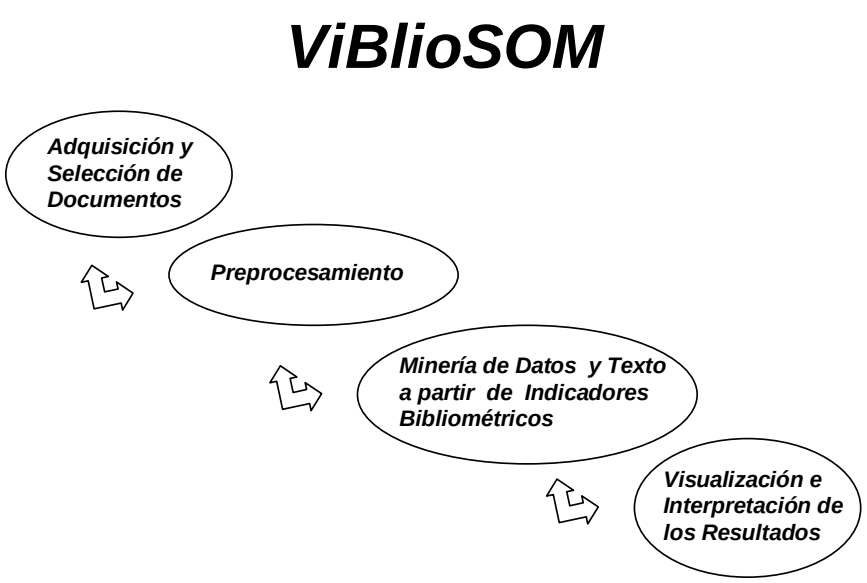


Figura I.0.15: Etapas de la Metodología ViBlioSOM.

Antes de iniciar la metodología ViBlioSOM es necesario precisar: el contexto del dominio de aplicación, formar una perspectiva de nuestro conocimiento; identificar la meta del análisis desde el punto de vista del

[33] i.e., usando técnicas de aprendizaje inteligente que examinan los datos a través de procesos automatizados

cliente. Además, se recomienda trabajar en colaboración con los especialistas del tema, es decir con las personas claves que liderean los temas de investigación, desarrollo o innovación[34].

En resumen, cada etapa consiste en:

- *Adquisición y selección de datos:* Los Análisis Bibliométricos requieren de seleccionar las bases de datos científicas o tecnológicas más representativas del tema. Con los artículos científicos o las patentes seleccionadas haremos un *almacén*[35] de datos. A partir de este almacén seleccionaremos los campos idóneos a nuestro análisis.

- *Preprocesamiento:* Algunos campos como son los nombres de autores, los nombres de instituciones, necesitan ser normalizados o estandarizados antes de iniciar cualquier análisis. Durante esta etapa algunas de las operaciones básicas del preprocesamiento como son la limpieza, la integración, la selección, etc., serán ampliamente utilizadas. La normalización beneficiará directamente a los algoritmos de la Minería de Datos.

- *Minería de Datos y Textos:* La tarea[36] de la minería de datos consiste en formar conglomerados de artículos o patentes. El SOM[37] es altamente eficiente para realizar dicha tarea.

- *Visualización e interpretación de resultados:* EL SOM proporciona un mapa bidimensional, el cual posee la característica de preservar la topología de los datos multidimensionales en forma de vecindades en el mapa. Esta característica permite al analista detectar relaciones intrínsecas de los datos. Lo anterior combinado con la retroalimentación con los especialistas en los temas que se analizan, ofrecerá una mejor visión de los análisis efectuados. Se recomienda que el conocimiento descubierto sea documentado, reportado o comparado con conocimientos obtenidos previamente.

Cada etapa emplea un determinado software. La Adquisición y Selección de Ficheros se hace con el software ProCite. El Preprocesamiento con el Macro de Excel Toolinf. Mientras que la Minería de Datos y Textos a

[34] A esta etapa previa se le conoce como "Comprensión del Campo de Aplicación".
[35] i.e. un Data Warehouses o un Data Marts.
[36] Ver sección 1.2
[37] Ver capitulo 4

partir de Indicadores Bibliométricos; y la Visualización e Interpretación de los Resultados se hace con el software Viscovery SOMine.

En el manual de ViBlioSOM [31] se expone de forma general el uso de los softwares antes mencionados para realizar la metodología ViBlioSOM. En el Laboratorio de Dinámica No Lineal se está desarrollando un nuevo sistema de software llamado ViBlioSOM. Este software integra todas las herramientas necesarias para realizar la metodología ViBlioSOM.

Como se habrá notado durante la exposición de las etapas que integran ViBlioSOM, se hizo uso de términos propios de la Bibliométria, de Bases de Datos Científico-Tecnológicas y por último del SOM. Debido al deseo de resaltar su función dentro de cada etapa.

En los tres capítulos siguientes se abordan (de manera general) estos temas con la finalidad de mostrar su importancia dentro de la metodología ViBlioSOM. Y finalmente, en el capitulo 5 se expone un ejemplo utilizando esta metodología.

Capítulo II

Cienciometría y Bibliometría

El análisis de la información forma parte esencial de las actividades de gestión y toma de decisiones de cualquier institución. Una vía para realizar este análisis son los indicadores bibliométricos y cienciométricos a través de los cuales se pueden descubrir conocimientos relevantes que se encuentran en grandes bases de datos.

La *Bibliometría*[38] [32] como concepto engloba el estudio de los aspectos cuantitativos de la producción, diseminación y uso de la información que se tiene registrada, para lo cual desarrolla modelos y medidas matemáticos que sirven para hacer pronósticos y tomar decisiones en torno a estos procesos. La *Cienciometría*[39] [32] es una disciplina que estudia los aspectos cuantitativos de la ciencia como disciplina o actividad económica. Es parte de la sociología de la ciencia y tiene aplicación en el establecimiento de las políticas científicas e incluye, entre otras, la de publicación, por lo que tiene cierta área común con la Bibliometría.

Vale la pena mencionar que existe además un campo más amplio relacionado con la Bibliometría y que a veces se considera que engloba a la Cienciometría llamado *Informetría*[40].

La *Informétria* [32] estudia los aspectos cuantitativos de la información en cualquier forma, no sólo la compilada en registros bibliográficos; y abarca cualquier grupo social, por lo que no se limita solamente al científico. Puede incorporar, utilizar y ampliar los diversos estudios de evaluación de la información que se encuentran fuera del área de interés de la Bibliometría y de la Cienciometría.

La figura II.0.1 presenta las interrelaciones de estas técnicas cuantitativas dentro del área de las ciencias de la información.

---

[38] Bibliometrics
[39] Scientometrics
[40] Informetrie

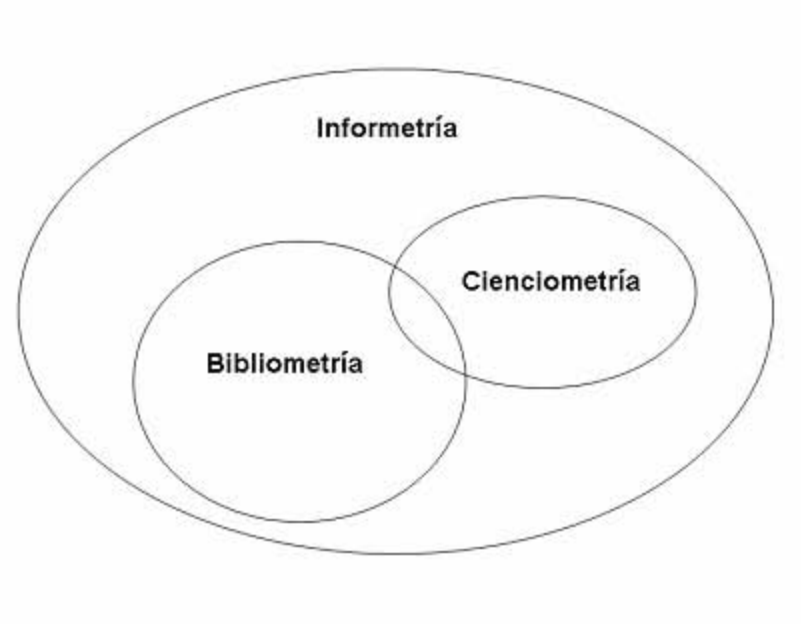


Figura II.0.1: Las técnicas cuantitativas.

Los especialistas que aplican los análisis bibliométricos y cienciométricos han orientado sus estudios, con los modelos y medidas matemáticos a áreas bien definidas, entre las que sobresalen [33]:

- Los aspectos estadísticos del lenguaje y la frecuencia de uso de las palabras y frases, tanto en textos redactados en lenguaje natural como en otros medios impresos y electrónicos.

- Las características de la productividad autoral, medida por el número de documentos publicados o por el grado de colaboración.

- Las características de las fuentes publicadas, incluyendo la distribución de los documentos por disciplinas.

- Los análisis de citas, teniendo en cuenta la distribución por autores, por tipo de documento, por instituciones y por países.

- El uso de la información registrada, a partir de su demanda y circulación.

- La obsolescencia de la literatura, en virtud de la medición de su uso y de la frecuencia con que se cita.

- El incremento de la literatura por temas.

2.1 Antecedentes históricos

Bibliometría

Tradicionalmente se le atribuye a Alan Pritchard [34] la paternidad del termino Bibliometría a partir de su trabajo de 1969 llamado *¿Bibliografía Estadística o Bibliometria?* Sin embargo, fue Otlet [35] y [36], en el año 1934, el primer investigador que aplicó el nombre de *Bibliometrie* a la técnica que trataba de cuantificar la Ciencia y a los Científicos. Éste teórico, pionero de las Ciencias de la Documentación, insiste en diferenciar la Bibliometría de la Bibliografía Estadística, ya que desde el origen, la medida o cuantificación de la ciencia se realizaba utilizando técnicas estadísticas que se aplicaban a las fuentes de información.

Resulta obligado mencionar aquí los aportes de J. D. de Solla Price, eminente científico norteamericano quien a partir de la década de 1960 complementó de forma decidida el marco teórico de la Bibliometría [37]. Estudioso de la *Ciencia de la Ciencia*[41], en la cual aplicó los recursos y métodos científicos al estudio de la propia ciencia. Para esto, se parte del hecho de que para su desarrollo, la ciencia necesita recursos (humanos, financieros, materiales e informativos) a partir de los cuales ofrece sus resultados, entre otros las publicaciones (artículos de revistas y patentes por solo mencionar dos de las publicaciones más paradigmáticas). Lo anterior dio lugar a lo que posteriormente los rusos Nalimov y Dobrov, acuñaran como *Naukometriya*[42].

Cienciometría

La Cienciometría como se le conoce actualmente, [38] es el resultado de la convergencia de dos movimientos. En la antigua URSS se desarrolló el movimiento *Naukovodemia*[43], con el objetivo de estudiar científicamente la

[41] A partir de los años sesenta, aparece la denominada "Ciencia de la Ciencia", que nace en la confluencia de la documentación científica, la sociología de la ciencia y la historia social de la ciencia, con el objeto de estudiar la actividad científica como fenómeno social y mediante indicadores y modelos matemáticos. Esta área dará origen a lo que hoy día se conoce como "Estudios Sociales de la Ciencia", campo de carácter claramente interdisciplinario, que se nutre de los recursos técnicos y conceptuales de distintas disciplinas, entre las cuales se encuentra la Bibliometría.

[42] Palabra rusa para referirse a Scientometrics.

[43] Nalimov dirige la escuela de Moscú mientras que Dobrov dirige la escuela de Kiev. Ambos tienen posturas distintas sobre la Ciencia de la Ciencia.

actividad de la investigación para favorecer su desarrollo. Su primera publicación tuvo lugar en 1926 con un artículo de Borichevski en el que se anuncia la constitución de un nuevo campo de investigación enfocado hacia el estudio de la naturaleza intrínseca de la ciencia. El campo se bautiza como *Naukometriya*, término acuñado por Nalimov y Dobrov[44].

Mientras que en los Estados Unidos de América, surge un movimiento similar al de contra parte soviética. El máximo representante de este movimiento es J. D. de Solla Price, quien en su obra titulada "Little Science, Big Science", propone toda una metodología para el análisis de la ciencia.

La Cienciometría [39] alcanzó su máxima popularidad en 1977, con el surgimiento de la revista *Scientometrics*. Inicialmente publicada en Budapest, Hungría, por la editorial Akadémiai Kiadó, y después en Amsterdam, Holanda, por la Editorial Kluwer Academic Publishers. Actualmente es una producción conjunta de ambas editoriales.

Las primeras definiciones consideraban a la Cienciometría [32] como "la medición del proceso informático" donde el término "informático" significaba, a diferencia de hoy, "la disciplina científica que estudia la estructura y las propiedades de la información científica y las leyes del proceso de comunicación".

Para Van Raan [40], la Cienciometría se dedica a realizar estudios cuantitativos en ciencia y tecnología y a descubrir los lazos existentes entre ambas, apuntando al avance del conocimiento y tratando de relacionar éste con cuestiones sociales y de políticas públicas. La Cienciometría tendría, por lo tanto, un carácter multidisciplinario en lo que se refiere a los métodos que utiliza. Tales métodos provienen tanto de las ciencias naturales como de las ciencias sociales y del comportamiento (estadística y otros métodos matemáticos, modelos sociológicos, investigaciones y métodos psicológicos de entrevista, informática, filosofía de la ciencia, lingüística, etc.)

Informetría

La Informetría[45] [32] estudia los aspectos cuantitativos de la información en cualquier forma, no sólo la compilada en registros bibliográficos, y abarca cualquier grupo social por lo que no se limita sólo al científico.

[44] Ambos, Nalimov y Dobrov fueron influenciados por el trabajo de Solla Price
[45] Propuesta por el director del Institut fur Informetrie de Alemania Otto Nacke en 1979.

Puede incorporar, utilizar y ampliar los diversos estudios de evaluación de la información que se encuentran fuera de la Bibliometría y de la Cienciometría. Sus aplicaciones prácticas son disímiles: la recuperación de información, la administración de bibliotecas, la historia de las ciencias y las políticas científicas de una institución o gobierno. Su alcance es, por tanto, teórico-práctico, pues si bien se enfatiza, en primera instancia, el desarrollo de modelos matemáticos, concentra también su atención en la obtención de medidas para los diferentes fenómenos que estudia.

En la actualidad existen proyectos a nivel internacional con el objetivo de seguir desarrollando los Análisis Bibliométricos, Cienciométricos e Informétricos. Por mencionar alguno, el proyecto europeo *CORTEX: Neuromimetic intelligence (project-team)* [41] dirigido por el Institut Nacional de Recherche en Informatique en Automatique, INRIA de Francia. Tiene el objetivo de desarrollar instrumentos automáticos que permitan análisis inteligentes no solamente de la literatura científica sino también de la literatura técnica.

En México, los Análisis Bibliométricos los realizan instituciones gubernamentales, entre las que se encuentran: La Academia Nacional de Medicina, Centro Medico la Raza, El Colegio de México, Instituto Politécnico Nacional, Universidad de Guanajuato, y varios institutos pertenecientes a la Universidad Nacional Autónoma de México. Los temas abordados principalmente para la realización de Análisis Bibliométricos son temas médicos [42].

Además, en la Universidad Nacional Autónoma de México, UNAM, está en desarrollo el *Proyecto universitario de tecnologías de la información y computación: tecnologías para la universidad de la información y la computación* que forma parte del Programa Transdisciplinario en Investigación y Desarrollo de la UNAM [43]. Entre los objetivos del proyecto está la creación de un Observatorio Informétrico. En donde se llevarán a cabo proyectos para el análisis de información e investigación cienciométrica de interés nacional y estudios sobre la universidad.

Las revistas internacionales [44] más importantes que publican trabajos relacionados con Bibliometría, Cienciometría e Informetría son las siguientes: Bulletin of the Medical Library Association, Information Processing & Management, Interciencia, International Journal of Scientometrics and Informetrics, International Society for Scientometrics and Informetrics Proceedings, Journal of Documentation. Journal of Information Science, Journal of the American Society for Information Science and Technology,

Rapport de l´Obervatoire des Sciences et des Techniques, Research Evaluation, Research Policy, Revista Española de Documentación Científica, Revue Francaise de Bibliometrie y Cahiers, Science & Public Policy, Scientometrics, Social Studies of Science.

El constante crecimiento de la información y de los conocimientos reflejados en publicaciones científicas y técnicas i.e. artículos, patentes, etc., les impone a sus usuarios (investigadores, agentes de servicios de información, etc.) nuevos requerimientos, marcados por la impronta de las nuevas tecnologías de la información. Felizmente, también estas tecnologías complementan a otras técnicas y metodologías, brindando la oportunidad de hacerle frente a tales desafíos. La Bibliometría se inserta en este marco como una disciplina relativamente nueva cuya actual denominación tiene escasamente treinta años. Aún cuando su gestación se inició a comienzos del Siglo XX, su auge es reciente. La misma ha influido sobremanera en la Cienciometría.

La Bibliometría ha capitalizado la esencia de la cuantificación de la información para diferentes fines. De una parte, se destacan aquellas aplicaciones relacionadas con la gestión y uso de la información en bibliotecas donde surgió. Por otra parte, de la Bibliometría se han derivado aplicaciones en el campo de la política científica en lo que se conoce como *Bibliometría Evaluativa*.

En su lógico desarrollo, mucho más recientemente la Bibliometría también aparece en la encrucijada de lo que se conoce como vigilancia tecnológica, como apoyo a la toma de decisiones en ambientes empresariales (producción y servicios), movidos por los desarrollos científicos y tecnológicos.

En ese ámbito, la Bibliometría han encontrado en el Descubrimiento de Conocimiento en Bases de Datos (en particular en la Minería de Textos) uno de sus modos de realización más prometedores en la actualidad. Para ello, se apropia de métodos estadísticos ya establecidos, y añade constantemente métodos más modernos, tales como las Redes Neuronales Artificiales para cumplir sus objetivos.

Lo anterior hace que la Bibliometría vaya de la mano de las tecnologías necesarias, que nos permitan aproximarnos no sólo a interpretar esa realidad que son los conocimientos (certificados) reflejados en las publicaciones, sino que nos alimenta para transformar dicha realidad. Para los hombres y las mujeres de ciencia y para quienes sirven a la ciencia como son los

bibliotecarios y los especialistas en información, resulta obligado conocer la Bibliometría como parte de la metodología de investigación, como medio para hacer a la investigación científica y al desarrollo tecnológico más productivos y eficientes. La tabla II.0.1 da una idea de los objetos de estudio, variables y métodos que utilizan estas técnicas cuantitativas [32].

TIPOLOGÍA	BIBLIOMETRÍA	CIENCIOMETRÍA	INFORMETRÍA
Objeto de Estudio	Libros, documentos, revistas, artículos, autores, usuarios.	Disciplinas, temas, áreas y campos científicos y tecnológicos. Patentes, disertaciones y tesis.	Palabras, documentos, bases de datos, comunicaciones informales (incluso en ámbitos no científicos),
Variables	Números en circulación (prestamos) y de citas, frecuencia de aparición de palabras, longitud de las oraciones, etc.	Aspectos que diferencian a las disciplinas y las subdisciplinas. Revistas, autores, trabajos, formar en que se comunican los científicos.	Difiere de la Cienciometría en los propósitos de las variables, por ejemplo, medir la recuperación, la relevancia, el recordatorio, etc.
Métodos	Ranking, frecuencia, distribución.	Análisis de conjunto y de correspondencia, coaparición de términos , expresiones, palabras claves, etc.	Modelos Vector-espacio, modelos voléanos de recuperación, modelos probabilísticos, lenguaje de procesamiento, abordajes basados en el conocimiento, tesauros.
Objetivos	Pronosticar y en la tomar decisiones Asignar recursos, tiempo, dinero, etc.	Identificar esferas de interés, donde se encuentran las materias; comprender cómo y con qué frecuencia se comunican los científicos.	Aumentar la eficiencia de la recuperación de información. Identificar estructuras y relaciones dentro de los diversos sistemas de información.

Tabla II.0.1: Tipología de las Ciencias Cuantitativas.

En resumen, es indudable la existencia de un alto nivel de solapamiento entre ellas, principalmente en el flujo del *conocimiento/información* y en los métodos y modelos matemáticos afines, sin embargo, cada una tiene su propio objeto y tema de estudio específico. Las divergencias se centran en torno a ciertos aspectos: los límites de la misma, los objetivos que pretende alcanzar y sobre la naturaleza y pertinencia de los datos sobre los que trabaja. Se concluye que los objetos de estudio de estas disciplinas se definen por las ciencias a las que sirven de instrumento. *La Bibliometría es la disciplina instrumental de la Bibliotecología, en tanto, la Cienciometría lo es de la Cienciología, y la Informetría, de las Ciencias de la Información* [39].

El estudio de la ciencia por medio de análisis cuantitativos, requiere un desarrollo teórico que explique su dinámica dentro de la comunidad científica y fuera de ella, es decir, en la sociedad. En esta sección se expone en forma general dos aspectos teóricos de la dinámica de la ciencia y posteriormente se expone un modelo teórico que nos permitirá comprender el funcionamiento de la ciencia en la sociedad.

2.2 Modelo estadístico de la ciencia

Los análisis cuantitativos de la ciencia, requieren desarrollar modelos estadísticos que permitan aplicar técnicas matemáticas al estudio de la ciencia. Muchos investigadores de esta área consideran *al artículo científico como un indicador de la producción de la investigación científica*[46]. Desde esta concepción, la literatura científica se presta para su conteo, su clasificación y su representación bajo la forma de series temporales [45] y [46].

El modelo de la ciencia que sirve de paradigma es el de su representación como una población de publicaciones donde cada documento es considerado como un átomo de conocimiento, en tanto que cada artículo representa un *"quantum"* de información científica. No obstante, resulta importante subrayar, que "*documento"* y "*conocimiento"* no constituyen entidades idénticas.

Otro punto importante que permite desarrollar un modelo estadístico de la ciencia, es su *estructura acumulativa*. Los investigadores han identificado cuatro estados de esta *estructura acumulativa*, que son los siguientes:

a) Existe un estado inicial, es decir, hubo precursores.

[46] Conocido como "Reduccionismo Bibliométrico".

b) La ciencia crece exponencialmente, esto se puede interpretar como un estado de expansión de la ciencia. Este estado se modela con la siguiente expresión matemática:

$$f(t) = ae^{bt}$$

Siendo a la dimensión inicial del corpus al tiempo $t = 0$ y b la tasa continua de crecimiento que refleja el porcentaje de crecimiento de corpus por unidad de tiempo.

c) Inmediatamente después del estado de expansión, la ciencia pasa a un estado de crecimiento lineal.

d) Por último, la ciencia entra a un estado de saturación. Este estado se modela con la siguiente expresión matemática llamada curva logística.

$$g(t) = \frac{k}{1 + ae^{bt}}$$

En donde $g(t)$ representa el tamaño del corpus al tiempo t y k representa el límite superior.

En la figura II.0.2 se muestra como se relacionan los cuatro estados. El estado de precursores es el origen. Después se inicia el estado de expansión de la ciencia. Inmediatamente de este estado la ciencia pasa a un estado de crecimiento lineal. Y por último, la ciencia entra a un estado de saturación.

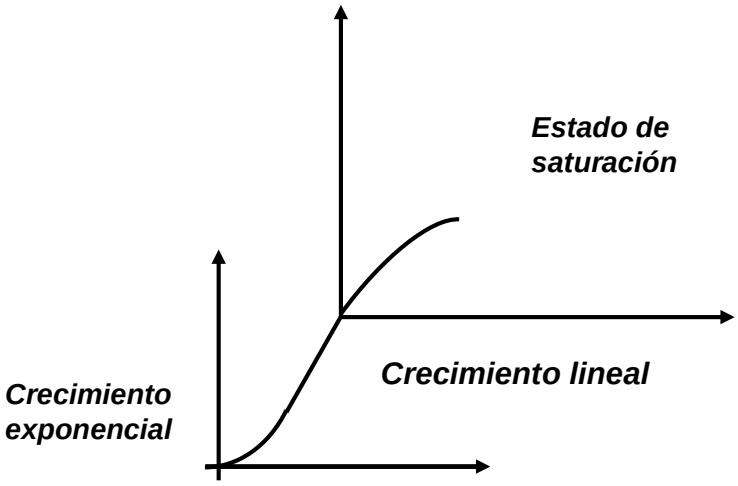


Figura II.0.2: Estados de la estructura acumulativa de la Ciencia.

Por otra parte, el *comportamiento estadístico* de la estructura acumulativa de la ciencia se modela con funciones hiperbólicas de la forma:

$$x^n y = k$$

Donde x, y son variables y n, k son constantes. De este modo, las distintas fases de generación, diseminación y utilización de la información tienen comportamiento estadístico.

Un principio muy importante relacionado con estas funciones es la Ventaja Acumulativa [37]. Con él se hace referencia al dicho: *el éxito, genera éxito*. Este principio establece que: "...en tanto que una fuente tiene más objetos, entonces es mayor la probabilidad de que esa fuente produzca otro objeto; aunque existe siempre una (pequeña) probabilidad de que una fuente sin objetos produzca un primer objeto".

El principio anterior subyace en el *Proceso de Producción de Información*. En este proceso intervienen dos elementos: *las fuentes, y los objetos que producen estas fuentes*. Así por ejemplo: los autores o grupos de autores (considerados fuentes) que producen determinado número de publicaciones (objetos o items); las fuentes (fuentes) que producen artículos (objetos); y los

artículos (fuentes) con las citas (objetos) que reciben en determinado período de tiempo, etc.

Otro aspecto importante lo constituye el *Sistema de Rangos* [37] utilizado. En lo fundamental se asume que el rango de las fuentes en un *Proceso de Producción de Información* está dado porque la fuente más productiva recibe el rango 1, la siguiente el rango 2 y así sucesivamente. El último rango T será de la fuente de menor producción.

Y para concluir con esta sección se exponen dos funciones hiperbólicas[47] que muestran este comportamiento estadístico de la ciencia. Para Gorbea [47] es más apropiado referirse a estas funciones como modelos matemáticos que como leyes.

EL modelo matematico de Bradford es simplemente la descripción de una relación cuantitativa entre revistas y artículos contenidos en una bibliografía especializada que necesariamente cubre un determinado periodo [48] y [49].

El modelo se deduce de la siguiente observacion: Si las revistas científicas se ordenan por la producción decreciente de artículos sobre una materia determinada, podría dividirse en un núcleo de publicaciones altamente dedicadas a la materia y en varios grupos o zonas que contengan el mismo número de artículos que el núcleo, siendo los números de revistas en el núcleo y en las zonas subsiguientes de la forma de 1, k, k^2,...

Según los expertos [33], el modelo plantea una forma longitudinal acumulativa de distribución de los documentos, por disciplinas en las publicaciones seriadas (Ventaja Acumulativa) e introduce la idea de una serie geométrica, que representa el número creciente de revistas desde el núcleo hacia las zonas adyacentes en una temática, donde el núcleo y las zonas contienen respectivamente igual número de documentos en orden decreciente según revista.

El modelo de Bradford se interpreta de la siguiente manera: los artículos sobre un tema se concentran en un número reducido de revistas (llamado *núcleo*) y el resto en una serie más amplia de ellas, muchas sin conexión directa con la disciplina (llamada *dispersión*). Es decir, que el núcleo contiene aquellas revistas que tienden a publicar el mayor número de artículos dedicados al asunto. El modelo se reformularse en la siguiente expresión matemática:

[47] Los modelos matemáticos de Bradford, de Lotka y de Zipf son los más conocidos.

$$N_r = k^r N_n$$

En donde, N_r representa el número de revistas en el *r-ésimo* grupo. k^r es una constante y N_n representa el número de revistas en el núcleo.

El modelo tiene un gran problema debido a que evalúa en las mismas condiciones revistas que son desiguales en varios aspectos; el período de participación, las frecuencias de publicación y el número de fascículos publicados. Por consiguiente, el modelo *homogeneiza* lo que es naturalmente *heterogéneo*, introduciendo una estática en un proceso que es dinámico y que está en permanente movimiento.

En las técnicas cuantitativas [50] la productividad generalmente se mide a través del número de publicaciones producidas por un científico, un grupo de científicos, una institución o un país, en un período. En este trabajo sólo se hablará de la productividad de los científicos. La medición de la productividad de los grupos o el de las instituciones, se basa en variantes del modelo que se presenta a continuación.

El modelo matematico de Lotka es simplemente la descripción de una relación cuantitativa entre los científicos y los artículos producidos en un campo dado y en cierto periodo. Y afirma que el número de científicos que hacen n contribuciones en una determinada área científica, es aproximado a $1/n^2$, el de aquellos que hacen una sola contribución. Lo anterior se reformula en la siguiente expresión matemática:

$$p(n) \cdot n^a = k$$

Donde k es constante y $p(n)$ representa el número de científicos que producen n artículos. Se ha encontrado que el valor de la constante k es aproximadamente igual a 0.6079, lo cual se traduce en que la proporción de aquellos científicos que publican un único artículo es de, más o menos, el 60%. (Siendo a aproximadamente igual a 2).

Si se considera que el modelo es adecuado, se afirma que los trabajos científicos no se distribuyen aleatoriamente, están concentrados en una porción de autores altamente productivos; cuántos más trabajos tiene un autor, más facilidad parece tener para producir otros (Ventaja Acumulativa); y existe un pequeño grupo de científicos muy productivos y una masa de científicos que lo son mucho menos.

El modelo de Lotka debe emplearse con sumo cuidado, pues en la mayoría de los casos suele confundirse productividad con calidad de la investigación publicada, lo cual, es una enorme equivocación. Por otra parte, resulta natural correlacionar el prestigio de un científico con su productividad, lo cual, puede inducir a clasificar erróneamente al científico.

2.3 El modelo bibliométrico de la ciencia

Una vez establecido un marco estadístico de la ciencia, se esta en condiciones para desarrollar indicadores que analizen en forma cuantitativa a la ciencia. Antes de hacer esto, se debe asumir el *Dogma Central de la Bibliometría*. Este dogma propone asumir ciertos hechos como verdaderos, los cuales, nos permitirán aceptar los resultados de la Bibliometría.

Se puede plantear que el *Dogma Central de la Bibliometría* [37] se basa en postulados y axiomas que subrayan la utilización y validez de los análisis bibliométricos[48]. Los dos postulados son los siguientes:

- Las publicaciones son el resultado de la actividad del pensamiento.

- Las publicaciones son el fruto de la comunión del pensamiento individual y del pensamiento colectivo.

Mientras que los axiomas son: *el axioma de la medida de actividad***,** esta medida plantea que el conteo de artículos y patentes ofrece indicadores válidos de la actividad de I + D (Investigación + Desarrollo) en las áreas temáticas de la temática de artículos y patentes, así como de las instituciones a partir de las cuales tales documentos se originan; *el axioma de la medida del impacto***,** plantea que la cantidad de veces que dichos artículos o patentes son citados en artículos o patentes subsiguientes, ofrece indicadores válidos del impacto o importancia de los artículos o los patentes citados; *el axioma de la medida de conexión***,** plantea que las citas que se hacen desde un artículo a otros artículos, de patentes a patentes y de patentes a artículos, ofrecen indicadores de la conexión intelectual entre las organizaciones que producen los artículos y las patentes, así como la conexión del conocimiento entre sus áreas temáticas.

*El axioma de la medida de relación***,** establece que los fenómenos que ocurren frecuentemente de forma conjunta en algún dominio, se asume que están relacionados y la fortaleza de esa relación se asume que está

[48] ver sección 2.5

relacionada con la frecuencia de la *coocurrencia*. Esta medida es empleada por muchos investigadores, pues permite diseñar mapas de la estructura y evolución de la ciencia a varios niveles de especificidad.

Estos axiomas tienen diferentes grados de validez, los cuales, pueden variar significativamente de acuerdo a los autores, disciplinas técnicas, y organizaciones. Razones histórico - culturales, temas relativos a las clasificaciones, propiedad corporativa, así como muchas otras causas pueden y de hecho contribuyen a que las fuentes públicas de la literatura tengan brechas sustanciales en la información documentada sobre la actividad actual y pasada en campos técnicos específicos. Mientras más puedan servir como muestra representativa del total de la literatura en una disciplina, las fuentes públicas de la literatura, estos axiomas son más valederos.

2.3.1 Los indicadores bibliométricos

De los postulados y medidas anteriores se derivan todos los indicadores que se pueden construir en Bibliometría. Hay que destacar que estos instrumentos se construyen con fines específicos, bajo circunstancias específicas y en ocasiones para casos y objetivos precisos, es decir, desarrollar indicadores en forma abstracta es inútil.

Los indicadores bibliométricos [51] se obtienen a partir de las estadísticas de publicaciones científicas y tecnológicas de los agentes del sistema de ciencia y tecnología, empresas, organismos públicos de investigación, universidades, departamento de estadística, organismos internacionales, etc.

Como estos indicadores se obtienen a partir de las estadísticas de publicaciones científicas y tecnológicas hay que tener en cuenta que: cada publicación no hace el mismo aporte al conocimiento científico y lo variable de los promedios de las publicaciones con respecto a la especialidad y al contexto institucional.

Los indicadores bibliométricos [52] se clasifican en generaciones según el nivel de complejidad que vayan alcanzando y la evolución (según su surgimiento) en el tiempo.

- Indicadores de actividad.
- Indicadores relacionales de primera generación.
- Indicadores relacionales de segunda generación.

En resumen, los indicadores de actividad están orientados a la parte de la evaluación de la investigación, a través de mediciones de la calidad y el impacto de las publicaciones. Mientras que indicadores relacionales de primera y segunda generación están orientados con los aspectos estructurales de la ciencia.

2.3.2 Los indicadores de actividad

Los indicadores de actividad se fundamentan en las técnicas escalares o unidimensionales, es decir, en las ocurrencias o en los simples recuentos de ciertos elementos bibliográficos, tales como las citas, las referencias, etc. pues en principio, los artículos publicados por un autor, las citas que recibe un autor y las referencias utilizadas por parte de un autor, pueden ser representadas por series de tiempo discretas. A través de estas ocurrencias los indicadores de actividad, proporcionan datos sobre el volumen y el impacto de las actividades de investigación.

Entre los indicadores de actividad más utilizados se encuentran [44]: *el número de publicaciones, el número de referencias, el número de citas, la obsolescencia de la literatura científica, el factor de impacto de las revistas, etc.*

El número de publicaciones. El ejercicio bibliométrico más sencillo y teóricamente más objetivo es el conteo del número de publicaciones científicas y/o tecnológicas [53]. El conteo del número de publicaciones científicas y/o tecnológicas, es generalmente, considerado como una medida de la producción científica de un autor, una institución, un sector de actividad, un área geográfica o de un país[49].

El análisis consiste en contar el número de publicaciones científicas y/o tecnológicas, que son atribuidas, a un autor, a una institución, a un sector de actividad, a un área geográfica o a un país, en un área específica. En cifras absolutas, este indicador nos permite efectuar comparaciones con la actividad de otros autores, instituciones, sectores, áreas o países, pues se hace necesario tener un *"marco de referencia"* en el que se pueda ubicar nuestro objeto de estudio. Como mencioné, en cifras absolutas este indicador puede ser interesante, pero la evolución temporal del indicador es mucho más significativa, por ejemplo, el crecimiento de cualquier campo de la ciencia, según la variación cronológica del número de trabajos que se publican en él; la evaluación cronológica de la producción científica, según el año de la publicación de los documentos, y además, la dinámica investigativa

[49] Axioma de la medida de actividad

de un país determinado puede monitorearse y seguirse sus tendencias a través del tiempo.

Entre las limitaciones del uso del número de publicaciones como indicador de producción científica, se pueden mencionar dos básicamente: el carácter cuantitativo del indicador, es decir, que sólo aporta información sobre la cantidad de publicaciones, pero no sobre la calidad intrínseca de la publicación e impacto de la publicación. La otra limitación, es consecuencia de las diferencias existentes, en cuanto a los hábitos de publicación y productividad de los autores, instituciones, etc., en las diversas áreas temáticas (ciencias sociales, ingeniería y tecnología, ciencias naturales y experimentales).

A causa de estas diferencias, no es siempre posible efectuar comparaciones entre autores, instituciones, sectores de actividad, áreas geográficas o países. Tampoco, hay que olvidar que estas diferencias existen también en las distintas disciplinas que componen un área.

La consideración del número de publicaciones como indicador de actividad científica ha desencadenado comportamientos como: el denominado *"síndrome de publicar o perecer"*[50], bajo este nombre se designa la situación actual en la que los científicos se ven presionados a publicar, no sólo para dar a conocer los resultados de su investigación, sino también como la única vía de justificar su actividad y obtener un reconocimiento; una tendencia a aumentar el número de autores por documento, hecho que no siempre se asocia a un aumento real de la colaboración entre autores, sino a la denominada *"autoría gratuita"* y la fragmentación de los trabajos en varias publicaciones que podrían haberse publicado en un solo artículo más completo y más coherente[51].

Entre las medidas tomadas para evitar estas conductas se pueden mencionar todas aquellas orientadas a premiar la calidad frente a la cantidad de publicaciones. En este sentido, es frecuente que en algunos procesos de evaluación de personal investigador, se solicite al científico que adjunte sólo sus tres o cinco publicaciones más relevantes con el fin de centrar la evaluación en la calidad de sus contribuciones.

[50] Síndrome POP o Publish or Perish
[51] Síndrome LPU o Least Publishable Unit

2.3.3 Los indicadores relacionales

Los indicadores relacionales se fundamentan en técnicas bidimensionales, es decir, en la aparición conjunta de ciertos elementos bibliográficos, tales como, las citas, las instituciones, los años de publicación, etc. Pues, en principio, la aparición conjunta de dos indicadores que pueden ser o no de la misma naturaleza, puede ser representada por series de tiempo discretas.

Indicadores relacionales de primera generación: estos indicadores rastrean los lazos y las interacciones entre investigadores y campos para describir los contenidos de las actividades científicas y su evolución. Los indicadores relacionales más destacados son: *las firmas conjuntas, las redes de citas, las referencias comunes, la colaboración en la investigación, las citas comunes, etc*.

Antes del ejemplo, veamos en que consiste una cita y una referencia. Según Price [54], si un trabajo A contiene una nota bibliográfica que utiliza y describe otro trabajo B, entonces, A contiene una *referencia* a B y B recibe una *cita* de A. Y las referencias son mostradas en las notas a pie de página y en la bibliografía del que se adjunta al final del documento A. Además, hay que tener en cuenta que las *referencias* que contienen las publicaciones científicas a trabajos previos son al propio tiempo *citas* desde el punto de vista de éstos. (Figura II.03).

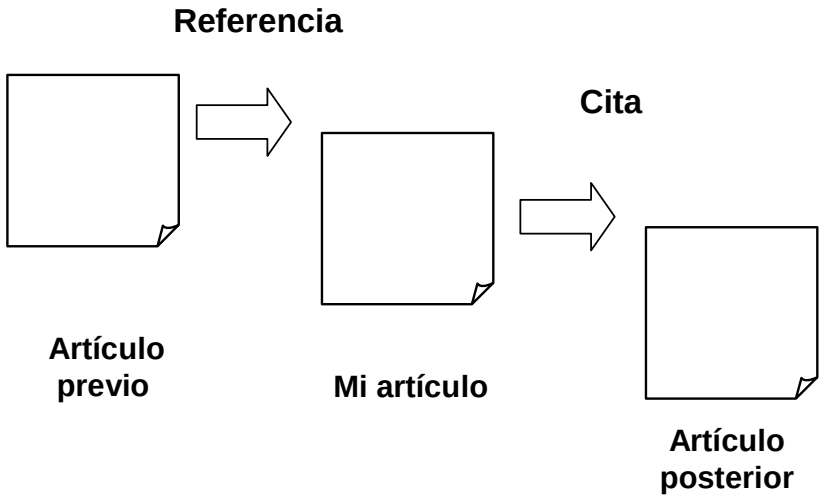


Figura II.0.3: Las referencia y las citas.

Las redes de citas: El conteo del número de citas que una publicación recibe de otras posteriores o el conteo del número de referencias que una publicación hace de otras anteriores, son indicadores validos según *el axioma de la medida del impacto*. Las redes de citas [51] suponen que las citas que hacen unos autores a otros, así como entre revistas o cualquier otro tipo de documentos pueden representarse mediante una estructura en red. (Ver figura II.0.4).

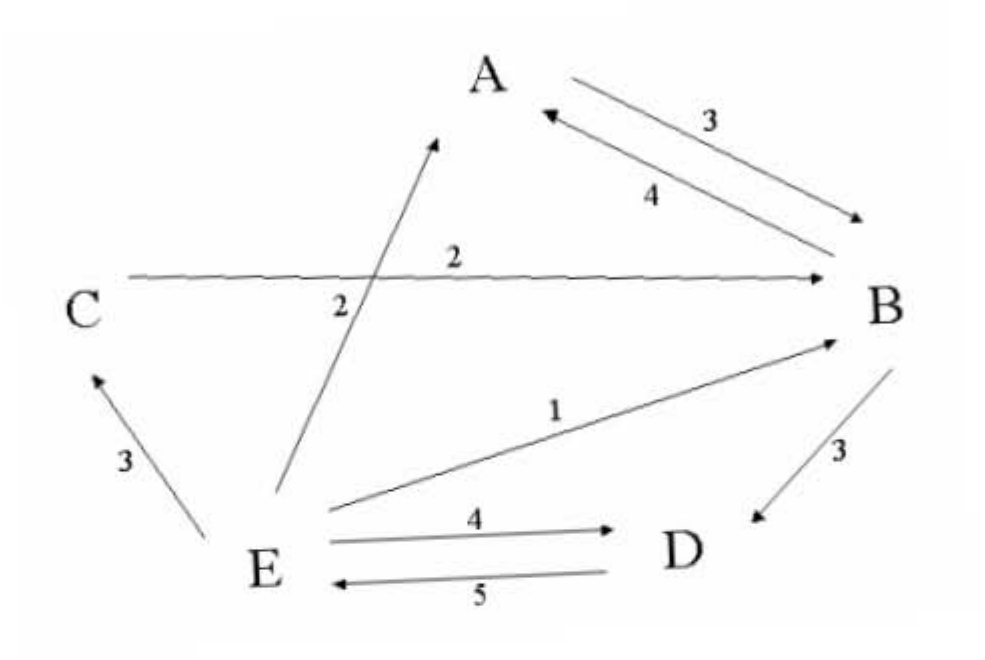


Figura II.0.4: Red de Citas de Autores.

Los elementos básicos de la red son los nodos y las flechas, los cuales se interpretan de la siguiente manera:

- Los nodos representan a las entidades (autores, revistas, patentes, etc.) que citan o son citados

- Las flechas indican la vinculación en el proceso de citación, apuntando hacia la entidad que es citada.

- Las flechas están acompañadas por un número que indica cuantas veces ha sido citada la entidad (nodo). Otros autores indican la frecuencia usando diferentes grosores de líneas.

Dependiendo de la entidad empleada para hacer la red, ésta servirá para: estudios detallados de la red de investigación; identificar transferencias y préstamos de técnicas experimentales; detectar líderes tecnológicos; identificar frentes de investigación; identificar tecnologías o teorías fundacionales; identificar estrategias tecnológicas de las empresas; alianzas tecnológicas; etc.

Indicadores relacionales de segunda generación: Estos indicadores se han creado con el objetivo de *entrar* al contenido de las publicaciones científicas. En general, la forma de *entrar* a los contenidos de los documentos, consiste en seleccionar un conjunto de palabras significativas de ciertos elementos bibliográficos, tales como los títulos de los artículos, los resúmenes, las palabras clave de artículos, los códigos de clasificación, y en última instancia, por relaciones semánticas en los textos. El siguiente indicador es el más utilizado para esta tarea:

Análisis de palabras comunes (Co-Word Analysis): El análisis consiste en la reducción del *conjunto de publicaciones científicas* a un *conjunto de palabras significativas*[52]*,* que describan al mismo tiempo el contenido de los documentos y sus relaciones con otros documentos.

La reducción [51] es posible debido a la siguiente hipótesis: un área de investigación puede ser identificada por su propio vocabulario[53] o más exactamente, por las particulares asociaciones que se establecen entre palabras, pudiendo ser utilizadas algunas de ellas (asociadas a otras palabras) en otros contextos sociales.

Hecho lo anterior, aplicamos *el axioma de la medida de relación*[54] a nuestro conjunto de palabras significativas. El axioma se interpreta de la siguiente manera: *la repetición conjunta de dos palabras significativas en gran cantidad de artículos, indica una relación entre ellas. Y la frecuencia de estas repeticiones mide el grado de relación entre dos documentos.*

[52] Este conjunto de palabras significativas, por lo general, lo vemos en forma de tesauro o códigos de clasificación. Pero ambas formas tienen el objetivo de describir parcialmente el contenido del documento.

[53] Este vocabulario es portador de los conceptos científicos, ideas y conocimientos

[54] Establece que los fenómenos que ocurren frecuentemente de forma conjunta en algún dominio, se asume que están relacionados y la fortaleza de esa relación se asume que está relacionada con la frecuencia de la coocurrencia.

El análisis de palabras comunes se utiliza para producir mapas[55], que describan las asociaciones más significativas de las palabras clave, en un conjunto dado de documentos de alguna especialidad [55]. Su utilidad en los estudios cuantitativos sobre la estructura y desarrollo de la ciencia, radica en predecir tendencias de cambio científico en instituciones o en investigadores; el ciclo de vida de los temas; etc.

Como se ha mencionado, el ViBlioSOM utiliza este indicador. En el capitulo 5 se utiliza dicho indicador con la finalidad de producir mapas, que describan el comportamiento de los temas matemáticos considerados por MedLine en la Biomedicina.

2.4 La actividad científica

La ciencia y la tecnología han adquirido una enorme importancia en la sociedad desde el siglo XX debido, en parte, a la gran influencia que ejercen en el desarrollo económico, político y cultural de los países. Esto hace que las expectativas de bienestar social estén fijadas en ellas, hasta el punto de que se produce una fuerte competencia entre los países por la carrera del desarrollo científico y tecnológico, considerándolo como una de las aspiraciones de la humanidad [55].

El estudio de la ciencia se basa en el modelo teórico llamado *Actividad Científica.* En términos generales se considera *Actividad Científica* a toda actividad sistematizada de impacto académico, social, político, cultural, económico, etc. Esto significa que los proyectos enmarcados en esta tipología, deberán delinearse teniendo en cuenta que la actividad a desarrollar debe formar parte (o ser un antecedente) de un futuro proyecto de investigación. Cabe destacarse que a nivel internacional, la *Investigación Científica* es un subsistema de *Actividad Científica*.

La *Actividad Científica* debe ser vista e interpretada dentro del contexto social en la que está enmarcada. Por ello, las evaluaciones del desempeño científico deben ser sensibles al contexto conceptual, social, económico e histórico de la sociedad donde se actúa. Esto significa que la ciencia no puede ser medida en una escala absoluta, sino en relación con las expectativas que la sociedad en la cual se desarrolla, ha puesto en ella.

La *Actividad Científica* puede ser considerada como un análogo a los modelos teóricos económicos de Entradas-Salidas (Input–Output). Esta

[55] Los sociólogos de la ciencia los llaman Mapas de la Ciencia debido a que permiten profundizar en el estudio de la estructura y dinámica de las áreas científicas.

forma de ver a la *Actividad Científica* puede resultar un poco simplista para algunos científicos pero tiene validez desde el punto de vista económico [56].

Desde este punto de vista, la producción de conocimiento, la investigación, la innovación y el desarrollo de nuevos productos o aplicaciones son procesos en los que hay una inversión a largo, mediano o corto plazo, y se obtienen unos productos (publicaciones, patentes, nuevos productos o procesos, etc.). Por lo tanto, estas actividades son susceptibles de estudios y mediciones.

Figura II.0.5: Entradas y salidas de la Actividad Científica.

Las entradas de la Actividad Científica, por lo general, se clasifican en:

- *Recursos humanos:* cantidad de técnicos académicos, científicos o ingenieros, cantidad de científicos por especialidades, cantidad de científicos por categorías, etc.

- *Recursos materiales:* valor de los inmuebles, número de instituciones dedicadas a la Investigación-Desarrollo, valor de los activos fijos, valor de los insumos, etc.

- *Recursos financieros:* cantidad de recursos dedicados a la actividad de Investigación – Desarrollo (I + D), salario, comunicaciones, información, mercado, etc.

Mientras que algunos resultados de la Actividad Científica son:

- *Publicaciones científicas:* artículos científicos, tesis doctorales, revisiones, cartas, notas, etc.
- *Publicaciones técnicas:* patentes, etc.
- *Premios y reconocimientos científicos de excelencia; trabajos presentados en eventos; invitaciones a eventos y conferencias; etc.*

Para algunos investigadores, las mediciones de esta actividad son sumamente complejas. Spinak [57] expone que la medición de la primera parte (Input) es una tarea más cercana a las ciencias de la economía, la estadística y la administración que, si bien no es simple, dispone desde hace tiempo de metodologías de una razonable aceptación y de manuales con definiciones y procedimientos usados internacionalmente como son el Manual de Frascati, el Manual de Oslo y el Manual de Canberra, publicados por la OCDE y la UNESCO.

Además, tanto Spinak [57] como Rosa Sancho [55] coinciden en mencionar que la medición de la segunda parte (Output) es la tarea más sofisticada y difícil. La evaluación de los resultados científicos no se ha resuelto todavía de forma definitiva, ya que supone medir el conocimiento generado en las tareas de investigación, así como su impacto o influencia en otros investigadores; y tanto el proceso científico como el de adquisición de conocimientos, son muy complejos por su carácter acumulativo y colectivo. Agregan además que el error clásico, consiste en suponer que los resultados de cualquier investigación, deben estar estrechamente relacionados con las inversiones realizadas.

En Mexico, El Consejo Nacional para la Ciencia y la Tecnología (CONACYT) publican anualmente los indicadores de actividades cientificas y tecnologicas referentes a las acciones que se desarollan en el pais de manera sistematica para la producción, diseminación y aplicación de los conocimientos en estas areas [58].

Es importante mencionar que los indicadores bibliométricos tiene cabida en la medicion del Output [52], pues utilizan datos extraídos de las publicaciones científicas asumiendo, que el resultado de la investigación es nuevo conocimiento que se da a conocer a través de publicaciones. Estas

mediciones complementan de manera eficaz las opiniones y los juicios emitidos por los expertos de cada área proporcionando herramientas útiles y objetivas en los procesos de evaluación de los resultados de la actividad científica.

2.5 Enfoque bibliométrico general

El primero [37] y más importante paso en la realización de los Análisis Bibliométricos de alta calidad para el estudio de la ciencia, consiste en la aceptación del *Dogma Central de la Bibliometría*, por parte del usuario potencial, esto es con el fin de que el usuario pueda validar la credibilidad del enfoque bibliométrico. Una vez superado este obstáculo, el segundo paso consiste en seleccionar el conjunto de indicadores más apropiados para alcanzar los objetivos del estudio, y paralelamente, seleccionar los productos (fuentes) con los indicadores primarios de la más alta calidad (datos y bases de datos).

El tercer paso consiste en aplicar análisis de alta precisión estadística a estos indicadores. Y el último paso, el cual (en última instancia) determina la credibilidad y utilidad de los resultados, consiste en la interpretación y presentación visual de los resultados.

Hay que tener en cuenta que los resultados del análisis más riguroso serán relativamente inútiles si no se sitúan en un contexto de evaluación apropiado y si no se muestran de forma concisa.

Capítulo III

Las Bases de Datos Bibliográficas

Por lo tanto, otro elemento importante en metodologías como lo es el ViBlioSOM son las Bases de Datos Científicas - Tecnológicas. Estas bases de datos son la principal fuente de información que se utiliza en los análisis bibliométricos.

En la actualidad existen bases de datos especializadas en todas las áreas científicas, lo que permite analizar cualquier de ellas a través de estas fuentes. Sin embargo, la validez del análisis bibliométrico dependerá en gran medida de que la base de datos seleccionada cubra de forma adecuada el área bajo estudio. Las distintas bases de datos difieren en cobertura temática, criterios de selección de revistas y/o documentos, sesgos geográficos y lingüísticos y todas estas características deben analizarse de forma previa a la realización de un análisis.

De todas las publicaciones científico-tecnológicas, solamente el artículo científico es considerado pieza clave para estudiar a la ciencia por medio de análisis cuantitativos [45] y [46], es decir a través de los análisis bibliométricos. El indicador bibliométrico palabras comunes[56] *(Co-Word Analysis)* elaborado a partir de los artículos seleccionados nos va a permitir detectar *conocimientos* que no existían explícitamente en ningún artículo de la colección, pero que surge de relacionar el contenido de varios de ellos.

Algunas características de este tipo de *conocimiento* son: es valido si y sólo si es entendible; se basa en un razonamiento inductivo; representa casos particulares; etc. Como se aprecia, este *conocimiento* es ligeramente distinto al *conocimiento* obtenido por la ciencia, pues este último es: valido si y solo si es probado; se basa en un razonamiento inductivo; es lo más general posible; es conciso; etc., [59] y [60].

El artículo científico tiene la característica de ser un texto estructurado [61] debido a que marca un orden lógico para la exposición de las ideas, unifica los criterios y facilita la tarea del lector. Además, se divide en varias

[56] Vea sección 2.3.3

partes y cada una de ellas tiene una misión informativa diferente. La estructura más general es la siguiente:

- Titulo.
- Autores.
- Resumen.
- Palabras clave.
- Introducción.
- Materiales y métodos.
- Resultados.
- Discusión de los resultados.
- Agradecimientos.
- Bibliografía.

Según los expertos [62] un artículo científico debe ser la primera divulgación y contener información suficiente para que los colegas del autor puedan:

- Evaluar las observaciones.
- Repetir los experimentos.
- Evaluar los procesos intelectuales.

Además, debe ser susceptible de percepción sensorial, esencialmente permanente, estar a la disposición de la comunidad científica sin restricciones y estar disponible también para su examen periódico por uno o más de los principales servicios secundarios reconocidos, por ejemplo, Biological Abstracts, Chemical Abstracts, Index Medicus, Science Citation Index, etc., en los Estados Unidos y servicios análogos en otros países.

3.1 Las publicaciones científicas

Las publicaciones científicas como son los artículos científicos, las patentes, las compilaciones, las monografías, las actas de congresos, las revisiones, los manuales, los boletines, las normas, las tesis de doctorado, los informes, los libros, los índices, las revistas, las revistas de resúmenes, las revistas anuales, los anuarios, los folletos, la literatura gris, etc., son medios de la *comunicación científica*[57].

[57] i.e. la difusión de los resultados del trabajo científico

Toda esta cantidad de publicaciones se clasifican de formas muy variadas. En este trabajo se emplea la clasificación que se proporciona en el artículo de Leticia Sánchez-Paus [63].

Por su origen:

- *Fuentes Primarias:* son fuentes inéditas, originales, escrita de primera mano por el autor. Tienen la característica de ser *autosuficientes*, es decir, son fuentes originales de la información. Ejemplo, los artículos científicos originales.

- *Fuentes Secundarias:* son documentos que contienen información sobre las fuentes primarias, son obras de referencia que no ofrecen conocimientos nuevos pero facilitan el acceso a las fuentes primarias. Son fuentes guía, no son *autosuficientes*. Ejemplo, las revisiones de la literatura, etc.

Por la información que aportan:

- *De primera mano:* como los artículos científicos originales.

- *De segunda mano:* recoge lo que otros autores han escrito, por ejemplo: un libro de texto.

Por la frecuencia de publicación:

- *No periódicas:* libros: monografías, compilaciones.
- *Periódicas:* revistas científicas y series.

Otros

- *La literatura gris:* trabajos no publicados o de circulación limitada como tesis doctorales, actas de congresos, informes técnicos, notas técnica, etc.

Otro aspecto importante de las publicaciones científicas suele ser su evaluación. Básicamente se evalúan tres aspectos: *la cantidad de información que nos suministra; la calidad y rigor de dicha información;* y por ultimo *la actualidad/accesibilidad.*

La cantidad de información que nos suministra se refiere a que la información debe ser completa y lo suficientemente amplia; *la calidad y rigor*

de dicha información se refiere a que la información debe ser fiable (con poco margen de error). Y la *actualidad/accesibilidad* se refiere a que las fuentes que tengan una regularidad y continuidad en la publicación y que sean de fácil acceso a través de canales comerciales, centros de venta de publicaciones oficiales, e instituciones de reconocido prestigio en el ámbito nacional e internacional.

3.2 Las bases de datos científicas - tecnológicas

Estas bases de datos[58] son la principal fuente de información que se utiliza en los análisis bibliométricos [64]. En ellas se almacenan los resultados de la *comunicación científica* e información que producen distintos organismos e instituciones científicas y tecnológicas. Además, estas bases de datos tienen la característica de desarrollan programas de gestión documental que se encarga de estructurar y controlar la información, para facilitar, en cualquier momento, su rápida y precisa localización y recuperación.

Los párrafos siguientes resumen aspectos importantes de estas bases de datos. Las áreas que cubre las bases de datos, en general, se clasificar en: *ciencias sociales y humanidades; ciencias exactas y naturales; tecnología y ciencias de la ingeniería; y en ciencias biológicas y de la salud.* Es por ello, que antes de realizar cualquier análisis bibliométrico debemos seleccionar la base de datos que cubra de forma adecuada el área objeto de estudio.

Además, se clasifican en dos grandes grupos: las *multidisciplinarias* ponen a disposición de los especialistas, una gran cantidad de información de las más diversas áreas del conocimiento, por ejemplo, el Science Citation Index, SCI, pone a disposición información de áreas como: la agronomía, astronomía, astrofísica, ciencias de la vida, ciencias de la tierra y el espacio, farmacología, física, matemáticas, química y tecnología. Y en *especializadas,* las cuales sólo ponen a disposición del especialista información especializada sobre alguna área del conocimiento.

Para acceder a éstas, los productores o distribuidores ofrecen las siguientes opciones: *En línea,* mediante este servicio, el usuario accede a la información en cualquier momento que lo desee, a través de las redes de Internet, de Intranet o de cualquier otra red de telecomunicaciones disponible. No obstante, se requiere un contrato de acceso, pagar una cuota de conexión y por lo general pagar también por el tiempo de utilización. Las

[58] Una base de datos es una colección de información organizada, de tal manera que un programa de computadora puede rápidamente seleccionar las piezas deseadas de datos.

ventajas de este tipo de acceso son: se dispone de la información más actualizada, y el usuario sólo paga por los servicios que realmente utiliza. La gran desventaja consiste en que si el usuario desea acceder a más servicios, el costo aumenta considerablemente; El *Proveedor o Host* ofrece acceso a una diversidad enorme de bases de datos, por medio de *paquetes* de bases de datos. Además, poseen sus propias herramientas, como son buscadores especializados, tutoriales técnicos o de divulgación, etc.; Y por último, a través de un *Disco Compacto,* cuando no es posible tener uno de los servicios anteriores, se aquieren los discos compactos que el distribuidor pone a la venta. El disco compacto tiene como ventajas su gran capacidad de almacenamiento y el hecho de que puede usarse tanto tiempo como se quiera. El principal inconveniente de este tipo de acceso, es que no se dispone de la información más reciente, ya que la actualización es lenta (puede oscilar desde un mes a un año según las Bases de Datos Bibliográficas).

Los principales problemas que caracterizan a estas bases de datos, se dan en tres grandes rublos que son: *la cobertura, la fiabilidad* y en *la recuperación de la información.* Veamos brevemente cada rublo.

En la cobertura:

- En pocas Bases de Datos, se considera a la llamada *literatura gris*.
- Sesgos geográficos y lingüísticos.
- Corriente Principal de la Ciencia[59].
- Las distintas Bases de Datos difieren en cobertura temática.

En la fiabilidad:

- Inconsistencias en las formas de registrar los nombres de los autores, las instituciones, las revistas, etc.
- Las políticas de publicación.
- Los criterios de selección de revistas y/o documentos.
- Los criterios de indización.

En la recuperación de información:

- Problemas de compatibilidad de la información en la misma Base de Datos, por ejemplo, en línea y en disco compacto.
- Falta de uniformidad cuando se trabajan diferentes Bases de Datos.

[59] Vea el apéndice A

- Las Bases de Datos sólo ofrecen la información que consideran pertinente.

Pero también, es importante considerar que algunos de estos problemas son causados por los propios autores, por ejemplo, hay autores que firma de varias maneras, autores que no proporcionan las referencias en forma apropiada, autores que tienen su trabajo segmentado en varias versiones o inclusive su trabajo está distribuido en varias lenguas.

Un hecho importante que hay que tener en cuenta es el siguiente: *unos cuantos países administran toda la información científica y tecnológica producida por las distintas organizaciones e instituciones científicas y tecnológicas* [25]. Algunos participantes que interviene en tal administración son los productores y los distribuidores de las bases de datos.

Las empresas u organizaciones que crean las bases de datos se denominan *productores de bases de datos*. Su misión consiste en convertir en información elaborada las publicaciones científicas provenientes de las distintas organizaciones e instituciones científicas y tecnológicas [65]. La tabla III.0.1 muestra algunos productores de bases de datos reconocidos mundialmente.

PRODUCTOR	BASE DE DATOS	ÁREAS	CAPACIDAD	COBERTURA
Instituto de la Información Científica y Tecnológica, CNSR, Francia	PASCAL	Ciencias, Tecnología y Medicina.	14.7 millones	1973
Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal	Latindex: Índice latinoamericano de publicaciones científicas.	Están consideradas todas las publicaciones seriadas en las disciplinas de las ciencias exactas, naturales, sociales y humanas	12,000 registros.	1997
Centro de Información y Documentación Científica, CINDOC, del Consejo Superior de Investigaciones Científicas, CSIC, España	ICYT -Ciencia y Tecnología	Astronomía, Astrofísica, Ciencias de la Vida, Ciencias de la Tierra y el Espacio, Farmacología, Física, Matemáticas, Química y Tecnología.	152,000 registros	1979
Biblioteca Nacional de Medicina de E.U.	MedLine	Medicina	12 millones de registros	1966
The Thomson Corporation	Web of Science	Varias disciplinas cientificas	Más de 5.4 millones de enlaces a documentos en texto completo	1960

Tabla III.0.1: Algunos productores de bases de datos.

El contenido de estas bases de datos varia, por ejemplo:

PASCAL: proporciona información sobre más de 6000 titulos de revistas, articulos de revista, procedimientos, disertaciones, libros, patentes, reportes.

Latindex: Brinda información basica sobre más de 12,000 titulos a traves del directorio. Ofrece información adicional sobre un conjunto seleccionado de revistas. Brinda también acceso a recursos electronicos a atraves del indice.

ICYT: proporciona información de 747 publicaciones periódicas editadas en España, fundamentalmente revistas además de monografías, actas de congresos, informes y tesinas.

MedLine: entre sus ventajas como fuente de información está su amplia cobertura de revistas, pues contiene aproximadamente 15 millones de citas bibliográficas que provienen de más de 4,600 revistas que cubren los temas de la biomedicina, principalmente medicina, enfermería, odontología, oncología, medicina veterinaria, salud pública, ciencias preclínicas y de otras áreas de las ciencias de la vida.

Web of Science: proporciona información de más de 8700 revistas a nivel internacional. Además, ofrece el Science Citation Index® (1900 - al presente), Social Sciences Citation Index® (1956 – al presente), Arts & Humanities Citation Index® (1975 – al presente), Index Chemicus® (1993 – al presente), y al Current Chemical Reactions® (1986 – al presente).

Para realizar búsquedas directamente en las bases de datos de los productores, es necesario aprender el *lenguaje de interrogación*[60]. Este lenguaje, por lo general, se basa en la combinación de la lógica booleana, es decir, combinaciones de tipo AND, OR y NOT.

Por ejemplo, para buscar información en MedLine se puede utilizar alguna de las siguientes sintaxis:

1. *search term [tag] BOOLEAN OPERATOR search term [tag]*
2. *asthma/therapy [mh] AND review [pt] AND child, preschool [mh] AND english [la]*
3. *arthritis NOT letter [pt]*

[60] Vea sección 1.2

La primera sintaxis muestra la forma general de combinar términos mediante los operadores booleanos. La segunda sintaxis muestra la forma de buscar una revisión en inglés que trate el tema de las terapias de asma en niños en edad preescolar. Mientras que la tercera sintaxis muestra cómo buscar el tema artritis pero sin buscar este tema en cartas.

Cabe destacar, que la rapidez de desarrollo de los *lenguajes de interrogación* no ha sido comparable a los avances producidos en el hardware o software. Los nuevos lenguajes de interrogación deben basarse en la lógica difusa y en el uso de lenguaje natural, que permitan formular las consultas utilizando las mismas palabras con las que hablamos para que simplifiquen la recuperación de la información.

El detalle de aprender los diversos lenguajes de interrogación, ha favorecido la aparición de los *proveedores* de bases de datos o mejor conocidos como *host*. Los distribuidores permiten el acceso a diversas bases utilizando un lenguaje de interrogación estándar, lo que constituye una gran ventaja para el usuario. Por ejemplo, Thomson Corporation ofrece el acceso a Dialog, Chemical Abstracts, MedLine, Compendex, Inspec, Claims, etc. El Centro de Información y Documentación, CINDOC ofrece el acceso a ICYT e ISOC.

Los distribuidores disponen de potentes computadoras que contienen la información de diversas bases de datos y de programas informáticos que hacen posible las consultas a las diversas bases con un lenguaje estándar. En la actualidad, la mayoría de los productores de bases de datos ceden su explotación a los distribuidores.

Para concluir este resumen, hay que destacar que *los países que crean y administran contenidos científicos–tecnológicos* se destacan por producir una gran cantidad de bases de datos de renombre y con reconocimiento mundial; tienen una amplia red de distribución de la información a través de Internet; desarrollan varios estudios e indicadores sobre varios aspectos relacionados con el análisis de información [51].

Mientras que *los países con bajos niveles en la creación y administración de contenidos científicos–tecnológicos* se destacan por tener oficinas de registro de la propiedad industrial poco organizadas; carecen de leyes adecuadas de protección y los fondos de documentos son deficientes; esfuerzos dispersos en la recopilación y creación de bases de datos; Tiene bases de datos con escasa difusión o de desarrollo permanente; carecen de una infraestructura

estadística de apoyo; falta de recursos y de política científica-tecnológica coordinada.

3.3 Ventajas de estas bases de datos para los análisis bibliométricos

Como se ha mencionado la validez del análisis bibliométrico dependerá en gran medida de que la base de datos bibliográfica seleccionada cubra de forma adecuada el área objeto de estudio. Algunas de las ventajas de estas bases de datos son:

- Debido a la gran capacidad de almacenamiento permiten actuar sobre grandes unidades de datos en cantidad suficiente.

- La estructura y organización de los datos en campos normalizados posibilita la presentación homogénea de las citas bibliográficas.

- El gran número de campos posibles: autores, titulo, editorial, nombre de revista, año de publicación, lugar de trabajo del autor, clasificación, descriptores o resumen, permite una gran variedad de elementos de recuperación.

3.4 MEDLINE®

En esta sección se describe en forma general como buscar y recuperar citas bibliográficas en MedLine.

MedLine [66] es una base de datos bibliográficos[61] producida por la Biblioteca Nacional de Medicina de los Estados Unidos. Entre sus ventajas como fuente de información está su amplia cobertura de revistas, pues contiene aproximadamente 15 millones de citas bibliográficas que provienen de más de 4,600 revistas que cubren los temas de la biomedicina, principalmente medicina, enfermería, odontología, oncología, medicina veterinaria, salud pública, ciencias preclínicas y de otras áreas de las ciencias de la vida.

Otra ventaja consiste en la asignación de palabras clave a documentos que tratan algún tema de biomedicina. A este proceso de asignación se le conoce como *indización* y es simplemente la enumeración sucesiva de los

[61] Las Bases de Datos Bibliográficas: son archivos de información organizada que contienen registros o referencias bibliográficas completas, que suelen ir acompañadas de los resúmenes de los artículos publicados en revistas científicas y que nos permiten obtener el documento completo.

diferentes términos del MeSH Vocabulary[62] que identifican el contenido o los contenidos de cada documento en MEDLINE. La indización es un proceso técnico que requiere de la aplicación de criterios uniformes como son la exhaustividad (multiplicidad), la especificidad, la coherencia, la imparcialidad, la fidelidad y el buen juicio [67].

El MeSH Vocabulary es un *tesauro* de palabras representativas sobre temas de biomedicina. Se integra por más de 33,000 palabras clave (términos), las cuales están clasificados en:

- *Los encabezados MeSH (MeSH Headings)* representan conceptos o temas generales que se encuentran en la literatura biomédica. Algunos ejemplos de encabezados MeSH son los siguientes: body weight, dental cavity preparation, radioactive waste, kidney, self medication, brain edema, etc.

- *Los subencabezados MeSH (MeSH Subheadings)*, son palabras o frases, con las cuales, se califica un *encabezado MeSH*, esto es, estas palabras o frases se usan para caracterizar a los temas generales en sus aspectos más específicos. Algunos ejemplos de subencabezados MeSH son los siguientes: diagnosis, surgery, metabolism, pathology, etc.

- *Los conceptos suplementarios (Supplementary Concepts Records)* son palabras o frases usadas para detallar los efectos farmacológicos de algunos químicos. Por ejemplo, “Aspirin” (Aspirina) posee los siguientes efectos farmacológicos:

    - Anti-inframmatory agents, non steroidal
    - Cyclooxygenase inhibitors
    - Fibrinolytic agents
    - Platelet aggregations Inhibitors

Un aspecto importante del MeSH es su *estructura jerárquica*. En esta estructura en forma de árbol (MeSH Tree Structure), los términos del MeSH se ramifican en series de términos cada vez más concretos o específicos.

La tabla III.0.2 muestra las 15 categorías en las que se organiza el MeSH Vocabulary.

[62] Acrónimo de **Me**dical **S**ubject **H**eadings Vocabulary

CATEGORIAS	
A	Anatomy
B	Organisms
C	Diseases
D	Chemical and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Biological Sciences
H	Physical Sciences
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology and Food and Beverages
K	Humanities
L	Information Science
M	Persons
N	Health Care
Z	Geographic Locations

Tabla III. 0.2: Categorías del MeSH Vocabulary.

Se detallan brevemente algunas categorías.

- La categoría A agrupa términos de anatomía referidos tanto a seres humanos como animales.

- La categoría B se refiere a organismos vivos.

- La categoría C agrupa enfermedades tanto experimentales como clínicas. Los términos relativos a una enfermedad se configuran en el siguiente orden: términos precoordinados como órgano/enfermedad («brain diseases», «skin diseases») o como organismo/enfermedad («salmonella infections», «trypanosomiasis»), órgano+término precoordinado u órgano+enfermedad («illeum, intestinal diseases», «conjunctiva, eye diseases»), síndrome+descriptivo («crying cat syndrome»), síndrome+epónimo («Korsakoff syndrome»), infecciones+términos generales precoordinados («Bordetella infections», «HIV-infections»), cáncer: tumor, cáncer y carcinoma son sinónimos; no se especifican diferencias entre tumores benignos y malignos; los tumores se indexan con términos que indican el tipo histológico («carcinoma, basal cell») y con términos que indican el órgano afectado («skin neoplasms»).

- La categoría D agrupa sustancias químicas, endógenas y exógenas.

- La categoría E agrupa métodos para diagnóstico, terapéutica y equipamiento técnico, entre otros. Las técnicas y métodos se indexan solamente si son la materia principal de un artículo o si son tratados en detalle. Por ejemplo, un artículo que verse sobre el EEG en la epilepsia será indizado como «epilepsy» y «electroencephalography».

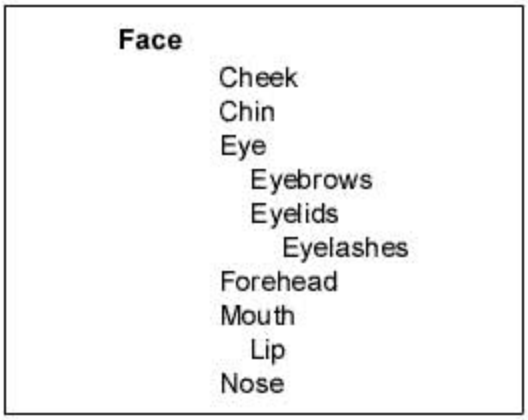


Figura III.0.1: Encabezados y subencabezados.

¿Cómo se relacionados los encabezados y subencabezados MeSH en el MeSH Tree Structure? En la figura III.0.1 se muestra una parte de la jerarquía de la categoría A, la cual incluye a “Face” (Rostro). Por ejemplo, “Eye” *(Ojo)* se considera encabezado mientras que “Eyebrows” *(Ceja)* y “Eyelids” *(Párpado)* son sus correspondientes subencabezados. También, se observa que “Eyelids” *(Párpados)* posee la palabra “Eyelashes” *(Pestaña)* como subencabezado. Note que los subencabezados están debajo de los encabezados.

Ahora se presenta en forma sencilla la manera en que utilizan el MeSH Vocabulary el equipo de indizadores de la Biblioteca Nacional de Medicina de Estados Unidos [68]. Suponga que las dos oraciones siguientes representan los conceptos que se discuten en dos documentos distintos.

- **Transport of aspirin to the brain in relation to the rate of pain relief.**

- **Mathematical models of oxidation reactions of morphine derivatives.**

Entonces, el equipo de indizadores de la Biblioteca Nacional de Medicina de los Estados Unidos asigna los siguientes encabezados y subencabezados a los dos documentos:

- **Transport of aspirin to the brain in relation to the rate of pain relief.**

  ASPIRIN / * pharmacokin / * ther use
  PAIN / * drug ther / * metab
  BRAIN / * metab
  BIOLOGICAL TRANSPORT / physiol

- **Mathematical models of oxidation reactions of morphine derivatives.**

  MORPHINE DERIVATIVES / * chem
  MODELS, CHEMICAL
  OXIDATION-REDUCTION

Se aprecia que los términos en mayúsculas representan a los encabezados MeSH. Mientras que los términos después de la diagonal invertida representan a los subencabezados MeSH. Un encabezado puede tener varios subencabezados. El símbolo asterisco (*) se utiliza para distinguir la importancia de los términos dentro de los documentos, es decir, el termino con asterisco representan al concepto principal que se discute en el documento, de ahí su nombre, *MeSH Major Topic*. La decisión de concederle a un MeSH ser un *"major"*, la toma el grupo de personas encargadas de la indización después de analizar el artículo.

3.5 El Sistema Entrez–Pubmed

Una vez visto los elementos, en forma general, que son utilizados por la Biblioteca Nacional de Medicina durante el proceso de indización. Se destaca otro punto muy interesante ¿Cómo buscar y recuperar citas bibliográficas mediante las *palabras clave*? Para ello, se expondrá en primer lugar el funcionamiento en forma general del Sistema Entrez – Pubmed y en segundo lugar se da un ejemplo de búsqueda y recuperacion de fichas en formato MedLine.

La Biblioteca Nacional de Medicina pone a disposición el sistema Entrez–Pubmed para la búsqueda y recuperación de las citas bibliográficas localizadas en Medline. Este servicio es gratuito y está disponible en la página electrónica de Pubmed. http://www.ncbi.nlm.nih.gov/.

El núcleo del Sistema Entrez–Pubmed es el algoritmo llamado *Mapeo Automático de Términos*[63]. Básicamente, el algoritmo se enfoca en encontrar coincidencias de términos o frases que son ingresados en los cuadros de búsqueda. Los términos o frases pueden ser nombres de temas, nombres de autores, nombres de revistas, instituciones, regiones, edades, términos técnicos, nombres químicos, etc.

El funcionamiento del mapeo automático de términos es el siguiente: los términos o frases son comparados (en este orden) contra lo siguiente:

1. *Tabla de traducción MeSH:* contiene una lista alfabética de las palabras clave del MeSH; los sinónimos, las referencias cruzadas y los términos de entrada para las palabras clave del MeSH; los tipos de publicaciones; términos derivados del *Sistema de Lenguaje Médico Unificado*[64]; Los conceptos de nombres suplementarios correspondientes a los nombres de sustancias y sus sinónimos.

2. *Tabla de traducción de revistas:* contiene un listado alfabético de todos los títulos de revistas; abreviaturas de revistas en formato Medline; el número de identificación unívoco de una revista[65].

3. *Lista de frases:* contiene frases derivadas del *Sistema de Lenguaje Médico Unificado*; nombres de sustancias.

4. *Índice de autores:* contiene una lista alfabética con los nombres de los autores.

Si existe una coincidencia en cualquier etapa, el algoritmo se detiene y muestra los resultados. Si el algoritmo no obtuvo coincidencias en su primer intento de búsqueda, entonces entra en la fase de descomposición del término o frase mediante el operador AND. De nuevo, el procedimiento se repite, pero esta vez, el algoritmo empleará la instrucción *All Fields*[66].

Por ejemplo, suponga que el algoritmo no encontró coincidencias para la frase *HIV Seropositive* en su primer intento. Ahora en su segundo intento descompone dicha frase en:

---

[63] Automatic Term Mapping
[64] Vea apéndice A
[65] International Standard Serial Numbers, ISSN
[66] La instrucción All Fields obliga al algoritmo a buscar coincidencias en los campos que integran el formato MEDLINE.

HIV AND *Seropositive.*

Debido a la opción descomposición se recomienda escribir entre comillas las frases que no se deseen que sean descompuestas, por ejemplo, *"rheumatic diseases".*

La tabla III.0.3 muestra algunos campos del formato MedLine. Los campos (Tags) de un registro bibliográfico, se identifica mediante una etiqueta de dos o más letras (calificadores de campo).

Estas etiquetas permiten realizar búsquedas muy especificas, por ejemplo, si se busca *"mycobacterium"* como un MeSH basta escribir lo siguiente *"mycobacterium bovis [mh]"*. Realizar búsquedas utilizando las etiquetas es muy complicado, es por ello, que solamente personas calificadas realizan este tipo de búsquedas.

- ✓ *(gastro*[JOUR] AND ranitidine[ALL]) AND (100[VOL] OR 150[VOL])*

Devolverá todos los artículos publicados en las revistas cuyo nombre empiece por *"gastro"*, que traten sobre la *"ranitidine"*, y se hayan publicado en los números *"100"* o *"150"* de la revista.

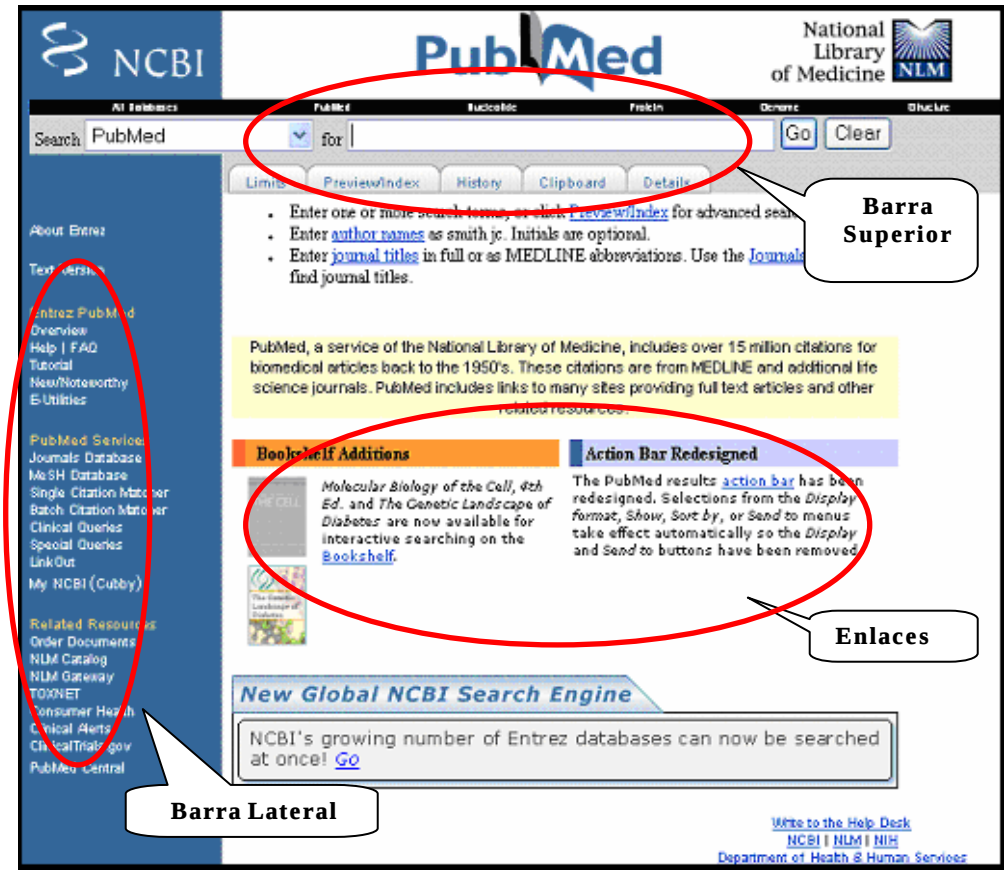
`

CAMPOS DEL FORMATO MEDLINE		
Tag	**Nombre**	**Descripción**
AB	Abstract	Resumen
AD	Affiliation	Filiación Institucional y dirección del primer autor
AU	Autor Name	Nombre de los autores
CY	Country	País de publicación de una revista
DP	Publication Date	Fecha en la que el artículo fue editado
EDAT	Entrez Date	Fecha en la que se incorporó en PubMed.
ID	Identification Number	Número que designa los trabajos financiados por la Agencia Americana del Servicio Público de Salud
IS	ISSN	Número de identificación unívoco de una revista.
JC	Journal Title Code	Código de identificación único compuesto de tres caracteres que adjudica Medline.
JID	NLM Unique ID	Número de identificación de revistas en el catálogo de la Biblioteca Nacional de Medicina.
LA	Language	Idioma del artículo
MH	MeSH Terms	Descriptores o palabras claves
MHDA	MeSH Date	Fecha en la que el término MeSH fue incorporado a la cita
PG	Page Number	Páginas del artículo
PMID	PubMed Unique Identifier	Número de identificación unívoco asignado a cada registro Pubmed
PT	Publication Type	Tipo de artículo
RN	EC/RN Number	Número asignado por la Comisión de Encimas o por el Servicio de Resumen Químicos.
TI	Title Words	Título del artículo
UI	MEDLINE Unique Identifier	Número unívoco asignado a cada registro Medline
VI	Volume	Volumen de la revista

Tabla III.0.3: Algunos campos del formato MEDLINE.

Otro aspecto importante del formato es el siguiente puede ser procesado con cualquier programa de gestión de documentos, por ejemplo el ProCite.

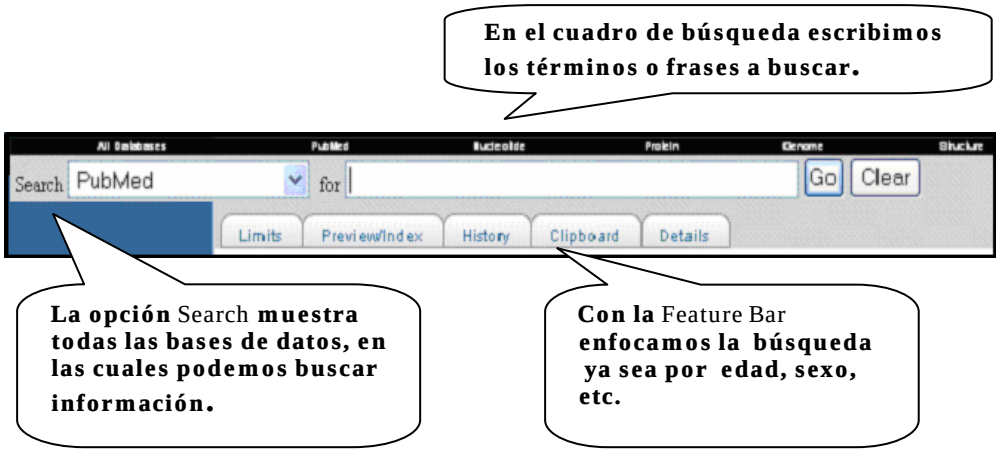
Ahora se presentan algunos elementos que integran la página de Pubmed. En la figura siguiente están señalados tres elementos básicos para iniciar búsquedas y recuperaciones de citas bibliográficas:



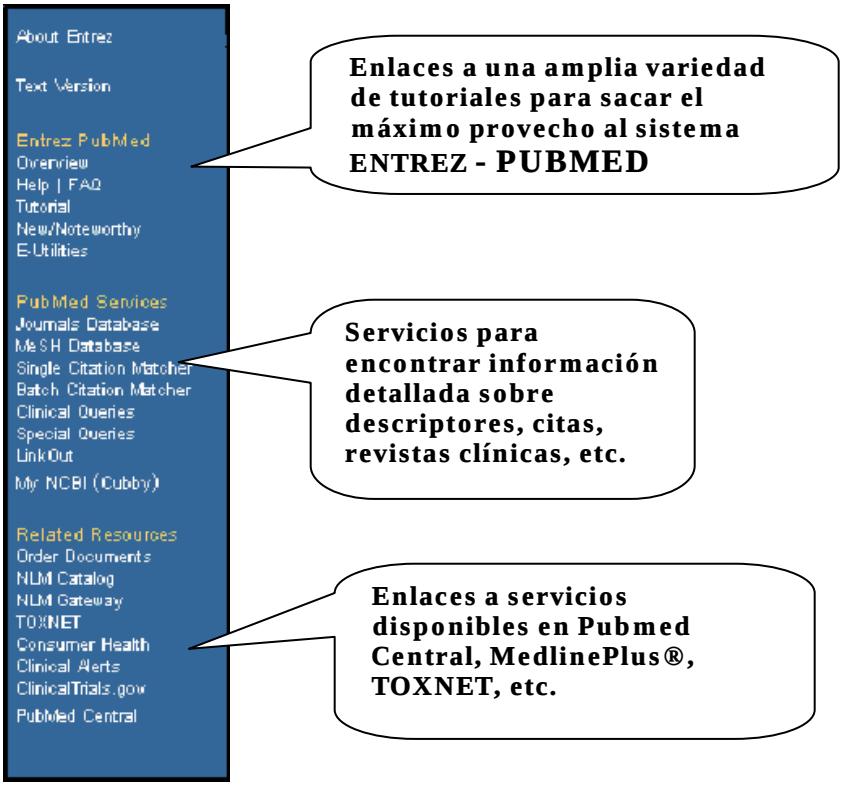
La barra superior se utiliza para buscar términos, frases, autores, revistas, etc. La barra lateral ofrece servicios especializados de búsqueda como son el MeSH Databases, Bath Citation Matcher, Cubby, Fact Sheet Medline, Tutoriales, etc.

Y por último, está la opción de búsqueda a través de enlaces a todas las páginas electrónicas de sitios pertenecientes a la Biblioteca Nacional de Medicina. Algunos sitios son Pubmed Central, MedlinePlus y las bases de datos no bibliográficas de la NCBI, etc.

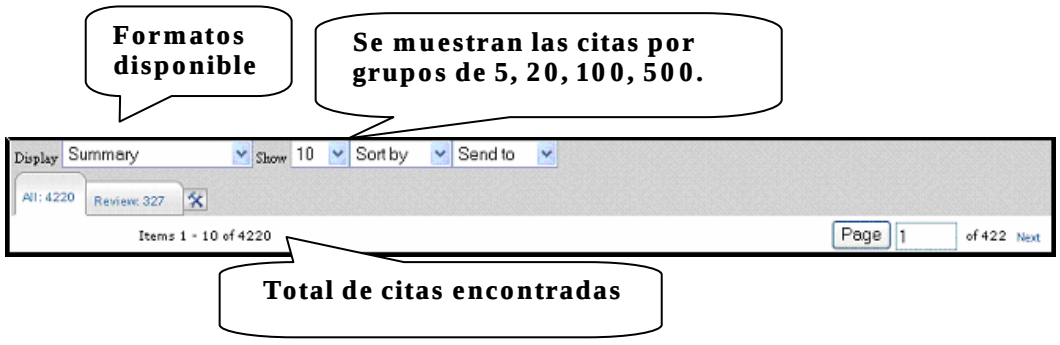
El funcionamiento de estos tres elementos es simple. En la barra superior solamente se escribe el término, la frase, el nombre del autor, etc. Y hacemos clic en *Go* para iniciar la búsqueda. En la figura siguiente están señaladas algunas opciones de la barra superior.



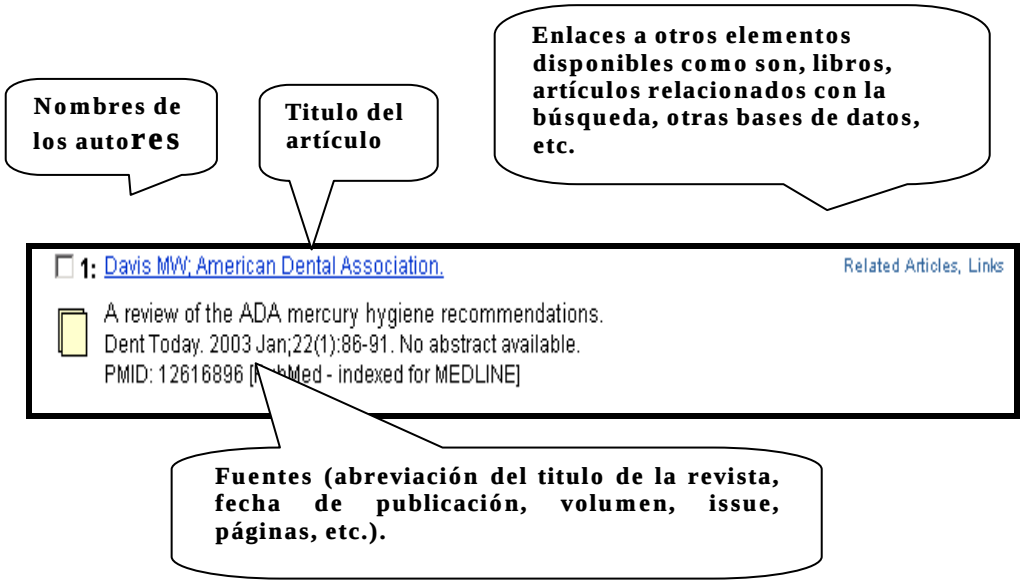
La barra lateral ofrece servicios especializados para todas aquellas personas que los requieran. Los servicios que ofrece la barra lateral se dividen en tres grupos: *Entrez Pubmed, Pubmed Services y Related Resources.* En la figura siguiente están señaladas algunas opciones de la barra lateral.



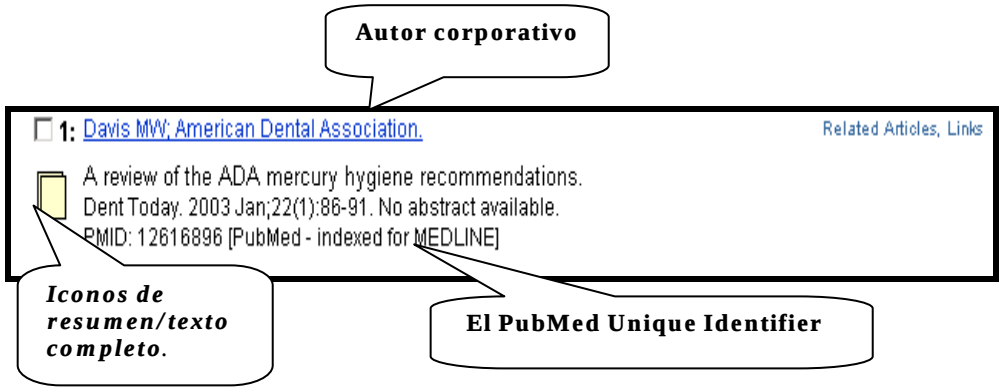
A continuacion se explica como salvar los resultados de una búsqueda. Cuando se realizan búsquedas por medio de cualquiera de los elementos anteriores, los resultados se muestran en una nueva página. Esta página contiene la barra de selección, la cual contiene todas las opciones necesarias para guardar resultados. Algunas opciones están señaladas en las dos figuras siguientes.



Los resultados de la búsqueda se muestran en el *formato Summary*[67]. Éste expone los elementos bibliográficos que integran las citas en forma sencilla que es fácilmente entendible por cualquier usuario. Las dos figuras siguientes muestran los elementos bibliográficos.



[67] Opción de default



De todos estos elementos bibliográficos se destaca solamente dos, *los iconos de resumen/texto completo y el PubMed Unique Identifier (PMID).* El primero indica la presencia o ausencia del resumen, el cual es vital si se desea comparar el contenido según el autor y el contenido según los términos del MeSH Vocabulary. La tabla siguiente muestra los iconos asociados para tal fin.

	Las citas no incluyen resumen.
	Las citas incluyen resumen.
	El texto completo está disponible en PubMed Central (PMC).
	Hay un enlace al texto completo sin ser necesaria una suscripción.

Mientras que el *PubMed Unique Identifier (PMID),* nos indica si la cita del artículo ha sido indizada con los términos del MeSH Vocabulary. La tabla siguiente muestra las etiquetas que acompañan al PMID:

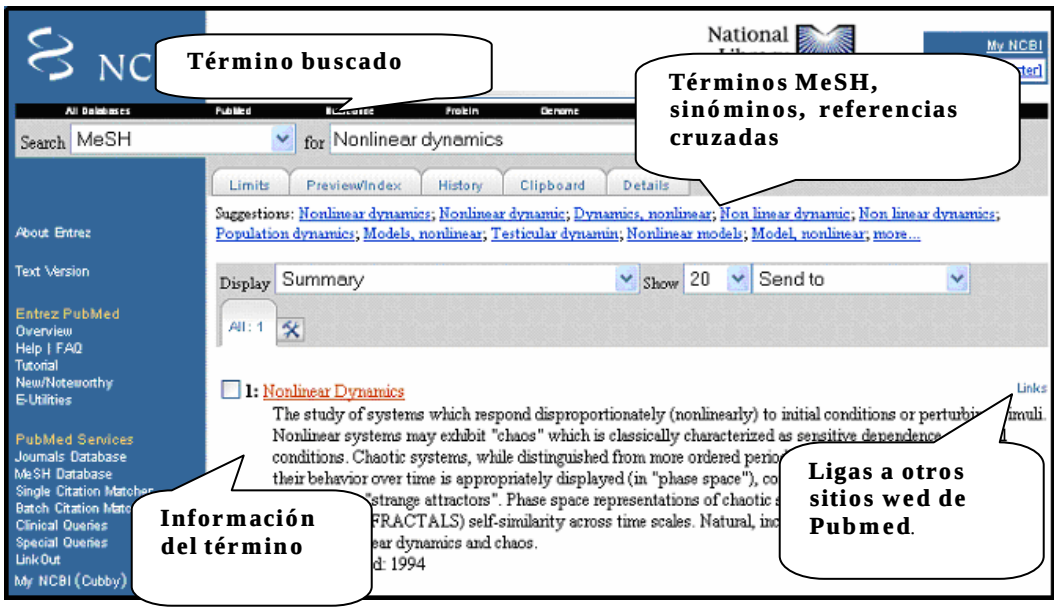
Publisher Supplied Citations	Son citas enviadas a PUBMED, por primera vez para su indización.
In Process	Son citas que están en el proceso de indización.
Indexed for MedLine	Son las citas indizadas exitosamente con el MeSH Vocabulary.
OLDMEDLINE for Pre1966	Son las citas localizadas en OLDMEDLINE provenientes de las décadas 1950s y 1960s. No están indizadas con el MeSH Vocabulary.
PubMed	Son las citas que no pudieron ser indizadas con el MeSH Vocabulary.

Ejemplo de Búsqueda: La barra lateral ofrece el servicio *MeSH Database*. Este servicio se utiliza para buscar documentos indizados con el MeSH Vocabulary. Antes de iniciar cualquier búsqueda se recomienda diseñar una *estrategia de búsqueda*.

Para diseñar la estrategia se debe tener en mente lo siguiente. En primer lugar, se debe limitar nuestra búsqueda, ya sea por tipo de documento (Clinical Trials, Editorial, Review, etc.[68]), el tipo de estudio (Humano vs. Animal, Masculino vs. Femenino, etc.), la edad de la población en estudio, etc. En segundo lugar, se debe seleccionar las palabras más representativas del tema de investigación.

Inicio: seleccione el MeSH Database en la barra lateral. Y en el cuadro de diálogo escriba ***Nonlinear Dynamics*** (Dinámica No Lineal), y finalmente se hace clic en *Go*. En la imagen siguiente se ven los resultados obtenidos.

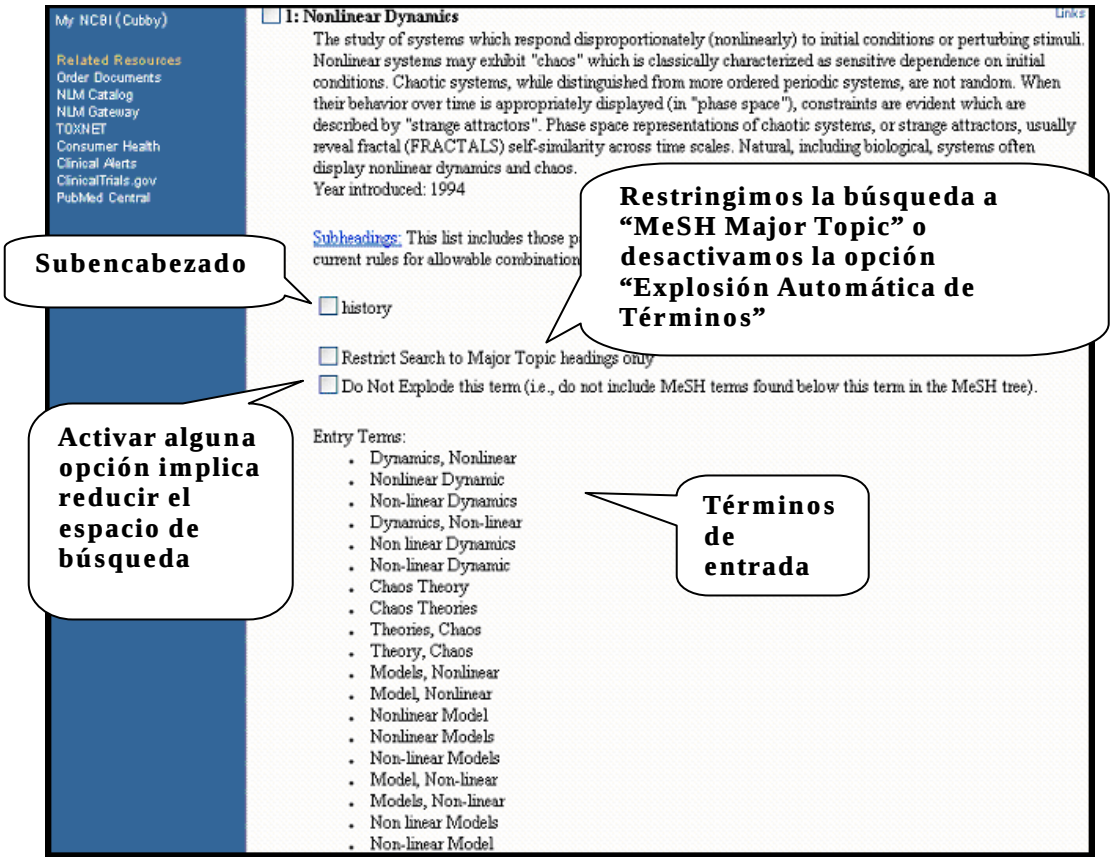
[68] Vea apéndice A



La frase que se ingreso resulto ser MeSH, si no lo hubiese sido, el sistema habría dicho que no encontró nada o simplemente muestra términos MeSH que se relacionan con el término introducido. Ahora se hace clic en ***Nonlinear Dynamics*** para ver sus *subencabezados*, *términos de entrada*, *sinónimos, etc.*, y además, *su posición en el árbol del MeSH*.

En la figura siguiente se muestra la nota de información y las sugerencias del término ***Nonlinear Dynamics*****.** La nota de información es:

Nonlinear Dynamics: The study of systems which respond disproportionately (nonlinearly) to initial conditions or perturbing stimuli. Nonlinear systems may exhibit "chaos" which is classically characterized as sensitive dependence on initial conditions. Chaotic systems, while distinguished from more ordered periodic systems, are not random. When their behavior over time is appropriately displayed (in "phase space"), constraints are evident which are described by "strange attractors". Phase space representations of chaotic systems, or strange attractors, usually reveal fractal (FRACTALS) self-similarity across time scales. Natural, including biological, systems often display nonlinear dynamics and chaos.



Hay que destacar que este término solamente posee el subencabezado *"history"*. Las siguientes citas tienen a *"history"* como subencabezado de Nonlinear Dynamics.

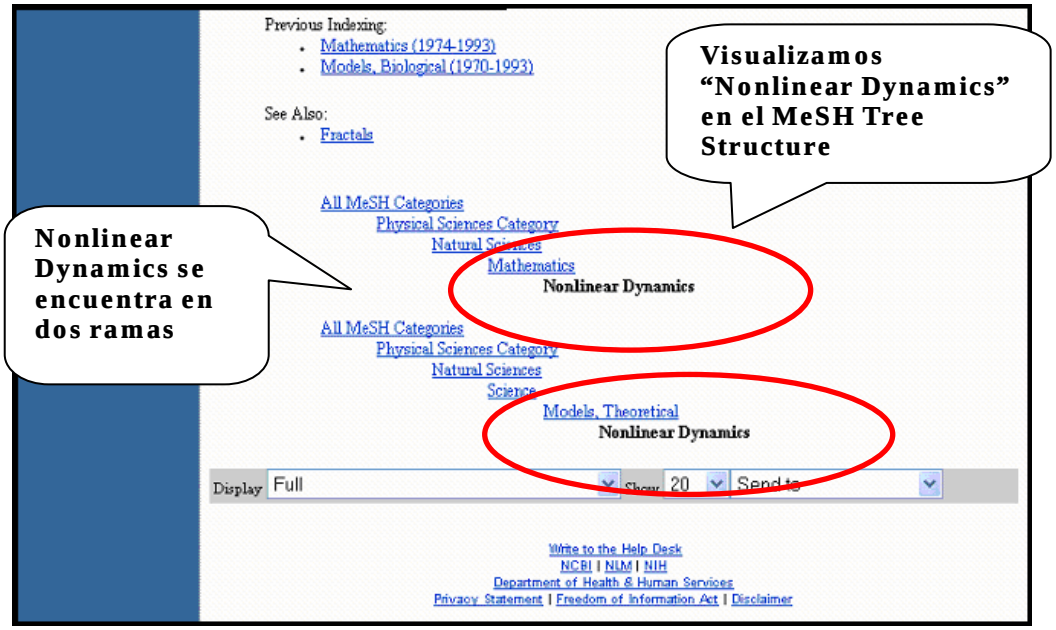
- Zador AM. "The basic unit of computation". Nat Neurosci. 2000 Nov;3 Suppl:1167.
- Barciszewski J. "Pioneers in molecular biology: Emil Fischer, Erwin Schrodinger and Oswald T. Avery". Postepy Biochem. 1995; 41(1): 4-6.
- Bogartz RS. "The future of dynamic systems models in developmental psychology in the light of the past". J Exp Child Psychol. 1994 Oct; 58(2): 289-319.

Posteriormente, se aprecian las dos opciones para limitar la búsqueda. Estas opciones son:

- ✓ *Restrict Search to Major Topic Headings Only*
- ✓ *Do Not Explode This Term.*

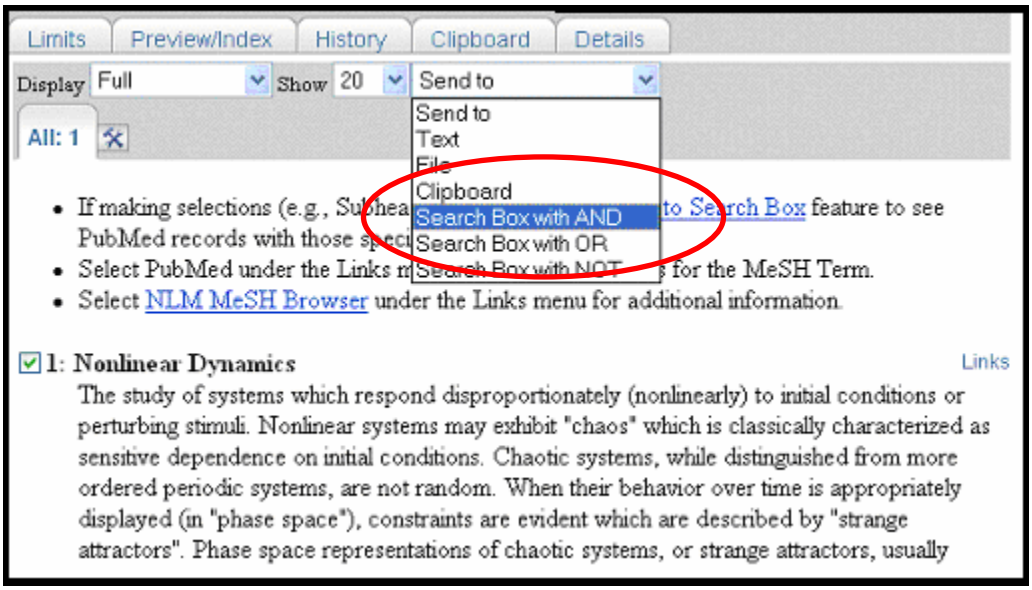
Como se desea recuperar todas las citas que traten algún tema relacionado con ***Nonlinear Dynamics*** no se activará ninguna opción. Y por ultimo, los términos de entrada (Entry Terms) representan alias del término ***Nonlinear Dynamics***.

Por otra parte, en indización previa (Previous Indexing) se aprecia que anteriormente ***Nonlinear Dynamics*** se localizaba en las ramas del MeSH Tree Structure correspondientes a *mathematics* (Matemáticas; 1974-1993) y a *Models, Biological* (Modelos, Biológicos; 1970-1993).



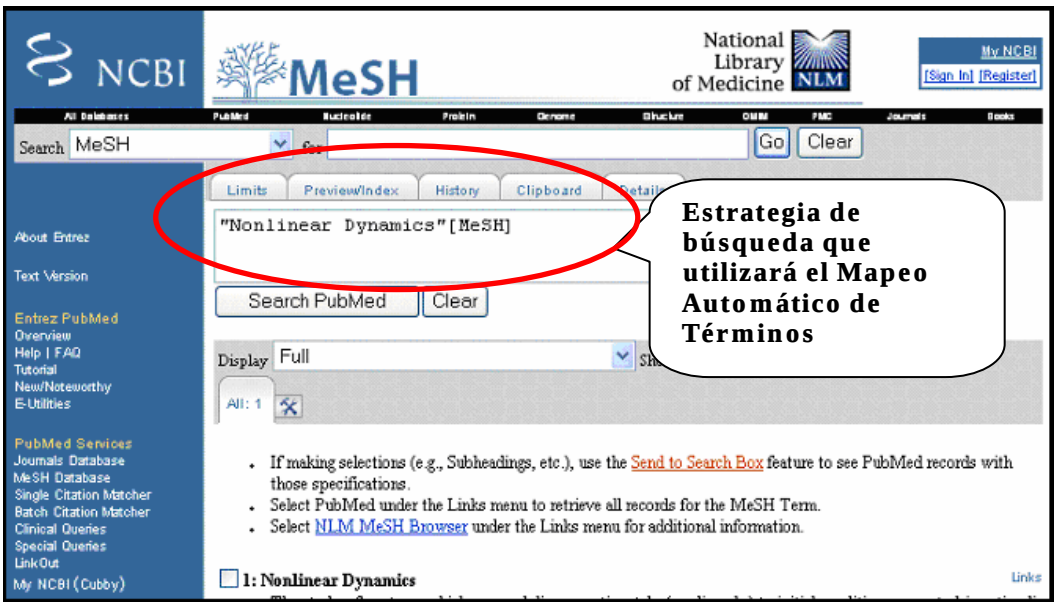
En la actualidad ***Nonlinear Dynamics*** se localiza en las ramas del MeSH Tree Structure correspondientes a *mathematics* (Matemáticas) y a *Models, theoretical* (Modelos Teóricos).

Como siguiente paso se selecciona la casilla de chequeo y en el *combobox* se selecciona a su vez, enviar a la caja de búsqueda con un AND (Search Box with AND). Esto permitirá agregar a la petición que busque aquellas citas que tienen como MeSH ***Nonlinear Dynamics***. En este caso, es la primera sentencia de búsqueda que se agrego. La figura muestra lo antes mencionado.



Se ha mencionado en la sección 3.2 que es muy importante entender como funcionan los operadores boolenos AND, OR, NOT. Pues estos permiten hacer búsquedas muy especificas.

Como resultado se obtiene la creación automática de la sentencia de búsqueda, la cual se agrega a la caja de búsqueda. En la siguiente figura se ve la sintaxis de búsqueda **“Nonlinear Dynamics” [MeSH]** que empleará el mapeo automático de términos.



Y finalmente, se hace clic en el botón con etiqueta *Search PubMed* para iniciar la búsqueda. En la siguiente figura se muestran los resultados en el formato predeterminado. En resumen, se recuperarón 4,220 citas que tratan algún tema relacionado con ***Nonlinear Dynamics.*** Y estas citas abarcan el periodo de 1970 al 2004.

Y por último, se deben guardar estas citas en la computadora. Para ello, en primer lugar, se selecciona el *formato MedLine* disponible al lado derecho de *Display*. En segundo lugar, se selecciona *File* y se hace clic en *Send to*. Se abrirá un cuadro de dialogo en donde se debe dar un nombre y una ubicación al archivo.

Con este ejemplo se concluye este capítulo. La bases de datos bibliográfica MedLine es idónea para realizar análisis bibliométricos de la literatura biomédica, pues ofrece una enorme cantidad de documentos. El formato MedLine ofrece una gran cantidad de campos en donde se puede recuperar los nombres de los autores, las instituciones, las palabras clave, los años de publicación, los resúmenes, etc.

Capítulo IV

Las Redes Neuronales y los Mapas Auto - Organizantes

El algoritmo SOM[69] se utiliza en el ViBlioSOM debido a su eficiencia para llevar a cabo las siguientes tareas:

- *Conglomerados (Clusters):* afortunadamente el SOM se ubica dentro del contexto de las redes neuronales artificiales de aprendizaje no supervisado. Esto lo hace idóneo para detectar conglomerados en conjuntos de datos multidimensionales.

- *Visualización:* el SOM hace una proyección no lineal de los conglomerados detectados a un mapa bidimensional. Esta proyección posee la característica de preservar la topología de los conglomerados en forma de vecindades en el mapa. Esta característica permite al analista detectar las relaciones intrínsecas de los datos.

Estas tareas son vitales en las etapas de minería de datos y en la visualización e interpretación de los resultados[70]. A continuación se exponen brevemente algunos conceptos generales relativos a la naturaleza y utilidad de las redes neuronales artificiales. Posteriormente se da una revisión general sobre la importancia del algoritmo SOM en la visualización de grande cantidades de datos multidimensionales.

4.1 Elementos de las redes neuronales

Después de la segunda mitad del siglo XX, los investigadores en las áreas de inteligencia artificial, aprendizaje máquina, etc., han tratado de imitar, por medio de simulaciones por computadora el funcionamiento del cerebro humano.

[69] Self-Organizing Maps
[70] Ver sección 1.5.3

Hasta la fecha, la mejor simulación del cerebro humano son las *Redes Neuronales Artificiales*[71]. Estas redes son modelos computacionales, los cuales, tienen al cerebro humano como modelo ideal, es decir, toman las características esenciales de la estructura neuronal del cerebro para crear sistemas que lo mimeticen, en parte, mediante sistemas computacionales (resumen de [69]). En la figura siguiente se muestran algunas partes que integran a una neurona biológica.

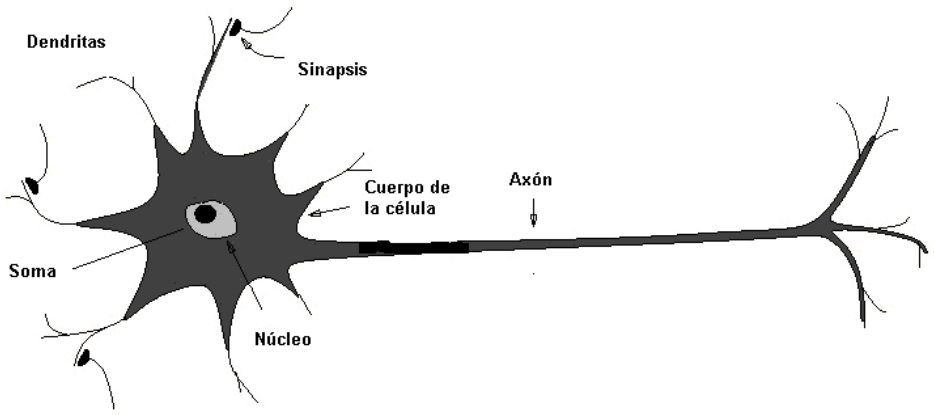


Figura IV.0.1: Esquema de una neurona biológica.

En el cerebro, la unidad básica es la neurona, célula que se caracteriza por su capacidad de conexión con otras neuronas a través de sinapsis, gracias a sus miles de receptores *(dendritas)* y salida *(axón)*. Con estos elementos de conexión que permiten la interrelación entre neuronas del cerebro, se forman redes neuronales. En el cerebro humano, las neuronas trabajan en equipo, es decir, grupos de neuronas se enfocan en procesar la información proveniente de la vista o el tacto, mientras que otros grupos se enfocan en el pensamiento abstracto, en la percepción estética, etc.

La figura siguiente muestra las partes que integran a una red neuronal artificial. Por su parte, en la neurona artificial, la unidad básica recibe el nombre de *nodo.* Al igual que su contra parte biológica, se puede conectar a otros nodos para formar redes.

El funcionamiento de cada nodo depende de dos partes básicas: una suma ponderada de las entradas x_i y una función de activación f.

[71] Ver apéndice A

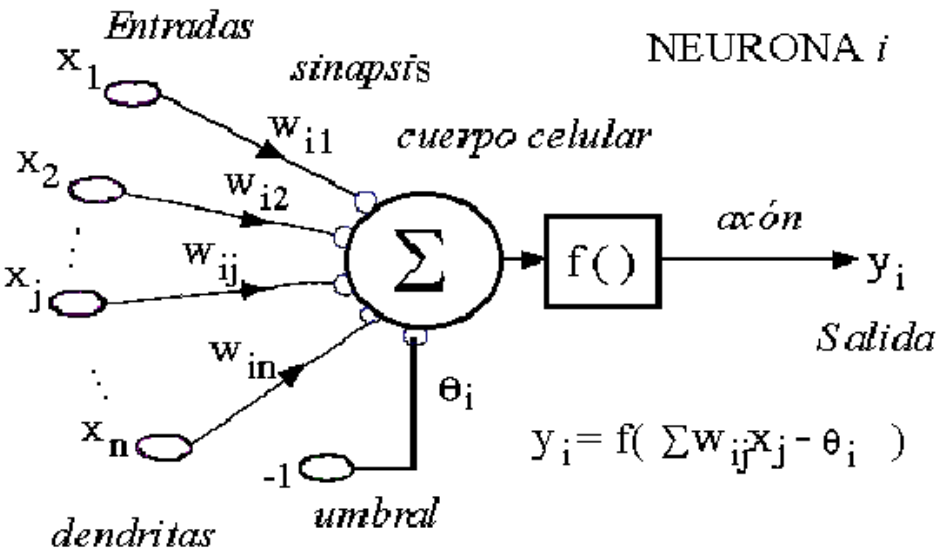


Figura IV.0.2: Modelado una red neuronal artificial.

4.2 Arquitecturas de las redes neuronales

Desde un punto de vista matemático, las conexiones entre los nodos pueden ser representadas por medio de gráficas dirigidas, es decir, por un conjunto de vértices unidos por segmentos dirigidos. *La arquitectura de la red* es simplemente el flujo que se le asigna a estas conexiones, es decir, el patrón de conectividad entre los nodos. Por medio del patrón de conectividad de la gráfica, se pueden definir dos categorías básicas de arquitecturas en las redes neuronales: las Redes Hacia - Adelante[72] y las Redes de Retroalimentación[73].

[72] Feedforward Networks
[73] Feedback Networks

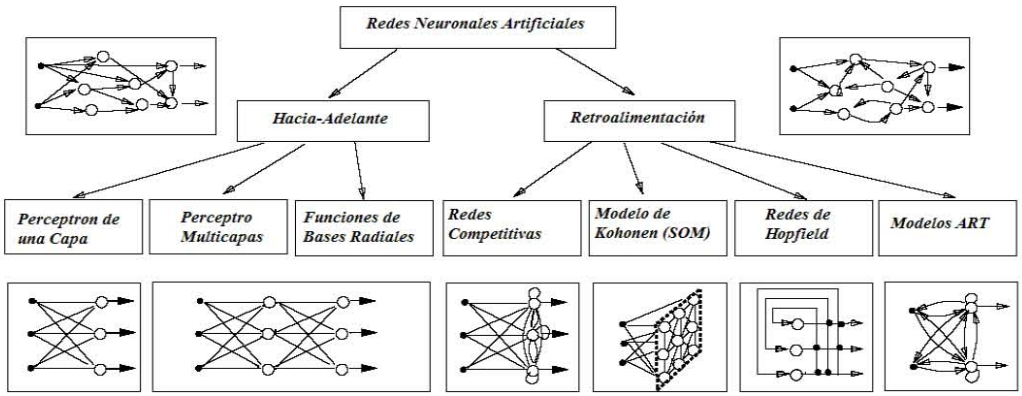


Figura IV.0.3: Las arquitecturas más representativas de cada categoría.

Como se aprecia en la figura IV.0.3, la diferencia básica entre las categorías consiste en el uso de ciclos en el patrón de conectividad. En la primer categoría no existen ciclos, mientras que en la segunda si. La existencia de estos ciclos permite a la red neuronal retroalimentarse o no durante el entrenamiento. Hay que tener en cuenta que las distintas conectividades dan comportamientos distintos al interior-exterior de las redes neuronales. En general, las redes hacia adelante son redes estáticas, es decir, dado un dato de entrada, estas producen un sólo conjunto de valores de salida y no una secuencia de éstos. Además, estas redes no tienen memoria ya que la respuesta de una de estas redes a un dato de entrada dado, es independiente de los estados previos de la red. Por el contrario, las redes de retroalimentación se consideran sistemas dinámicos, debido a que cada vez, que se presenta un dato de entrada las respuestas de las neuronas son computadas, por medio de las conexiones de retroalimentación, de manera que los vectores de pesos de las neuronas son modificados. Lo anterior hace que la red se modifique hasta que alcance algún tipo de equilibrio o convergencia.

4.3 El proceso de aprendizaje

A groso modo, las redes neuronales artificiales [70] simulan al proceso de aprendizaje del cerebro humano, por medio de pesos sinápticos w. Veamos brevemente en que consiste la simulación. Supongamos que tenemos una neurona i como la de la figura IV.0.4, el peso sináptico w_{ij} representa la intensidad de interacción o probabilidad de que la neurona i sea activada por la neurona j. La neurona i tiene un número n de sinapsis o entradas provenientes de otras neuronas x_n, cada una con su peso sináptico

w. *El potencial postsináptico* de la neurona i vendrá dado por la suma ponderada de los productos de las entradas x_j por su peso sináptico w_{ij}. La salida y_i de la neurona i es una función de esa suma ponderada, según la fórmula que aparece en la figura IV.0.4, siendo θ_i el umbral.

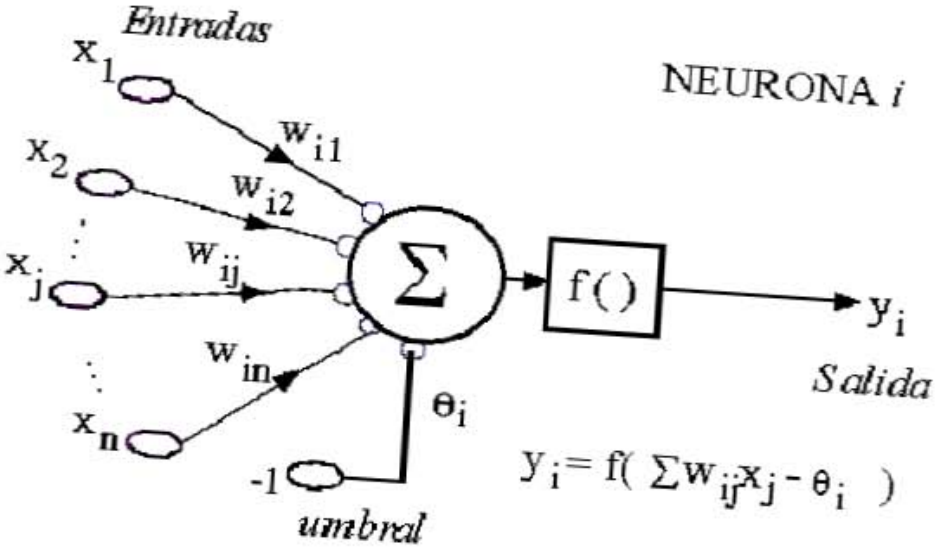


Figura IV.0.4: La neurona artificial.

Esta simulación se sustenta en las ideas de Hebb [65]. Según Hebb la variación de la capacidad, probabilidad, frecuencia o tendencia que una neurona tiene para activar a otra es también la base de todo aprendizaje, el cual, consiste en asociar y disociar estímulos y respuestas. Para Hebb, cuando se estimulan conjuntamente dos neuronas aumenta la probabilidad de que vuelvan a hacerlo simultáneamente en una ocasión subsecuente, es decir, refuerzan sus sinapsis, y todo este proceso tiene como consecuencia que el sistema sináptico de la red se modifique en forma continua.

Por otro parte, el proceso de aprendizaje en una red neuronal artificial se considera desde el punto de vista matemático como un problema de actualización iterativa de los pesos sinápticos durante el proceso de entrenamiento. Básicamente, el problema se ha tratado de resolver simulando dos paradigmas de aprendizaje que los humanos utilizamos frecuentemente. *El paradigma del aprendizaje supervisado*, se asemeja al método de enseñanza tradicional con un profesor que indica y corrige los errores del alumno hasta que éste aprende la lección. Mientras que en *el*

paradigma del aprendizaje no supervisado, no hay un profesor que corrija los errores al alumno; recuerda más al autoaprendizaje. El alumno dispone del material de estudio pero nadie lo controla.

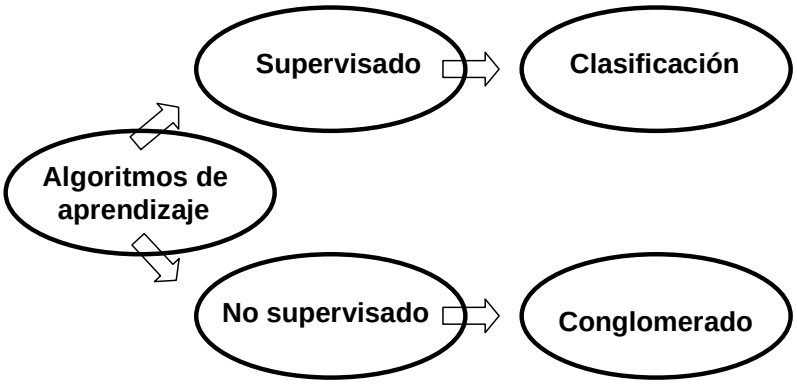


Figura IV.0.5: Paradigmas de Aprendizaje.

Desde el punto de vista matemático, lo anterior significa que durante el aprendizaje supervisado debemos proporcionarle a la red neuronal parejas de datos del tipo *"entrada–salida"*, a partir de los cuales se realizaran las actualizaciones. Mientras que en el aprendizaje no supervisado, únicamente debemos suministrar datos de entrada a la red neuronal, a partir de los cuales se realizaran las actualizaciones.

4.4 El éxito de las redes neuronales

El éxito de las redes neuronales artificiales en campos como la economía, las finanzas, el reconocimiento de patrones, etc., radica principalmente en su robustez más que en su desarrollo teórico.

A continuación se mencionan dos problemas para los cuales, las redes neuronales han sido empleadas exitosamente. Posteriormente menciono algunas ventajas de las redes neuronales sobre otros modelos computacionales.

Clasificación. La clasificación consiste en distribuir el conjunto de datos, en un número de categorías posible dado a priori por un experto. El paradigma de aprendizaje supervisado se ajusta muy bien a la clasificación, ya que para llevarla a cabo es necesario especificar las características de las distintas categorías y el número de las mismas, además de tener que proporcionarle un conjunto preparado de datos. Usualmente estos datos pertenecen a las distintas categorías; así el sistema aprende a que categorías pertenecen cada tipo de datos y generaliza para clasificar nuevos conjuntos de datos. Una vez aprendido los clasificadores crean una estructura propia o reglas en base a los casos que le han sido presentados y los aplican a los nuevos casos.

Conglomerado[74]. Un conglomerado o conglomeración consiste en la partición del conjunto de datos en conjuntos ajenos que se forman de acuerdo a una métrica previamente establecida. El conglomerado permite la identificación de topologías o grupos, en los cuales, los elementos de un mismo grupo guardan similitud entre sí y se diferencian de los elementos de otros grupos. El paradigma de aprendizaje no supervisado, se ajusta muy bien a la partición de los datos en conglomerados, pues en este caso no se le proporciona ninguna información al sistema, el sistema aprende por si mismo. No se parte de un conjunto prefijado de categorías sino que a través del análisis de los datos mismos y de su naturaleza, esta técnica agrupa dichos datos en las distintas categorías.

La diferencia sustancial entre las dos técnicas es que, en la clasificación se conoce el número de categorías y la naturaleza de los datos que la forman, mientras que en el conglomerado no se conoce a priori la naturaleza de las categorías ni los atributos de datos que influirán en la formación de grupos.

Veamos algunas ventajas de las redes neuronales sobre otros modelos computacionales.

Aprendizaje Adaptable. Esta característica es una de las propiedades más atractivas de las redes neuronales; las neuronas artificiales aprenden a llevar a cabo ciertas tareas mediante un entrenamiento con ejemplos ilustrativos. Como las redes neuronales pueden aprender a diferenciar patrones mediante ejemplos y entrenamiento, no es necesario que elaboremos modelos a priori ni necesitamos especificar funciones de distribución de probabilidad.

[74] Cluster

Tolerancia a fallos. Las redes neuronales son los primeros métodos computacionales con la capacidad inherente de tolerancia a fallos. Comparados con los sistemas computacionales tradicionales, los cuales pierden su funcionalidad en cuanto sufren un pequeño error de memoria, en las redes neuronales, si se produce un fallo en un pequeño número de neuronas, aunque el comportamiento del sistema se ve afectado, no sufre una caída repentina. Hay dos aspectos distintos respecto a la tolerancia a fallos: primero, las redes neuronales pueden aprender a reconocer los patrones con "ruido", distorsionados o incompletos, esta es una tolerancia a fallos respecto a los datos. Segundo, pueden seguir realizando su función (aunque con cierta degradación) si se destruye parte de la red. La razón por la que las redes neuronales son tolerantes a fallos es que tienen la información distribuida en las diversas conexiones entre neuronas y existe cierto grado de redundancia en esta forma de almacenamiento. La mayoría de las computadoras algorítmicas y sistemas de recuperación de datos, almacenan cada pieza de información en un espacio único, localizable y direccionable. Las redes neuronales artificiales, a semejanza de las biológicas, almacenan información no localizada. Por tanto, la mayoría de las interconexiones entre los nodos de la red tendrán unos valores en función de los estímulos recibidos, y se generará un patrón de salida que represente la información almacenada.

Operación en tiempo real. Una de las prioridades de las áreas de aplicación, es la necesidad de procesar grandes cantidades de datos de forma muy rápida. Las redes neuronales se adaptan bien a esta situación, debido a su implementación paralela. Para que la mayoría de las redes neuronales puedan operar en el momento en el que se requiere, la necesidad de cambio de los pesos de las conexiones o entrenamiento es mínima. Por tanto, las redes neuronales son una excelente alternativa para el reconocimiento y clasificación de patrones en tiempo real.

Fácil inserción dentro de la tecnología existente. Debido a que una red neuronal puede ser rápidamente entrenada, comprobada, verificada y trasladada a una implementación hardware de bajo costo, es fácil insertar redes neuronales para aplicaciones específicas dentro de sistemas existentes.

4.5 Las redes de Kohonen

Los Mapas Auto Organizados (*Self-Organizing Maps, SOM),* fueron presentados en 1982 por Teuvo Kohonen desde entonces se han producido miles de artículos de investigación y ha sido aplicado en una amplia variedad

de campos de investigación (Resumen [69]). La principal razón de la popularidad del SOM es su capacidad de presentar de manera automática un mapa en el cual se puede observar una descripción intuitiva de la similitud entre los datos; el despliegue bidimensional tiene la propiedad de presentar la información contenida en los datos de manera ordenada y resaltando las relaciones mencionadas. A continuación se exponen algunos conceptos generales relativos a la naturaleza y utilidad del algoritmo SOM.

4.5.1 La estructura del SOM

La estructura interna del SOM consiste de dos capas de neuronas, una capa de entrada y otra capa de procesamiento. En la figura IV.0.6 se aprecia que las capas están conectadas entre si por medio de los *pesos sinápticos*.

La idea básica del modelo es, a partir de de un espacio multidimensional de entrada, crear una imagen en un espacio de salida de menor dimensión. Las neuronas de la primera capa se limitan a recoger y a canalizar los datos de entrada. La segunda capa está conectada a la primera a través de los pesos sinápticos y realiza la tarea importante: una proyección no lineal del espacio multidimensional de entrada, en un espacio de menor dimensión, preservando las características esenciales de estos datos, en forma de relaciones de vecindad. El resultado final es la creación del llamado *mapa autoorganizado* donde se representan los rasgos más sobresalientes del espacio de entrada.

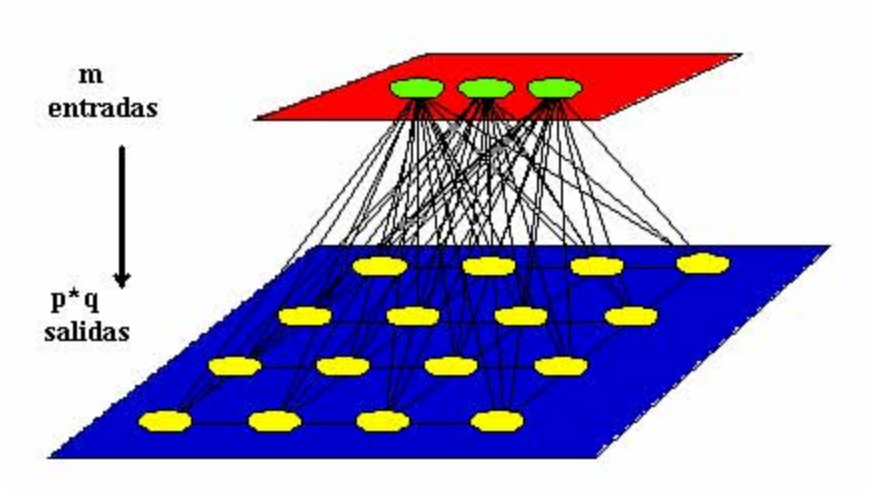


Figura IV.0.6: La estructura del SOM.

Ahora veamos brevemente el funcionamiento del algoritmo SOM. Para una revisión detallada del algoritmo vea [69], [71], [72] y [73].

El punto de partida del SOM es un conjunto $\Im = (\eta_1,\ldots,\eta_N)$ de neuronas todas ellas con las mismas propiedades: se conectan de manera idéntica a la entrada $x \in \mathbb{R}^n$ --normalmente se considera que $U \subset \mathbb{R}^n$ -- e interactúan entre ellas por medio de relaciones laterales que se activan durante la actualización de los pesos.

Estas relaciones responden a la relación (figura IV.0.7) de distancia física entre una neurona y sus vecinas.

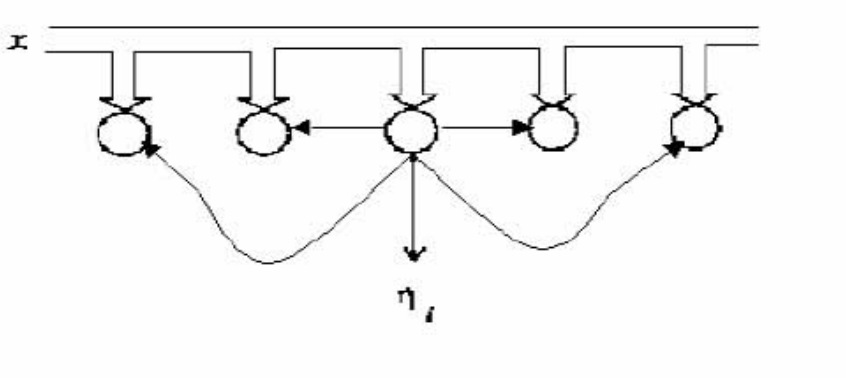


Figura IV.0.7: Representación de una neurona y sus conexiones con la entrada x y las neuronas vecinas.

Durante el proceso de entrenamiento competitivo, la entrada x se considera como una variable en función de t (donde t es la coordenada de tiempo discreto) que toma valores del conjunto de datos de entrada X, por tal motivo es necesario indexar a los elementos del conjunto X de la siguiente manera:

$$X = \{x(t): \ t = 1,\ldots,m\}$$

Cuando el valor de t sobre pasa al número m, el conjunto X es reciclado y sus elementos son reindexados manteniendo el orden de la primera presentación.

Normalmente, *la arquitectura de la red* tiene la siguiente característica:

- Las neuronas se distribuyen a lo largo de una retícula bidimensional.
- Cada neurona constituye a un nodo de la retícula.

- La configuración o tipo de retícula puede ser definida como rectangular, hexagonal o incluso irregular.
- La localización de la neurona sobre la retícula está representada por su vector de localización $r_i = (p_i, q_i) \in \mathbb{N}^2$
- Cada neurona es asociada a un vector de pesos $w_i \in \mathbb{R}^n$. En el caso del SOM este vector es también llamado *vector de referencia*.

En la figura IV.0.8 se muestran las configuraciones o tipo de retícula más usados con los correspondientes $r_i = (p_i, q_i)$ en cada nodo. Cabe señalar que la configuración hexagonal es más conveniente para efectos de visualización.

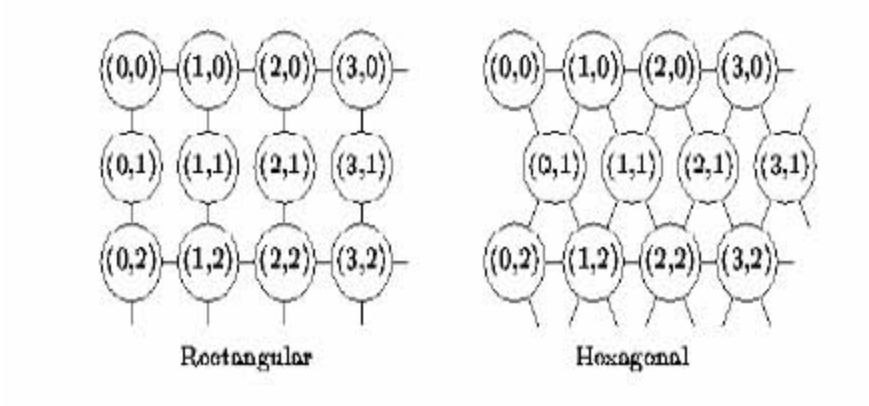


Figura IV.0.8: Configuraciones más comunes en la retícula del SOM.

En el algoritmo SOM básico, las relaciones topológicas entre los nodos (hexagonal o rectangular) y el número K de neuronas son fijados desde el principio. Normalmente se definen las distancias entre las unidades del mapa de acuerdo a la distancia euclidiana entre los vectores de localización, sin embargo, en ocasiones es más práctico usar otras funciones de distancia.

Ahora se detalla en que consiste el entrenamiento. El entrenamiento se lleva a cabo mediante un proceso de aprendizaje competitivo, en el cual las neuronas se vuelven gradualmente sensibles a diferentes categorías de los datos de entrada. En cada momento t del proceso de entrenamiento, un vector de entrada $x(t) \in \mathbb{R}^n$ es conectado a todas las neuronas en paralelo vía los vectores de referencia w_i de cada neurona.

Las neuronas compiten para ver cual de ellas es capaz de representar de mejor manera al dato de entrada *x(t)*. Dado cualquier $x \in X$ la competencia consiste en encontrar la neurona tal que su vector de referencia w_c cumpla con:

$$\|x - w_c\| = Min_{i=1}^{N} \{\|x - w_i\|\} \qquad [[1]]$$

A la neurona ganadora η_c se le define como el nodo que mejor representa al dato x. Nótese que el subíndice c es función de x; para cada x existe un $w_{c(x)}$. En caso de que este índice no esté bien definido, es decir cuando para un dato x existan dos w_e, $w_d \in \Im$, tal que:

$$d(x, \eta_c) = Min \{ d(x - \eta_i) : i = 1,..., k \} = d(x, \eta_d)$$

La selección de un único *c(x)* debe hacerse de manera aleatoria. Por simplicidad se adoptará la siguiente notación:

$$x : \eta \Leftrightarrow \eta = \eta_{c(x)}$$

Generalmente, se utiliza la distancia euclidiana para determinar el nodo que mejor representa a un dato en [[1]]. La distancia euclidiana se define para elementos del espacio $\Re^n$ como sigue:

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

También, se tiene la opción de usar otras normas si el problema lo requiere. Por ejemplo, las llamadas *normas* - L_r que son funciones de la forma:

$$d(x, y) = \|x - y\|_r = \left(\sum_{i=1}^{n} | x_i - y_i |^r \right)^{\frac{1}{r}}$$

Donde r es un número real positivo. Para el caso $r = 1$ la métrica es llamada métrica Manhatan.

Para variables que son medidas en unidades que no son comparables en términos de escala, la siguiente función de distancia es especialmente apropiada:

$$d(x, y) = \left((x - y)^t B^{-1} (x - y) \right)^{\frac{1}{2}}$$

Donde B es una matriz de $n \times n$ invertible definida positiva, esta función es conocida como la distancia de Mahalanobis. Nótese que la distancia de Mahalanobis generaliza a la norma Euclidiana ya que esta última se obtiene cuando $B = I$ donde I es la matriz identidad en $\mathbb{R}^n$.

Para cada tiempo t se realiza la competencia [[1]] de manera que se puede definir $c = c(t)$ tal que $x(t) \sim \eta_{c(t)}$, aquellas neuronas que se encuentran dentro de una vecindad de $\eta_{c(t)}$ en el arreglo bidimensional (ver figura 3) aprenderán de la misma entrada $x(t)$. La vecindad de $\eta_{c(t)}$ sobre la retícula se define a partir del vector de localización $r_{c(t)}$ de la siguiente manera:

$$N_{c(t)} = \{ i \in \mathbb{N} \ : \ \left\| r_{c(t)} - r_i \right\| \ \leq \ \rho(t) \} \qquad [[2]]$$

Donde $\rho(t)$ es el radio de la vecindad en el tiempo t. Como se observa en [[2]], el radio de la vecindad varía en función de t. Para efectos de la convergencia del algoritmo, la variación del radio a través del tiempo debe cumplir las siguientes condiciones:

a) Si $t_i \leq t_j \ \Rightarrow \ \rho(t_i) \geq \rho(t_j)$

b) Si $\rho(t) \rightarrow 0$ cuando $t \rightarrow \infty$

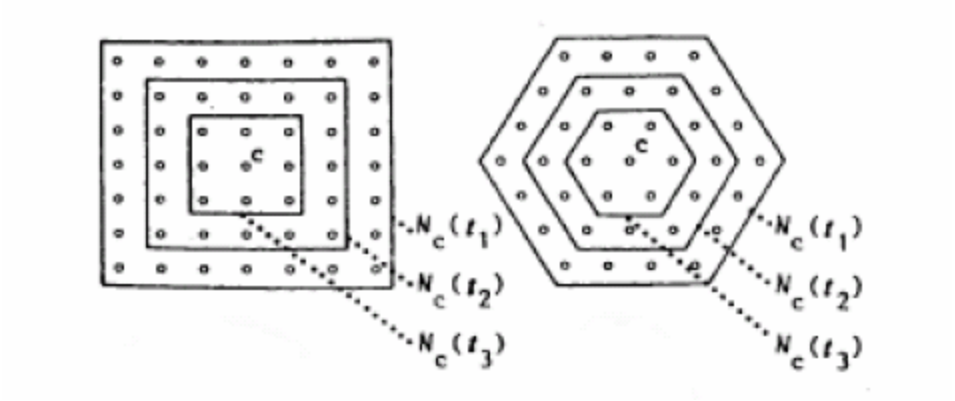


Figura IV.0.9: Variación en el tiempo del radio de la vecindad.

Debe tenerse cuidado al escoger el tamaño inicial de $\rho(0)$, si desde un comienzo la vecindad es muy pequeña, el mapa no se ordenará globalmente, lo cual implicará que el mapa generado se verá como un mosaico de parcelas entre las cuales el ordenamiento cambia discontinuamente. Para evitar este fenómeno $\rho(0)$ puede comenzar siendo más grande que la mitad del diámetro de la red. Para iniciar el proceso de aprendizaje se utilizan valores aleatorios para los vectores de referencia $w_i(0)$. En las versiones más simples del SOM los valores sucesivos para los vectores de referencia se determinan recursivamente por el siguiente mapeo de iteraciones:

$$w_i(t+1) = w_i(t) + h_{ci}(t) \cdot [x(t) - w_i(t)]$$

La función $h_{ci}(t)$ desempeña un papel fundamental en este proceso. A esta función se le conoce como función vecindad. En la literatura es común encontrar que esta función tenga la forma:

$$h_{ci}(t) = h(\|r_{c(t)} - r_i\|, t) \qquad [[3]]$$

Lo cual implica que el valor de la función depende de la distancia entre la neurona η_i y la neurona ganadora $\eta_{c(t)}$ en el tiempo t. El ancho promedio $\rho(t)$ y forma de $h_{ci}(t)$ definen la rigidez del mapa que será asociada a los datos. Independientemente del cual sea la forma explícita de la función [[3]], debe ser tal que $h_{ci}(t) \rightarrow 0$ mientras $||r_{c(t)} - r_i||$ se incrementa. Una de las definiciones más simples que se encuentran de la función vecindad es la siguiente:

$$h_{ci}(t) = \begin{cases} \alpha(t) & si \quad i \in N_c(t) \\ 0 & si \quad i \notin N_c(t) \end{cases} \qquad [[4]]$$

El valor de $\alpha(t)$ se define como factor de aprendizaje el cual cumple con la condición $0 < \alpha(t) < 1$ y usualmente $\alpha(t)$ es una función monótona decreciente. Otra forma común de la función vecindad está dada en términos de la función Gaussiana:

$$h_{ci}(t) = \alpha(t) \cdot exp\left(\frac{\|r_c - r_i\|^2}{2 \cdot \sigma^2(t)}\right) \qquad [[5]]$$

Donde $\alpha(t)$ es el factor de aprendizaje y el parámetro $\sigma(t)$ corresponde al ancho promedio de $N_c(t)$, en este caso $\rho(t) = \sigma(t)$**.** Tanto $\alpha(t)$ como $\sigma(t)$ son

funciones escalares decrecientes con respecto al tiempo. La definición de estas funciones debe tener como consecuencia del cumplimiento de básicamente dos etapas del proceso durante el proceso de entrenamiento: ordenamiento global y refinamiento.

El entrenamiento consiste de las siguientes dos etapas (en general): *Una etapa de ordenamiento global* y *una etapa de refinamiento.*

Ordenamiento Global: Según lo reportado por Teuvo Kohonen durante aproximadamente las primeras 1000 competencias se lleva acabo el ordenamiento de los datos a lo largo y ancho del mapa. Este ordenamiento consiste en establecer los pesos de cada neurona para que estas sean capaces de identificar cierto subconjunto característico dentro del conjunto de datos X y para que las relaciones de cercanía entre las distintas neuronas del mapa reflejen cercanía de los datos correspondientes en el espacio multidimensional del cual provienen. Si los valores iniciales de los pesos han sido seleccionados de manera aleatoria, durante estos primeros 1000 pasos los valores de $\alpha(t)$ deben comenzar siendo razonablemente grandes (cercanos a la unidad) e ir descendiendo hasta llegar a valores cercanos a 0.2.

En general, la forma de $\alpha(t)$ no es importante, puede ser lineal, exponencial o inversamente proporcional a t. Es importante señalar que la selección óptima de estas funciones y sus parámetros sólo pueden ser determinadas experimentalmente; ya que no existe algún resultado analítico que garantice dicha selección óptima.

Refinamiento: Después de la fase de ordenamiento los valores de $\alpha(t)$ deben ser pequeños y decrecer lineal o exponencialmente durante la fase fina. Dado que el aprendizaje es un proceso estocástico, la precisión final del mapa dependerá del número de pasos en esta etapa final de la convergencia, la cual debe ser razonablemente larga. El número de pasos debe ser del orden de 100000, sin embargo en ciertas aplicaciones, como el reconocimiento de voz, es de alrededor de 10000. Por otro lado, cabe señalar que la cardinalidad del conjunto X no es relevante para determinar este número de pasos. Nótese que el algoritmo es computacionalmente ligero y que el conjunto X puede ser reciclado para lograr tantos pasos como sea necesario.

Una vez concluido el proceso de entrenamiento, el SOM define una regresión no-lineal que proyecta un conjunto de datos de dimensión alta en un conjunto vectores de referencia, por lo que dicho conjunto sirve para

obtener una representación del conjunto de datos en una red adaptable *("elástica")* de dos dimensiones en la cual se pueden observar las relaciones de similitud y la distribución de los datos. De esta manera es posible construir una representación bidimensional de un conjunto de datos multidimensional.

4.5.2 Visualización de Información

Una problemática frecuente en el análisis de datos es que por un lado se cuenta con grandes cantidades de datos multidimensionales y por otro lado no se cuenta con información acerca de las relaciones y las estructuras subyacentes del conjunto de los datos; mucho menos se cuenta con una función de distribución o modelo matemático que describa estas estructuras; lo único con lo que se cuenta es con un gran volumen de datos multidimensionales y con una forma de medir la similitud entre ellos. (Resumen de [69]).

Una alternativa para la solución a esta problemática es la utilización de *Redes Neuronales de Aprendizaje No Supervisado*. Estas redes neuronales son capaces de encontrar y descubrir, de manera automática, patrones de similitud dentro del conjunto de datos de entrenamiento y agrupar a los elementos de este conjunto en conglomeraciones, de manera que datos similares se agrupen dentro del mismo conglomerado. Estos descubrimientos pueden realizarse sin ningún tipo de retroalimentación con el medio externo y sin la utilización de información a priori.

Dentro del contexto de los procesos cognitivos del cerebro, la forma de situar al aprendizaje no supervisado es considerándolo semejante a los procesos inconscientes, en los cuales ciertas neuronas del cerebro aprenden a responder a un conjunto específico y recurrente de estímulos provenientes del medio externo, de esta manera se construyen los llamados *mapas sensoriales* en el cerebro.

Desde hace tiempo es sabido que varias áreas del cerebro, especialmente la corteza cerebral, están organizadas de acuerdo a distintas modalidades sensitivas: hay áreas que se especializan en algunas tareas específicas (Figura IV.0.10), ejemplos de estas tareas son: control del habla y análisis de señales sensoriales (visual, auditivo, somatosensorial, etc.).

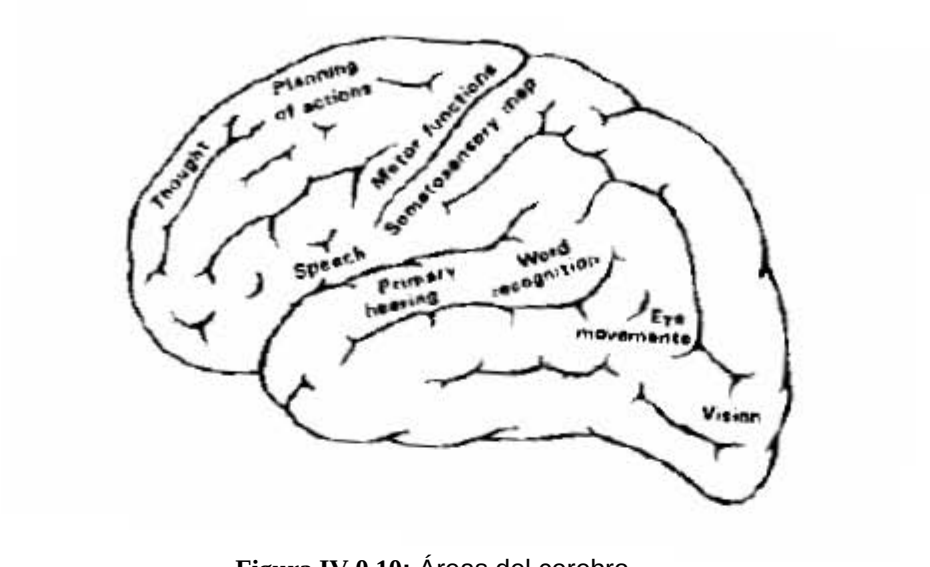


Figura IV.0.10: Áreas del cerebro.

Distintas regiones de un *mapa sensorial* aprenden a reconocer estímulos específicos del medio ambiente. Como consecuencia, la información que cada cúmulo de neuronas reconoce se ubica dentro de cierta categoría dentro de los estímulos que se reciben del exterior.

La manifestación más clara del sentido fisiológico en el aprendizaje no supervisado de las redes neuronales artificiales es que *"el aprendizaje puede suceder únicamente cuando hay redundancia en la presentación de los datos"*. En la práctica esta redundancia se obtiene mediante la utilización iterada (reciclaje) de un mismo conjunto de datos, a lo largo de todo el proceso de entrenamiento. *"Sin una retroalimentación con el exterior sólo la redundancia puede proveer de información útil acerca de las propiedades del espacio de entrada"*.

En resumen, el mecanismo de *auto-organización* que se propone en el SOM consiste de una red neuronal que usa su capacidad de aprendizaje para representar la estructura geométrica (orden topológico) subyacente en el conjunto de datos de entrenamiento, la representación es posible gracias a la auto-organización topográfica de las neuronas de acuerdo a las relaciones de similitud entre los datos representadas por la cercanía entre las neuronas y los vectores de referencia correspondientes. En este sentido, el SOM constituye un mecanismo que brinda la posibilidad de producir automáticamente una representación del conjunto de datos en una estructura bidimensional.

De manera que en dicha representación se haga evidente la emergencia de propiedades que ayuden a entender el orden geométrico subyacente en el conjunto de datos.

La emergencia de comportamientos complejos en un sistema de elementos que interactúan es uno de los fenómenos más fascinantes observados en la naturaleza. Ejemplos pueden ser observados en casi cualquier campo de interés científico, desde la formación de patrones en los sistemas físicos y químicos, el movimiento de enjambres de animales en biología hasta el comportamiento de grupos sociales. Estas investigaciones conducen a la hipótesis de que existe una forma común por medio de la cual describir la formación de estas estructuras.

A pesar de que no existe una definición de *auto-organización* comúnmente aceptada, en este trabajo entenderemos que: *"La auto-organización es el proceso por medio del cual en un sistema de unidades individuales, por medio de interacciones cooperativas, emergen nuevas propiedades en el sistema que trascienden a las propiedades de sus partes constitutivas".*

El cerebro no escapa a la presencia de autoorganización, ésta puede observarse durante el establecimiento de las conexiones entre las neuronas en el cerebro y el sistema nervioso. Una idea fundamental que se deriva a partir de la presencia de auto-organización en el cerebro es que la información no está concentrada en una simple neurona, reside distribuida en distintas áreas, la memoria de un hecho corresponderá a la activación de una familia específica de neuronas. Por lo anterior, algunos investigadores manejan la hipótesis de que el conocimiento es representado por el cerebro a partir de la emergencia de organización en las conexiones neuronales.

En el caso de las redes neuronales artificiales de entrenamiento competitivo (como el SOM) la ausencia de información previa hace necesario contar con algún mecanismo de auto-organización. Este mecanismo debe estar basado en algún criterio de similitud para que así, la organización de los datos corresponda a grupos de datos semejantes entre sí. De esta manera la evolución de la red neuronal, durante el proceso de entrenamiento, estará dirigida a hacer emerger una representación de las relaciones derivadas a partir de la similitud entre los datos.

Partiendo del marco conceptual de los procesos cognitivos del cerebro y las nociones expuestas previamente se pueden deducir algunas consecuencias que corresponden a propiedades de los *mapas sensoriales* y en

la forma en que la información es organizada. Entre otras cosas se puede concluir que:

- Es posible representar la organización de la información a través de las relaciones entre las neuronas de una red.

- Como se dispone de un número finito de neuronas, la representación de la información debe corresponder al orden natural --estructura subyacente en el conjunto de datos-- y hacerse de manera que se utilice eficientemente el número de neuronas.

- La pérdida de neuronas no implica que se pierda la representación de la información.

Dadas estas propiedades es pertinente plantearse la posibilidad de construir formas visuales que representen la organización natural de la información. *"Los mapas de conocimiento son representaciones gráficas de las conexiones hechas por el cerebro en el proceso de entendimiento de los hechos".* Dichos mapas constituyen un medio visual en el cual ideas complejas puedan ser expuestas de manera rápida y en un orden lógico. La representación proporcionada por los mapas resulta de gran utilidad en el descubrimiento de características presentes en el conjunto de datos, de las que no se tenía conocimiento previo.

En una gran cantidad de aplicaciones los mapas topográficos que se producen a partir del SOM resultan ser poderosas herramientas de análisis; el algoritmo SOM tiene la capacidad de producir medios visuales que representen las relaciones y estructuras de similitud entre los datos. En consecuencia, el despliegue visual de las relaciones de similitud provee al analista de una visión que es imposible obtener al leer tablas de resultados o simples sumarios de estadísticas.

Por lo tanto, los mapas generados a partir del SOM resultan ser útiles para el descubrimiento de información previamente desconocida y relevante en la comprensión del fenómeno correspondiente al conjunto de datos. En este sentido, *"el SOM representa una herramienta que puede ser utilizada para la generación automática de mapas del conocimiento".* Esta utilidad es aprovechada por el ViBlioSOM.

La virtud del algoritmo SOM es la regresión no-lineal del conjunto ordenado de vectores de referencia dentro del espacio de entrada. Los

vectores de referencia forman una red elástica de dos dimensiones que sigue a la distribución de los datos.

A continuación se hacen algunas especificaciones de las propiedades del SOM que lo destacan como una herramienta útil y eficiente en el análisis de grandes conjuntos de datos multidimensionales.

Visualización del ordenamiento del conjunto de datos: El ordenamiento producido por la regresión permite el uso de los mapas como un despliegue de los datos. Cuando los datos son mapeados a aquellas unidades en el mapa que tienen los vectores de referencia más cercanos, las neuronas vecinas serán similares a los datos mapeados dentro de ellas. Este despliegue ordenado de los datos facilitará la comprensión de las estructuras subyacentes en el conjunto de datos. El mapa puede ser usado como un campo de trabajo ordenado en el cual los datos originales pueden ser dispuestos en su orden natural. Estas disposiciones han sido discutidas en las variables se aplanan localmente en el mapa, lo cual ayuda a penetrar en las distribuciones de los valores del conjunto de datos. Este mapa es mucho más ilustrativo que tablas de columnas con estadísticas linealmente organizadas. Estas características de los mapas generados por el SOM, permiten que el SOM sea útil para la generación de mapas de conocimiento los cuales son de gran utilidad en los análisis bibliométricos.

Visualización de cúmulos: El mapa generado para el análisis del conjunto de datos puede ser usado para ilustrar la densidad de las acumulaciones en diferentes regiones en el espacio U en las cuales es posible observar relaciones de similitud. La densidad de los datos del conjunto de entrada X es representada por su acumulación en los vectores de referencia. En las áreas de acumulación los vectores de referencia serán cercanos y el espacio vació entre ellos se hará cada vez más escaso. Por lo tanto, la estructura del cúmulo en el conjunto de datos puede vislumbrarse por la disposición de las distancias entre los vectores de referencia de las unidades vecinas. El diagrama de acumulación resultante es muy general en el sentido de que no se necesita asumir nada acerca del tipo de cúmulo. Sin embargo, para lograr definir los cúmulos es necesaria la aplicación de algún algoritmo de conglomerados sobre los vectores de referencia. Algunos métodos de conglomerados utilizados son el *SOM-Ward Clusters, SOM-Single-Linkage Clusters*[75]*, etc.*

[75] Ver manual ViBlioSOM

Datos faltantes: Algunos métodos estadísticos (métodos de conglomerados y de proyección) tienen problemas si alguno de las componentes de los datos no está disponible o no es definible. En el caso del SOM el problema de datos faltantes puede ser tratado como sigue: cuando se escoge la unidad ganadora por [[1]] el vector de entrada x puede ser comparado con los vectores de referencia w_i usando sólo aquellos componentes que están disponibles en x. Nótese que en los vectores de referencia no hay datos ausentes, de tal forma que si únicamente una pequeña porción de las componentes está ausente, el resultado de la comparación será estadísticamente completo. Cuando los vectores de referencia son adaptados sólo las componentes que están disponibles en x serán modificadas. Se ha demostrado que se obtienen mejores resultados si se aplica el método antes descrito que si se opta por descartar los datos con componentes faltantes. Sin embargo, para datos en los cuales la mayoría de las componentes faltan, no es razonable asumir que la selección del ganador es adecuada.

Otra alternativa consiste en descartar durante el proceso de aprendizaje los datos cuyas componentes ausentes exceden una porción determinada Sin embargo, las muestras descartadas pueden ser dispuestas en el mapa después de que ha sido organizado.

Datos extremos: En la medición de los datos pueden existir datos extremos, que son datos ubicados muy lejos del cuerpo principal del conjunto de datos. Los datos extremos pueden resultar a partir de la medición de los errores o registrando los errores hechos mientras se insertan las estadísticas dentro de la base de datos. En estos casos es deseable que datos no afecten el resultado del análisis. En el caso en el que el mapa producido por el algoritmo SOM: cada dato extremo afecta únicamente una unidad del mapa y su vecindad, mientras que el resto del mapa puede ser usado para inspeccionar el resto de los datos. Más aún, los datos extremos pueden ser fácilmente detectados basándose en la distribución del conjunto de entrada X dentro del mapa. Si es deseado, los datos extremos pueden ser descartados y el análisis puede continuar con el resto del conjunto de datos.

Capítulo V

Análisis Bibliométrico

En este capítulo se utiliza la metodología ViBlioSOM para extraer *conocimiento* de los documentos indizados por MedLine. El indicador bibliométrico Palabras Comunes *(Co-Word Analysis)* elaborado a partir de los documentos seleccionados, nos va a permitir detectar *conocimiento* que no existía explícitamente en ningún documento de la colección, pero que surge de relacionar el contenido de varios de ellos [59] y [60]. Este *conocimiento* proviene de mapas que describen las relaciones más significativas de las *palabras clave*, en un conjunto de documentos sobre biomedicina que se ha seleccionado. Estos mapas [55] son idóneos para detectar tendencias de cambio científico en instituciones o en investigadores; el ciclo temporal en que ciertos temas permanecen vigentes, etc.

Antes de emplear la metodología ViBlioSOM se hace una breve revisión de la Categoría de Ciencias Fisicas y la subcategoría de Mathematics. Esta última contiene los términos que hacen referencia a temas matemáticos.

5.1 La Biblioteca Nacional de Medicina y el proceso de indización

Desde el año de 1950 hasta el primer semestre del año 2005, la Biblioteca Nacional de Medicina de Estados Unidos ha indizado alrededor de 15, 301, 768 documentos relacionados con temas de Biomedicina (medicina, enfermería, odontología, oncología, medicina veterinaria, salud pública, ciencias preclínicas y de otras áreas de las ciencias de la vida). Todos los documentos que indiza la Biblioteca Nacional de Medicina son de los siguientes tipos: *Crinical Trial, Editorial, Letter, Meta-Analysis, Practice Guideline, Randomized Controlled Trial, Review*[76].

En la figura V.0.1 se muestra el número de documentos indizados por año en los últimos 55 años en MedLine.

[76] Vea el apéndice A para una breve definición

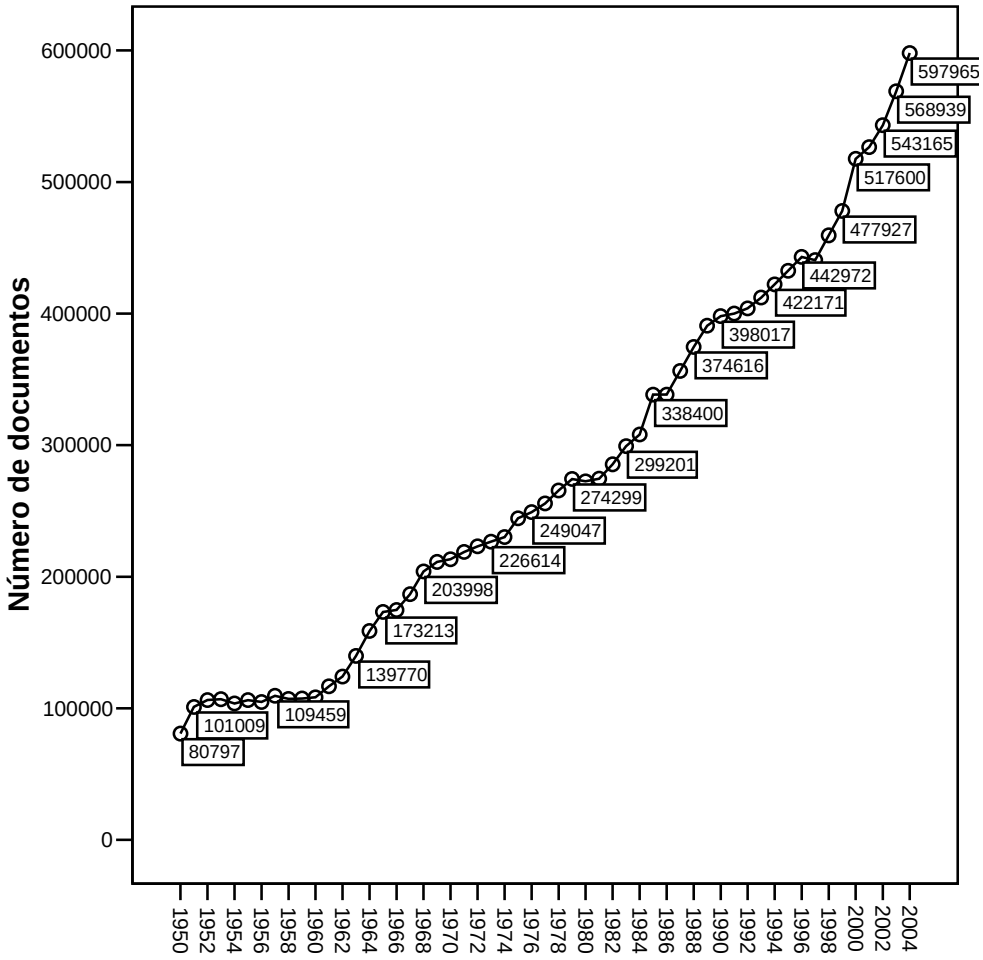


Figura V.0.1: Comportamiento de la indización por años en MedLine. (Cantidad Aproximada de Documentos 15, 200, 000 hasta finales del 2004)

El proceso de indización en la Categoría de Ciencias Físicas

Se ha mencionado en la sección 3.4. que la Biblioteca Nacional de Medicina de Estados Unidos utiliza el *MeSH Vocabulary* para indizar la literatura biomédica. El tesauro es revisado anualmente por un equipo de profesionales. El *MeSH Vocabulary* está organizado en 15 categorías principales y cada categoría se ramifica en series de subcategorías cada vez más concretas o específicas. La Categoría de Ciencias Físicas *(Physical Sciences Category)* es una de las 15 categorías que integran el *MeSH Vocabulary.*

En esta categoría se encuentran subcategorías relacionadas con temas físicos, químicos, matemáticos, biológicos, etc. En la figura V.0.2 se muestra la proporción de documentos que han sido indizados con algún término correspondiente de esta categoría.

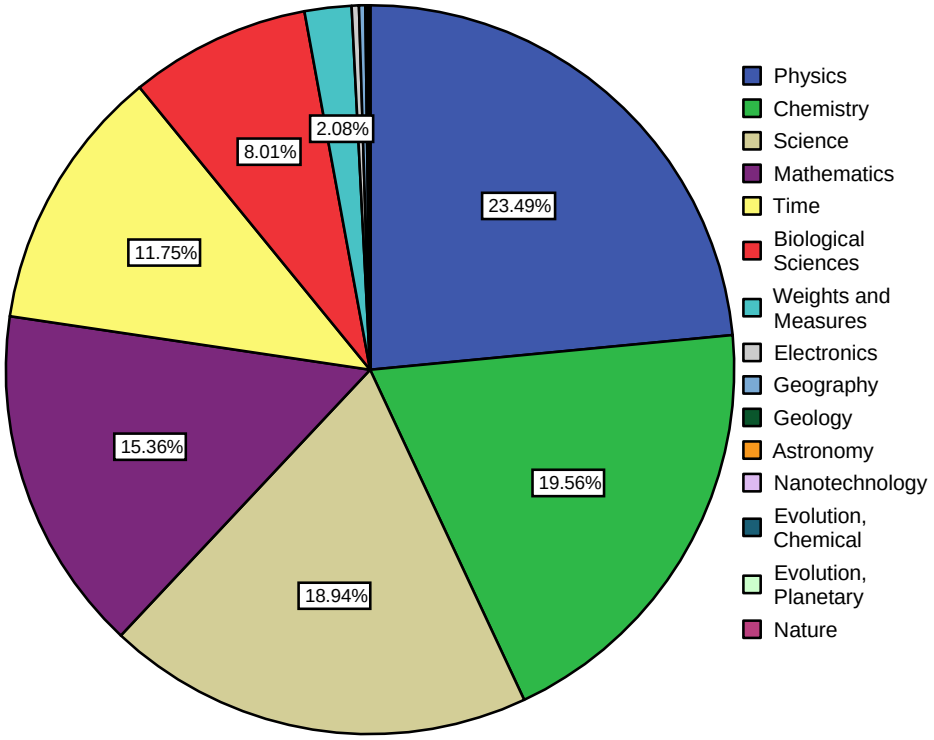


Figura V.0.2: Proporción de documentos indizados con algún termino perteneciente a la Categoría de Ciencias Físicas (Total de documentos 6, 411, 641 hasta el primer semestre del 2005).

Se aprecia que *Physics, Chemistry, Science, Mathematics, Time,* y *Biological Sciences* poseen porcentajes altos. Mientras que *Astronomy, Electronics, Evolution, Chemical, Evolution, Planetary, Geography, Geology, Nanotechnology, Nature* y *Weights and Measures* poseen porcentajes bajos.

El proceso de indización en *Mathematics*

En este trabajo solamente se consideran los términos pertenecientes a *Mathematics*. Esto se debe al interés que existe en el Laboratorio de Dinámica No Lineal de la Facultad de Ciencias de la Universidad Nacional Autónoma de México, por observar el comportamiento de los términos MeSH que integran la subcategoría de Ciencias Biologicas desde la perspectiva de *Mathematics*. Los términos de la subcategoría de Ciencias Biologicas contituyen las variables mientras que los términos de *Mathematics* constituyen las componentes de las variables.

Solamente, ocho términos son considerados por La Biblioteca Nacional de Medicina de Estados Unidos para realizar el proceso de indización de la literatura biomédica. Los términos que integran *Mathematics* son los siguientes: *Algorithms, Finite Element Anayisis, Fourier Anayisis, Fractals, Game Theory, Mathematical Computing, Nonlinear Dynamics, Statistics*[77].

Las figuras V.0.3 y V.0.4 muestran el número de documentos que se indizan por año con los términos que integran *Mathematics*.

[77] Ver Apéndice B para una breve descripción de cada término

En la figura V.0.3 se aprecia que el número de documentos indizados con *Finite Element Anayisis, Fourier Anayisis, Fractals, Game Theory, Nonlinear Dynamics* han estado creciendo unos más que otros.

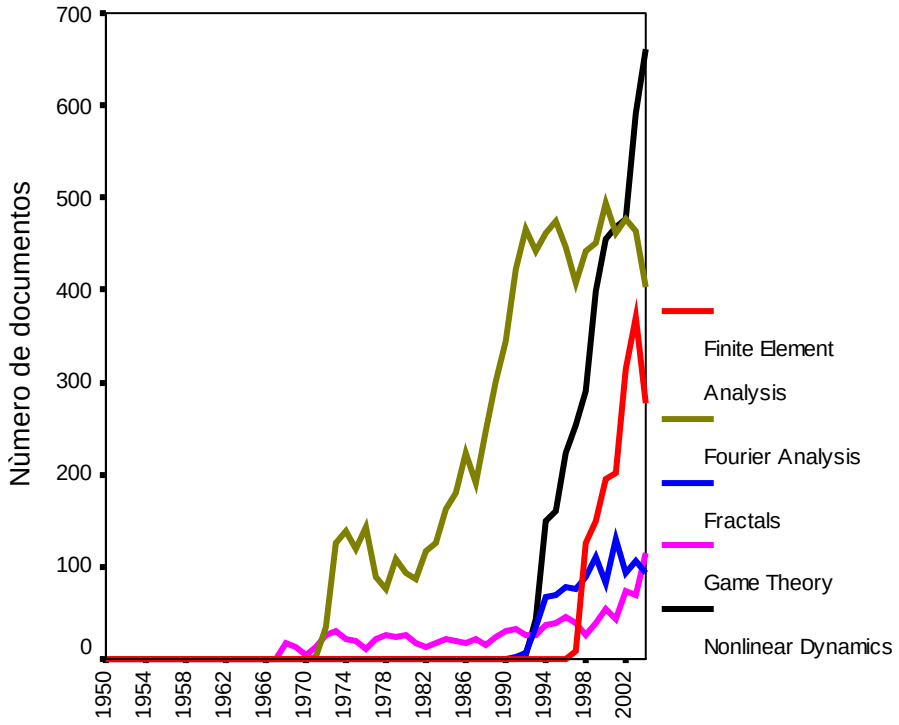


Figura V.0.3: Desarrollo de los términos que integran Mathematics a partir de 1950.

En la figura V.0.4 se aprecia que que el número de documentos indizados con *Algorithms* y *Mathematical Computing* ha estado incrementandose a partir de mediados de la década de los ochenta. Mientras que el número de documentos indizados con *Mathematics* y *Statistics* ha estado decreciendo.

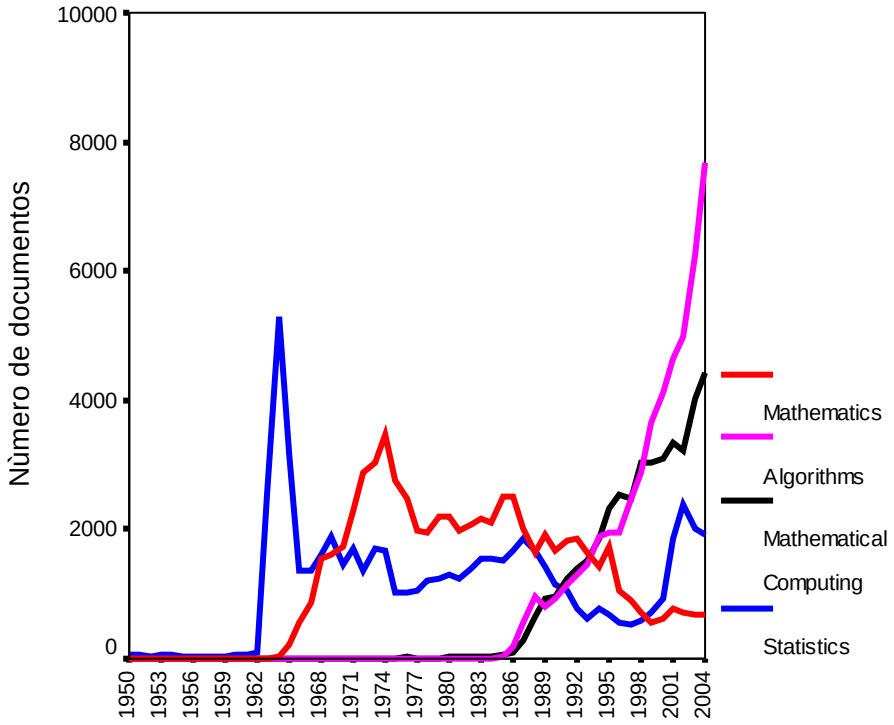


Figura V.0.4: Desarrollo de los términos que integran Mathematics a partir de 1950.

En resumen, estos términos se utilizan para:

Mathematics contituye el nombre de la subcategoría pero también es un término. El término *mathematics* incluye los temas de aritmética, de geometría o de cálculo. Se observa que el número de documentos indizados con *mathematics* se mantuvo en crecimiento hasta el año de 1974 y a partir de esta fecha ha estado decreciendo.

Mientras que el número de documentos indizados con *Algorithms* y *Mathematical Computing* se disparó desde mediados de la década de los ochentas. Estos términos indican que en cierta investigación biomédica se utilizo algún algoritmo y/o algún programa de computadora, respectivamente.

Se aprecia que el número de documentos indizados con *Nonlinear Dynamics* se ha mantenido creciendo desde principios de la década de los setentas. Este término hace referencia a la modelación de comportamientos caóticos, periódicos e irregulares de algunos sistemas dinámicos, por ejemplo, del corazón.

Además, se aprecia que el número de documentos indizados con los términos correspondientes a *Fourier Analysis, Finite Element Analysis, Game Theory* y *Fractals* se incrementó a partir de la década de los noventa. El término *Fourier Analysis* se emplea para el análisis de moléculas, el ADN, etc. Mientras que *Finite Element Analysis* se emplea para analizar el comportamiento de estructuras. Por otra parte, *Game Theory* se emplea para el análisis de situaciones en donde hay interpolación entre jugadores. Y finalmente, tenemos que *Fractals* se emplea para el análisis de patrones provenientes de algunas estructuras biológicas.

Y por último, el término *Statistics* hace referencia a métodos estadísticos que se emplean en la investigación biomedica. Estos métodos estadísticos no estan incluidos en la tabla V.0.1. Al igual que *mathematics*, el término *Statistics* consituye un término en si y el nombre de la subcategoría cuyos términos que la integran se muestran en la tabla V.0.1.

Todas estas “temáticas” se utilizan en la investigación biomédica, por ejemplo, en epidemiología para detectar variaciones en datos; construir intervalos de confianza, etc.; a través del diseño de un experimento, una prueba de hipótesis, etc.

Temáticas de Estadística
Statistics +
a) Actuarial Analysis + • Life Tables • Quality-Adjusted Life Years b) Analysis of Variance + • Multivariate Analysis c) Biometry d) Cluster Analysis + • Small-Area Analysis • Space-Time Clustering e) Confidence Intervals f) Data Interpretation, Statistical g) Discriminant Analysis h) Factor Analysis, Statistical i) Matched-Pair Analysis j) Monte Carlo Method k) Principal Component Analysis l) Probability + • Bayes Theorem • Likelihood Functions • Markov Chains • Odds Ratio • Predictive Value of Tests • Proportional Hazards Models • Risk • Logistic Models • Risk Assessment • Risk Factors • Uncertainty m) Regression Analysis + • Least-Squares Analysis • Linear Models • Logistic Models • Proportional Hazards Models n) ROC Curve o) Sensitivity and Specificity p) Statistical Distributions + • Binomial Distribution • Chi-Square Distribution • Normal Distribution • Poisson Distribution q) Statistics, Nonparametric r) Stochastic Processes + • Markov Chains s) Survival Analysis + • Disease-Free Survival
+ indica desglose

Tabla V.0.1: "Temáticas" de Estadística

Para dar una idea del uso de estas “temáticas” en la indización vea la Figura V.0.5, la cual muestra el porcentaje de documentos que han sido indizados con algúna “temática” perteneciente a *Statistics.*

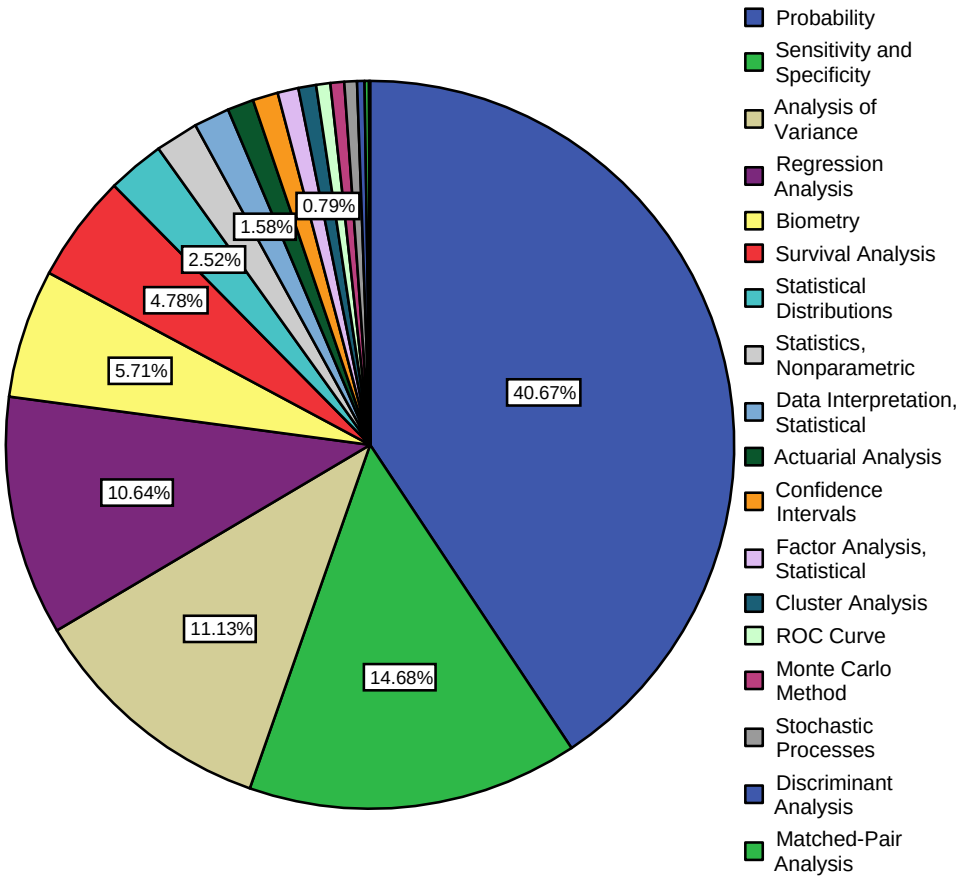


Figura V.0.5: Proporción de documentos indizados con algún termino perteneciente a Statistics (Total de documentos 932,144 hasta el primer semestre del 2005).

Se observa que *Probability* es la “temática” dominante. Mientras que *Sensitivity and Specificity, Analysis of Variante* y *Regression Analysis* son “temáticas” que poseen porcentajes medios.

Y para concluir esta revisión, vea la Figura V.0.6, la cual muestra el porcentaje de documentos que han sido indizados con alguno de los términos que integran *Mathematics.*

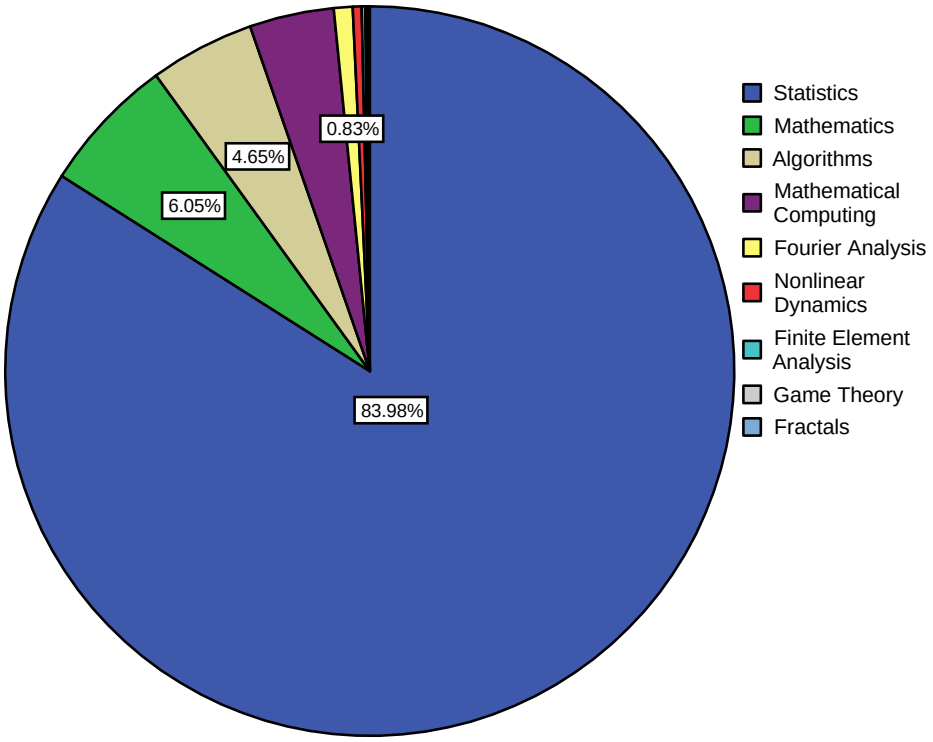


Figura V.0.6: Proporción de documentos indizados con algún termino perteneciente a ***Mathematics*** (Total de documentos 1, 037, 883 hasta el primer semestre del 2005)

Se aprecia que la proporción de documentos indizados con *Statistics* o alguna de sus "temáticas" es alta. También, se aprecia que la proporción de documentos indizados con *Mathematics* y *Algorithms* es baja comparada con *Statistics*. Y para *Mathematical Computing, Fourier Analysis Nonlinear Dynamics*, *Finite Element Analysis, Game Theory* y *Fractals* los porcentajes son bajos.

En resumen, la subcategoría *Mathematics* ocupa un lugar importante dentro de la Categoría de Ciencias Físicas. Otro aspecto importante, consiste en el enorme volumen de documentos que son indizados al año con alguna "temática" de *Statistics*. En contraste, con el volumen de documentos que son

indizados con los restantes términos que integran *Mathematics*. Esto no significa que estos últimos no sean importantes sino que su uso depende en gran medida del tipo de investigación que se realice.

Una investigación que desee comprobar hipótesis; seleccionar los datos más representativos de una muestra; intervalos de confianza; utilizará métodos estadísticos. Mientras que una investigación que trate de modelar algun tipo de fenómeno utilizará herramientas más apropiadas que las estadísticas, por ejemplo, los sistemas dinámicos y la matemática computacional.

En el apéndice C se muestran las estadísticas que se utilizaron en la elaboración de las figuras de esta sección.

5.2 ViBlioSOM

Así, como el *MeSH Vocabulary* contiene una categoría dedicada a las Ciencias Físicas *(Physical Sciences Category)* también contiene una categoría dedicada a las Ciencias Biologícas *(Biological Sciences Category),* la cual está integrada por las siguientes subcategorías:

001 Biological Sciences
002 Health Occupations
003 Environment and Public Health
004 Biological Phenomena, Cell Phenomena, and Immunity
005 Genetic Processes
006 Biochemical Phenomena, Metabolism and Nutrition
007 Physiological Processes
008 Reproductive and Urinary Physiology
009 Circulatory and Respiratory Physiology
010 Digestive, Oral, and Skin Physiology
011 Musculoskeletal, Neural, and Ocular Physiology
012 Chemical and Pharmacologic Phenomena
013 Genetic Phenomena
014 Genetic Structures

En el apéndice D se muestra una porción de esta categoría. Note que esta categoría abarca una enorme cantidad de temas biológicos.

El Análisis Bibliométrico tiene por objetivo mostrar el comportamiento de los términos MeSH pertenecientes a la subcategoría “Biological Sciences” desde la perspectiva de la subcategoría de Matemáticas. Para ello, se hacen los siguientes tres análisis bibliométricos:

- Distribución de Terminos.
- Recuperación de Información.
- Descubrimiento de Conocimiento.

Para realizar los Análisis Bibliométricos se usa la metodología ViBlioSOM, la cual contiene las siguientes etapas[78]:

1) Adquisición y Selección de Documentos.
2) Preprocesamiento.
3) Minería de Datos y Textos a partir de Indicadores Bibliométricos.
4) Visualización e Interpretación de los Resultados.

Adquisición y Selección de Ficheros

Los análisis bibliométricos se hacen en el almacén de matemáticas debido a que en los otros dos almacenes el poder de computo requerido es alto. Para esta etapa de la metodología BiVlioSOM se construyeron los siguientes *almacenes*.

El almacén de matemáticas contiene documentos que han sido indizados exclusivamente con los siguientes términos matemáticos: *Mathematics; Algorithms; Finite Element Analysis; Fourier Analysis; Fractals; Game Theory; Games, Experimental; Mathematical Computing; Decisión Support Techniques; Decision Theory; Decision Trees; Nomograms; Neural Networks (Computer); Nonlinear Dynamics.*

El almacén de estadística contiene documentos que han sido indizados exclusivamente con las "temáticas" de Estadística. (Ver tabla 1).

El almacén de intersección contiene los documentos que poseen términos matemáticos y "temáticas" estadísticas, es decir, documentos que se encuentran en la intersección de los almacenes de matemáticas y estadísticas.

El almacén de matemáticas contiene un total de 116, 612 documentos mientras que el almacén de estadística contiene un total de 887, 877 documentos y el almacén de intersección contiene un total de 51, 311 documentos. En la figura siguiente se muestra el tamaño de estos almacenes

[78] Vea sección 1.5.3

en porcentajes. Observe el porcentaje bajo de documentos que se encuentran en la intersección.

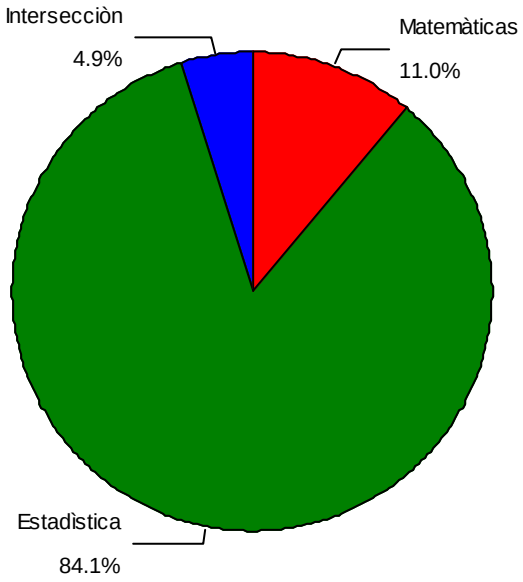


Figura V.0.7: Almacenes (Total de documentos 1, 055, 800)

Los almacenes se obtuvieron con las siguientes estrategias de búsqueda.

Almacén de Matematicas

("Mathematics"[MeSH:NoExp] NOT "Statistics"[MeSH]) OR ("Algorithms"[MeSH] NOT "Statistics"[MeSH]) OR ("Finite Element Analysis"[MeSH] NOT "Statistics"[MeSH]) OR ("Fourier Analysis"[MeSH] NOT "Statistics"[MeSH]) OR ("Fractals"[MeSH] NOT "Statistics"[MeSH]) OR ("Game Theory"[MeSH] NOT "Statistics"[MeSH]) OR ("Mathematical Computing"[MeSH] NOT "Statistics"[MeSH]) OR ("Nonlinear Dynamics"[MeSH] NOT "Statistics"[MeSH])

Almacén de Estadistica

((“Statistics”[MeSH]) NOT (“Mathematics”[MeSH] OR “Algorithms”[MeSH] OR “Finite Element Analysis”[MeSH] OR “Fourier Analysis”[MeSH] OR “Fractals”[MeSH] OR “Game Theory”[MeSH] OR “Games, Experimental”[MeSH] OR “Mathematical Computing”[MeSH] OR “Decisión Support Techniques”[MeSH] OR “Decision Theory”[MeSH] OR “Decision Trees”[MeSH] OR “Nomograms”[MeSH] OR “Neural Networks (Computer)”[MeSH] OR “Nonlinear Dynamics”[MeSH])).

Almacén de Intersección

("Mathematics"[MeSH:NoExp] AND "Statistics"[MeSH]) OR ("Algorithms"[MeSH] AND "Statistics"[MeSH]) OR ("Finite Element Analysis"[MeSH] AND "Statistics"[MeSH]) OR ("Fourier Analysis"[MeSH] AND "Statistics"[MeSH]) OR ("Fractals"[MeSH] AND "Statistics"[MeSH]) OR ("Game Theory"[MeSH] AND "Statistics"[MeSH]) OR ("Mathematical Computing"[MeSH] AND "Statistics"[MeSH]) OR ("Nonlinear Dynamics"[MeSH] AND "Statistics"[MeSH]).

Los documentos en los almacenes están en el formato MedLine y abarcan el periodo entre los años entre 1950 y el 2004. En el capítulo 3, sección 3.3 se muestra un ejemplo de búsqueda y recuperación de documentos utilizando el término *Nonlinear Dynamics* (Dinámica no Lineal).

2) Preprocesamiento

El formato MEDLINE, contiene todos los términos que fueron asignados a cada documento por el equipo de indización de la Biblioteca Nacional de Medicina de Estados Unidos. Estos términos *"sintetizan"* el contenido de cada documento. En esta etapa, utilizamos el indicador bibliométrico palabras comunes (sección 2.2.3), para obtener mapas que describan las asociaciones más significativas de los términos seleccionados en el conjunto de documentos. Para resaltar las asociaciones que muestre el indicador se utilizan los siguientes dos criterios de normalización:

La coocurrencia normalizada de acuerdo al criterio de Jaccard: resalta la relación cognitiva entre dos términos de la matriz de coocurrencia. La formula matemática del índice correspondiente es la siguiente:

$$\frac{C_{ij}}{f_i + f_j - C_{ij}}$$

Donde C_{ij} es la coocurrecia entre los términos i *y* j. Mientras que f_i y f_j son las frecuencias de los términos i y j respectivamente. En términos generales, el criterio de Jaccard toma valores entre el 0 y el 1. Alcanza el valor 0 cuando existe una relación nula entre dos temas. Alcanza el valor 1 cuando existe una relación fuerte entre dos temas Y está entre los valores 0 y 1 dependiendo de la relación existente entre dos temas.

La coocurrencia normalizada de acuerdo al criterio de Courtial: se define a través del coeficiente de aproximación e entre términos. La fórmula matemática del criterio es la siguiente:

$$e = \frac{C_{ij}^{2}}{f_i f_j}$$

Donde C_{ij} es la coocurrecia entre los términos i y j. Mientras que f_i y f_j son las frecuencias de los términos i y j respectivamente. En términos generales, el valor de e será 1 cuando coincidan dos términos y 0 si no coinciden nunca. La formula se rescribirse como:

$$e = \frac{C_{ij}}{f_i} x \frac{C_{ij}}{f_j}$$

Donde el primer factor es la probabilidad de tener la palabra i cuando se tiene la palabra j, y el segundo es la probabilidad de tener la palabra j cuando se tiene la palabra i. Este índice es una medida de la relación "Y" entre las palabras i y j.

A continuación se presentan algunos ejemplos numéricos de estos criterios.

Caso 1: Ambas palabras poseen frecuencias grandes pero aparecen conjuntamente poco en los documentos, es decir, coocurren poco.

Coocurrencia	**Palabra j**	**Frecuencias**		**Criterios**	
				Jaccard	**Courtial**
Palabra i	100	**Palabra i**	500	0.11111111	0.04
		Palabra j	500		

Tabla V.0.2: Caso 1

En este caso ambos criterios dan valores pequeños, por lo tanto el análisis de palabras comunes considera una relación débil entre las palabras.

Caso 2: La palabra i posee frecuencia alta y la palabra j posee una frecuencia menor que la palabra i pero aparecen conjuntamente poco en los documentos, es decir, coocurren poco.

Coocurrencia	Palabra j	Frecuencias		Criterios	
				Jaccard	Courtial
Palabra i	25	Palabra i	500	0.02941176	0.00333333
		Palabra j	375		

Tabla V.0.3: Caso 2

Al igual que en el caso 1 ambos criterios dan valores pequeños, por lo tanto el análisis de palabras comunes seguirá considerando una relación débil entre las palabras.

Caso 3: Ambas palabras tiene frecuencias parecidas pero aparecen conjuntamente mucho en los documentos, es decir, concurren mucho.

Coocurrencia	Palabra j	Frecuencias		Criterios	
				Jaccard	Courtial
Palabra i	450	Palabra i	500	0.81818182	0.81
		Palabra j	500		

Tabla V.0.4: Caso 3

En este caso los valores de los criterios son altos y por lo tanto el análisis de palabras comunes considerara una relación fuerte entre las palabras.

Caso 4: Casos extremos.

En el primer caso extremo ambas palabras poseen la misma frecuencia y aparecen conjuntamente en los mismos documentos, es decir, concurren mucho. En este caso ambos criterios alcanzan su valor máximo y por lo tanto el análisis de palabras comunes considera la unión fuerte entre las palabras.

Matriz	Palabra j	Frecuencias		Criterios	
				Jaccard	Courtial
Palabra i	500	**Palabra i**	500	1	1
		Palabra j	500		

Tabla V.0.5: Casos extremos 1

El segundo caso extremo corresponde a la siguiente situación. Ambas palabras poseen frecuencia uno y poseen coocurrencia uno, es decir, ambas palabras están en el mismo documento.

Coocurrencia	Palabra j	Frecuencias		Criterios	
				Jaccard	Courtial
Palabra i	1	**Palabra i**	1	1	1
		Palabra j	1		

Tabla V.0.6: Caso extremo 2

En este caso ambos criterios también alcanzan su valor máximo y por lo tanto el análisis de palabras comunes considera la unión fuerte entre las palabras. Pero, obviamente el análisis de palabras comunes se equivoco. Para que esto no suceda debemos considerar Ley de Zipf. Dicha ley dice que la frecuencia de aparición de palabras en un texto es muy baja en la mayoría de los casos, por lo que la mayor parte de las palabras serán poco abundantes y pueden ser despreciadas [74].

Para resaltar aun más el poder de estos criterios se puede establecer un umbral, es decir, los criterios toman el valor cero si los términos coinciden por debajo de algún umbral preestablecido.

En definitiva, mediante el uso de estos criterios el análisis de palabras asociadas es capaz de discernir qué palabras y qué asociaciones son realmente relevantes en los analisis bibliometricos y por su puesto eliminar aquellas que por su baja co-ocurrencia relativa o su elevada generalidad no lo son.

En las etapas correspondientes a la *Minería de Datos y Textos a partir de Indicadores Bibliométricos* y a la *Visualización e Interpretación de los Resultados* de la metodología ViBlioSOM se utiliza el software *Viscovery SOMine*

Enterprise Edition Version 4.0, el cual, utiliza el algoritmo SOM para la generación de mapas.

Financiamiento de la investigación

La importancia de la investigación biomédica para el desarrollo científico de los países y para el bienestar de su población es un hecho incuestionable en la actualidad. Sin embargo, también es una realidad que la investigación conlleva cada vez castos más altos, por su creciente especialización y complejidad (véase por ejemplo el caso de la investigación en genoma humano) y además, los recursos económicos que pueden destinarse a ella son limitados.

La Biblioteca Nacional de Medicina de los Estados Unidos indica si los documentos provienen de alguna investigación financiada. Para ello, pone las siguientes etiquetas en los documentos:

Research Support, Non-U.S. Govt.
(No financiada por el gobierno de E. U. Sin embargo, la investigación fue financiada por alguna sociedad, institución, universidades, organizaciones privadas, etc.)

Research Support, U.S. Govt, P.H.S[79].
(Financiada por la agencia gubernamental P.H.S.)

Research Support, U.S. Govt, Non-P.H.S.
(No financiada por la agencia gubernamental P.H.S. Sin embargo, la investigación fue financiada por otra agencia gubernamental)

Research Support, N.I.H[80]., Extramural.
(Financiada por la N. I. H. Extramural)

Las etiquetas fueron identificadas en los almacenes de matematicas e intersección. A continuación se muestra la cantidad de documentos que recibieron algún financiamiento o por el contrario no recibieron financiamiento.

[79] Preventive Health Service
[80] National Institute of Health

Almacén de Matematicas	
Tipo de financiamiento	**Cantidad de documentos**
Research Support, Non-U.S. Govt.	30, 177
Research Support, U.S. Govt, P.H.S.	18, 675
Research Support, U.S. Govt, Non-P.H.S.	9, 752
Documentos que no tienen etiqueta	58, 008
Total de documentos en el almacén	**116, 612**

Tabla V.0.7: Financiamiento en el almacén de matemáticas

Almacén de Traslape	
Tipo de Financiamiento	**Cantidad de documentos**
Research Support, Non-U.S. Govt.	16, 283
Research Support, U.S. Govt, P.H.S.	7, 898
Research Support, U.S. Govt, Non-P.H.S.	3, 834
Documentos que no tienen etiqueta	23, 296
Total de documentos en el almacén	**51, 311**

Tabla V.0.8: Financiamiento en el almacén de intersección

Las cifras globales son:

Tipo de Financiamiento	**Cantidad de documentos**
Research Support, Non-U.S. Govt.	46, 460
Research Support, U.S. Govt, P.H.S.	26, 573
Research Support, U.S. Govt, Non-P.H.S.	13, 586
Research Support, N.I.H., Extramural	154
Documentos que no tienen etiqueta	86, 773
Total de documentos en el almacén	**168, 077**

Tabla V.0.9: Cifras globales de financiamiento

Se aprecia que sociedades, instituciones, universidades, organizaciones privadas, etc., financian muchos trabajos del área biomédica, además de todas las agencias del gobierno de Estados Unidos como la P. H. S., que tienen una participación primordial en el financiamiento de este tipo de trabajos.

5.3 Mapas Auto - Organizantes

Se presentan los mapas auto – organizantes para los almacenes de matemáticas, intersección y unión. Este último es el resultado de la unión de los dos primeros almacenes. La presentación se lleva a cabo en la siguiente forma: en primer lugar se exploran los mapas y en segundo lugar se presentan las propiedades generales de los mapas.

La exploración de los mapas se puede entender como el proceso por medio del cual, se tiene la intención de extraer información valiosa, partiendo de la inspección visual de una gama de mapas. Estos mapas se llaman mapa de regiones y mapas de componentes.

En el mapa de regiones, las regiones representan conglomerados del conjunto de datos multidimensional. El despliegue de los mapas de componentes tiene la particularidad de representar la distribución de los valores de cada variable de los datos en un mapa. Este mapa denominado "Component Picture" representa el promedio de los valores de la variable correspondiente a los datos asociados a cada nodo. La distribución de estos promedios se puede visualizar por medio de una escala de color, que corresponde al rango de valores que los datos toman en la variable correspondiente. Los valores mínimos están representados en color azul, los intermedios en verde y amarillo, y los valores máximos en rojo.

Las propiedades generales de los mapas pueden ser utilizadas durante el proceso de exploración y son significativas en casi cualquier aplicación. Estas propiedades son la preservación de la topología y la distribución de los datos en un despliegue ordenado; basándose en ellas es posible el establecimiento de relaciones entre variables, la visualización de “clusters” y la inspección de relaciones de vecindad entre los nodos en el mapa.

Los tres análisis bibliométricos se llevan a cabo con los mapas auto – organizantes del almacén de matemáticas.

5.3.1 Mapas Auto – Organizantes del Almacén de Matemáticas

Datos generales

Los datos que se utilizaron durante el entrenamiento de la Red Neuronal de Kohonen provienen de la matriz de concurrencia. Esta matriz se construyo de la siguiente manera:

Cada fila de la matriz de concurrencia representa un término de la Categoría de Ciencias Biológicas[81]. Esta categoría esta integrada por las siguientes subcategorías (En el apéndice D se muestra una porción de esta categoría)

001 Biological Sciences
002 Health Occupations
003 Environment and Public Health
004 Biological Phenomena, Cell Phenomena, and Immunity
005 Genetic Processes
006 Biochemical Phenomena, Metabolism and Nutrition
007 Physiological Processes
008 Reproductive and Urinary Physiology
009 Circulatory and Respiratory Physiology
010 Digestive, Oral, and Skin Physiology
011 Musculoskeletal, Neural, and Ocular Physiology
012 Chemical and Pharmacologic Phenomena
013 Genetic Phenomena
014 Genetic Structures

Cada columna de la matriz de concurrencia representa un término de la subcategoria de Mathematics. Esta categoría esta integrada por los siguientes términos: *Mathematics; Algorithms; Finite Element Analysis; Fourier Analysis; Fractals; Game Theory; Games, Experimental; Mathematical Computing; Decisión Support Techniques; Decision Theory; Decision Trees; Nomograms; Neural Networks (Computer); Nonlinear Dynamics*. En otras palabras, cada fila es una variable y cada columna una componente de la variable.

Matriz de coocurrencia

Población de documentos: 116, 612
Tamaño de la matriz de coocurrencia: 1970 x 14

A) Mapas Auto - Organizantes bajo el Criterio de Jaccard

Una vez que el entrenamiento ha finalizado, la Red Neuronal de Kohonen muestra un mapa, el cual, esta dividido en regiones. Estas regiones representan conglomerados del conjunto de datos.

Mapa de regiones: A continuación se presenta el mapa que muestra las 13 regiones que representan conglomerados del conjunto de datos. Las regiones están etiquetas con C1, C2, C3,…, C13.

[81] Biological Sciences Category

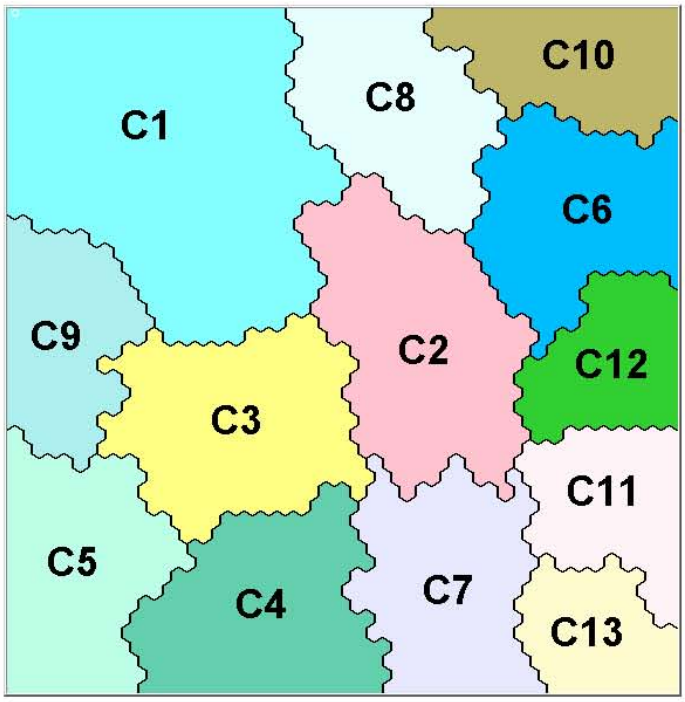


Figura V.0.8: Mapa auto – organizante que representa los conglomerados del conjunto de datos

Mapas de componentes: la exploración de los mapas de componentes ayuda a establecer relaciones entre las distintas variables. A continuación se presentan los mapas auto - organizantes de los 13 componentes. Los componentes corresponden a los siguientes términos de la subcategoría de "Mathematics": *Mathematics; Algorithms; Finite Element Analysis; Fourier Analysis; Fractals; Game Theory; Games, Experimental; Mathematical Computing; Decisión Support Techniques; Decision Theory; Decision Trees; Nomograms; Neural Networks (Computer); Nonlinear Dynamics*.

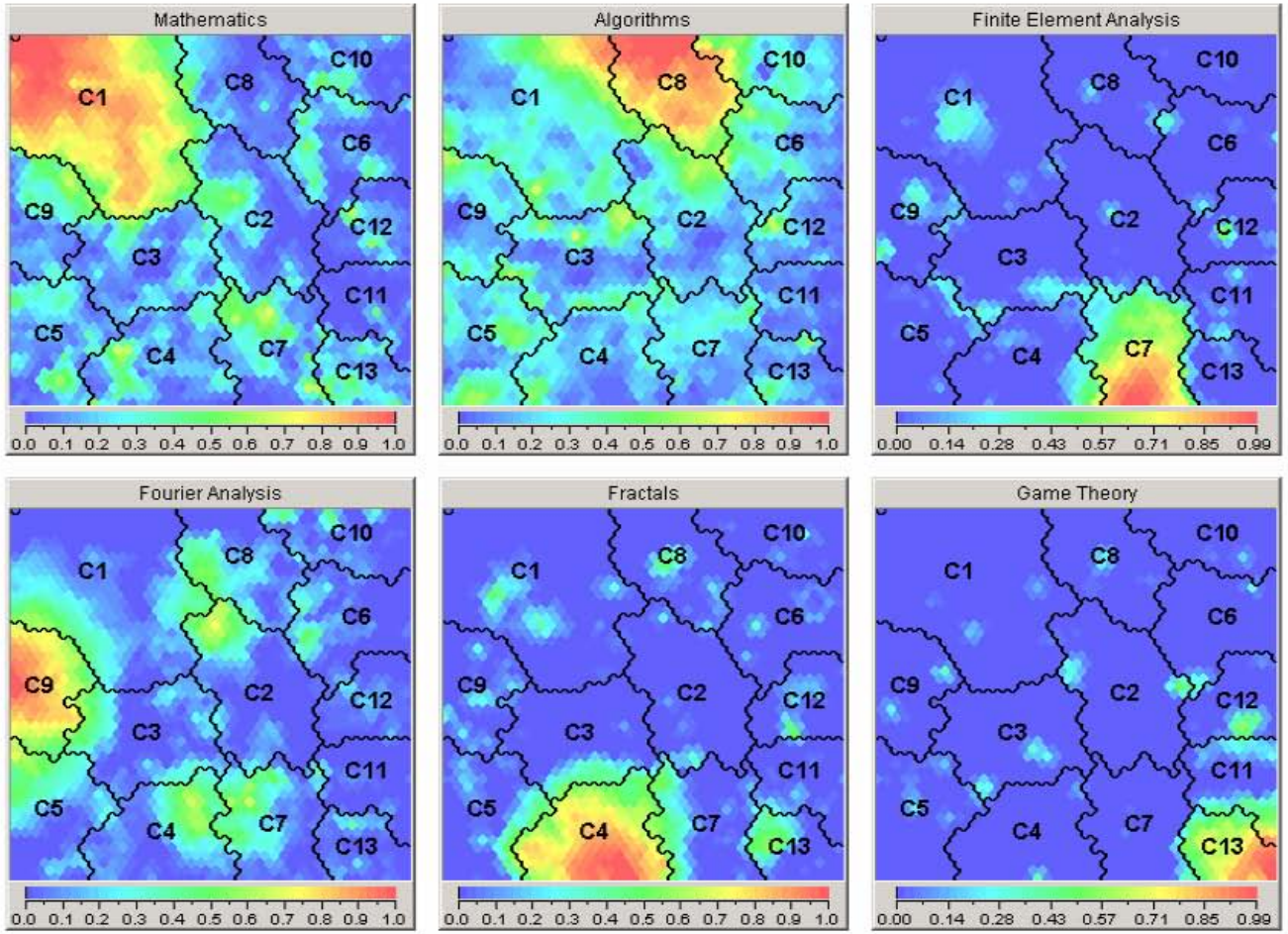


Figura V.0.9: Mapas de Componentes bajo el criterio de Jaccard

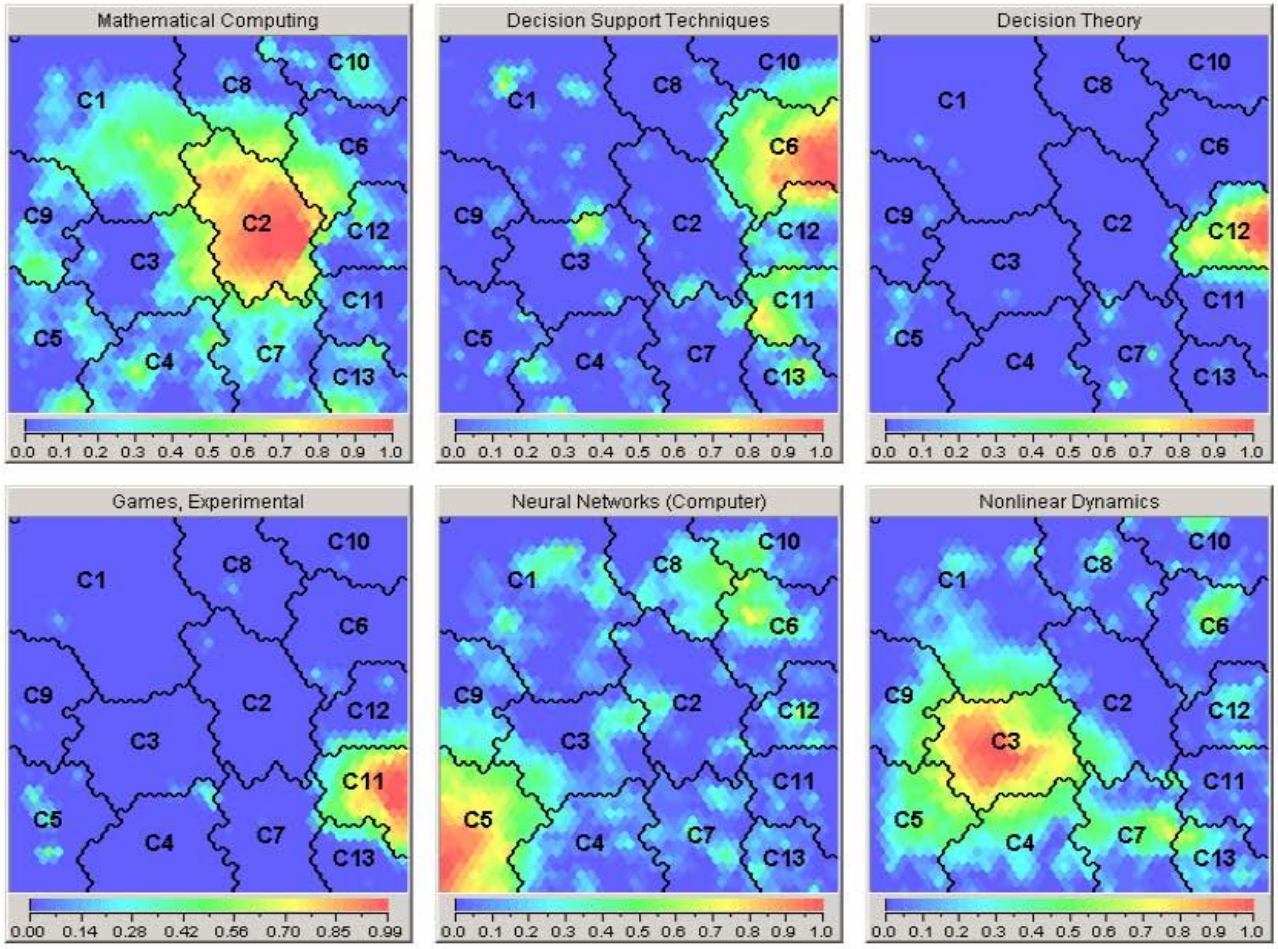


Figura V.0.9: Mapas de Componentes bajo el criterio de Jaccard

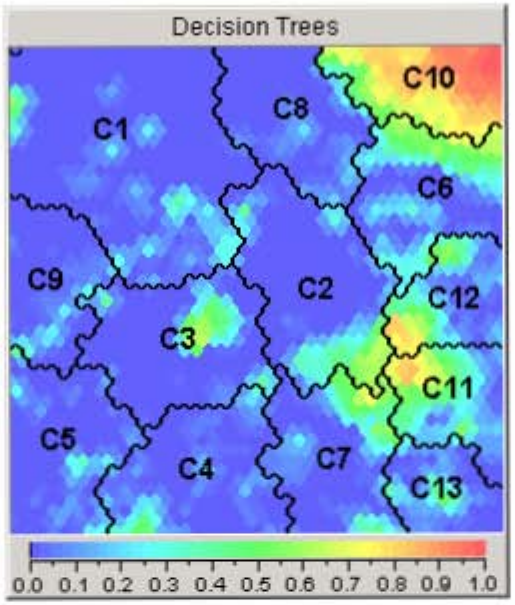


Figura V.0.9: Mapas de Componentes bajo el criterio de Jaccard

Observe que cada componente muestra una zona roja (núcleo) distinta. Esto se debe al comportamiento que cada variable (i.e., término) tiene en cada componente, es decir, hay variables que dependen más de algunas componentes que otras, por ejemplo, observe la componente Mathematics. Hay muchas variables que están influenciadas por esta componente.

Observe que cada región en el mapa está determinado por una componente.

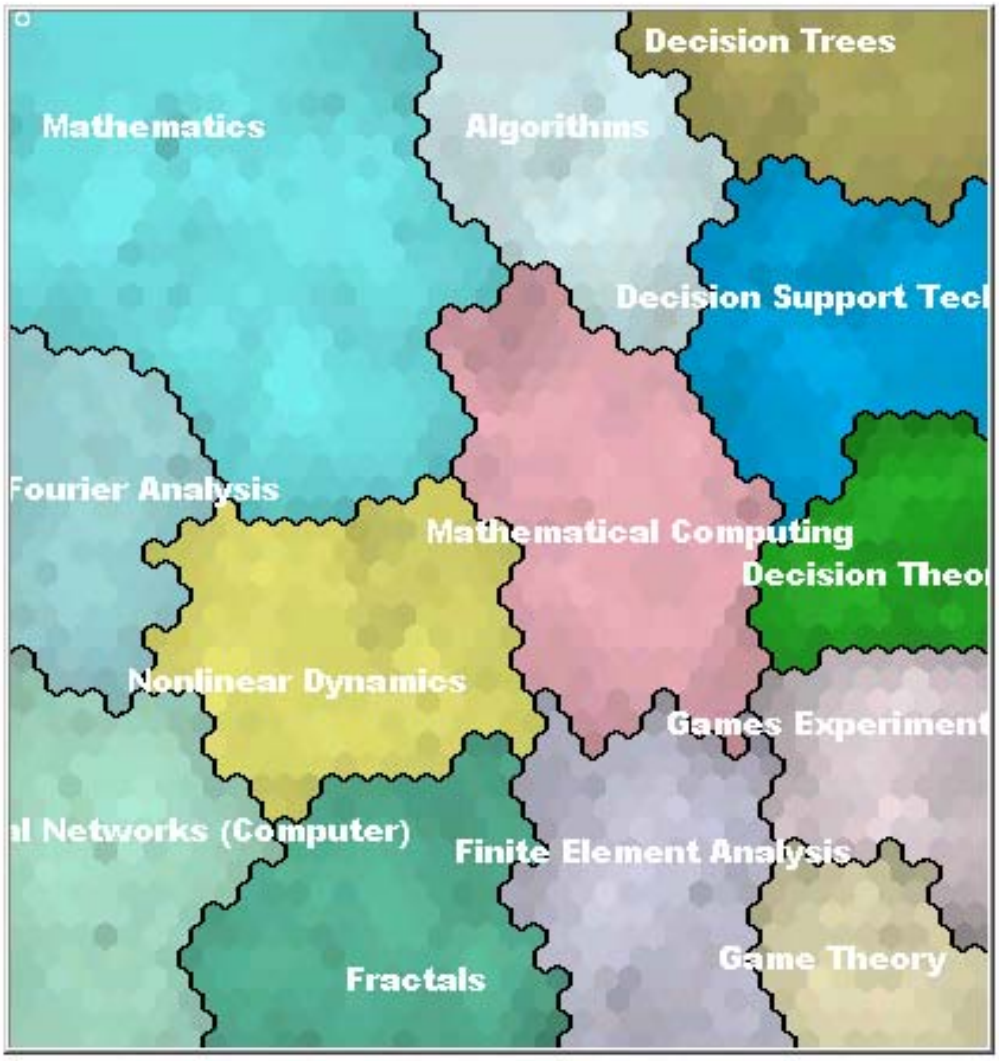


Figura V.0.10: Identificación de regiones bajo el criterio de Jaccard

B) Mapas Auto - Organizantes bajo el Criterio de Courtial

Siguiendo el mismo orden de presentación de mapas que en la sección anterior. Ahora se presentan los mapa auto - organizantes bajo el criterio de Courtial. En primer lugar se presenta el mapa de regiones y posteriormente se presentan los mapas de componentes.

Mapa de regiones: Observe que se obtuvieron 13 regiones. Tanto el criterio de Jaccard como el de Courtial muestran una region grande.

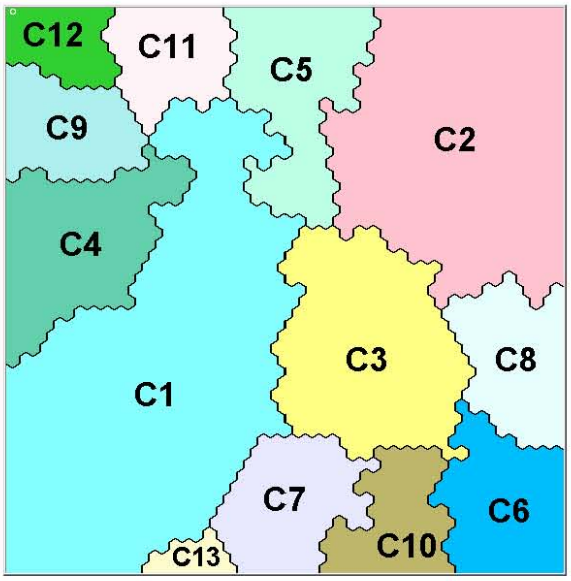


Figura V.0.11: Mapa auto – organizante que representa los conglomerados del conjunto de datos.

Mapas de Componentes: A continuación se presentan los mapas auto-organizantes de los 13 componentes. Los componentes corresponden a los siguientes términos de la subcategoría de "Mathematics": *Mathematics; Algorithms; Finite Element Analysis; Fourier Analysis; Fractals; Game Theory; Games, Experimental; Mathematical Computing; Decisión Support Techniques; Decision Theory; Decision Trees; Nomograms; Neural Networks (Computer); Nonlinear Dynamics*.

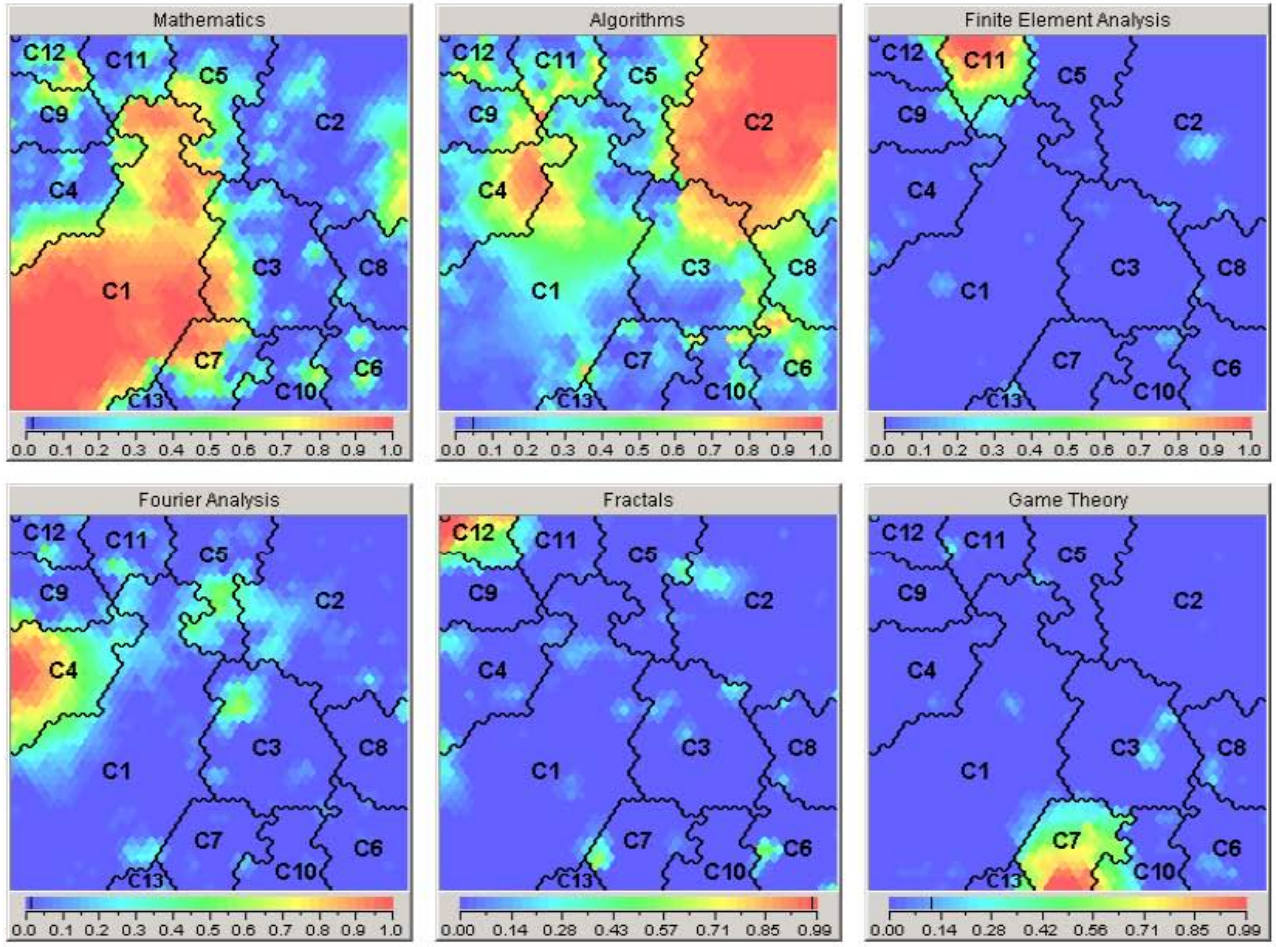


Figura V.0.12: Mapas de Componentes bajo el criterio de Courtial

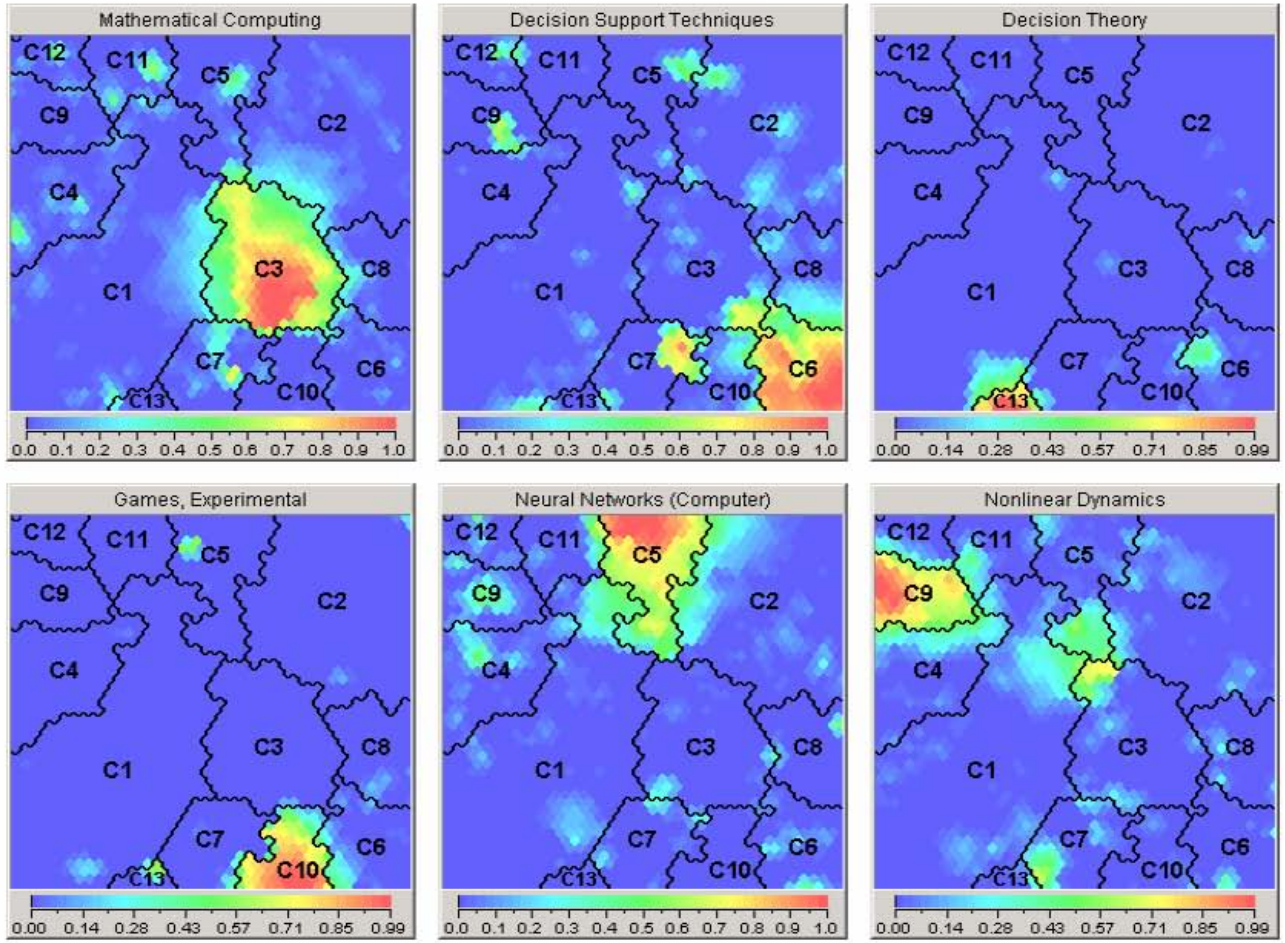


Figura V.0.12: Mapas de Componentes bajo el criterio de Courtial

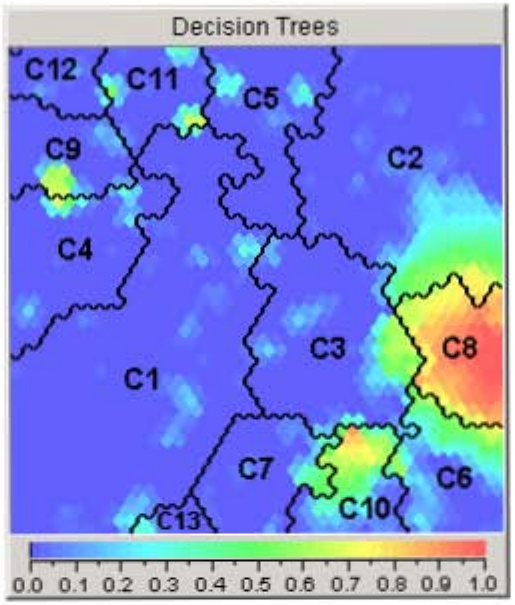


Figura V.0.12: Mapas de Componentes bajo el criterio de Courtial

Observe que Courtial muestra regiones más grandes en las componentes Mathematics y Algorithms que su contraparte Jaccard.

En este caso cada region en el mapa esta determinado por una componente.

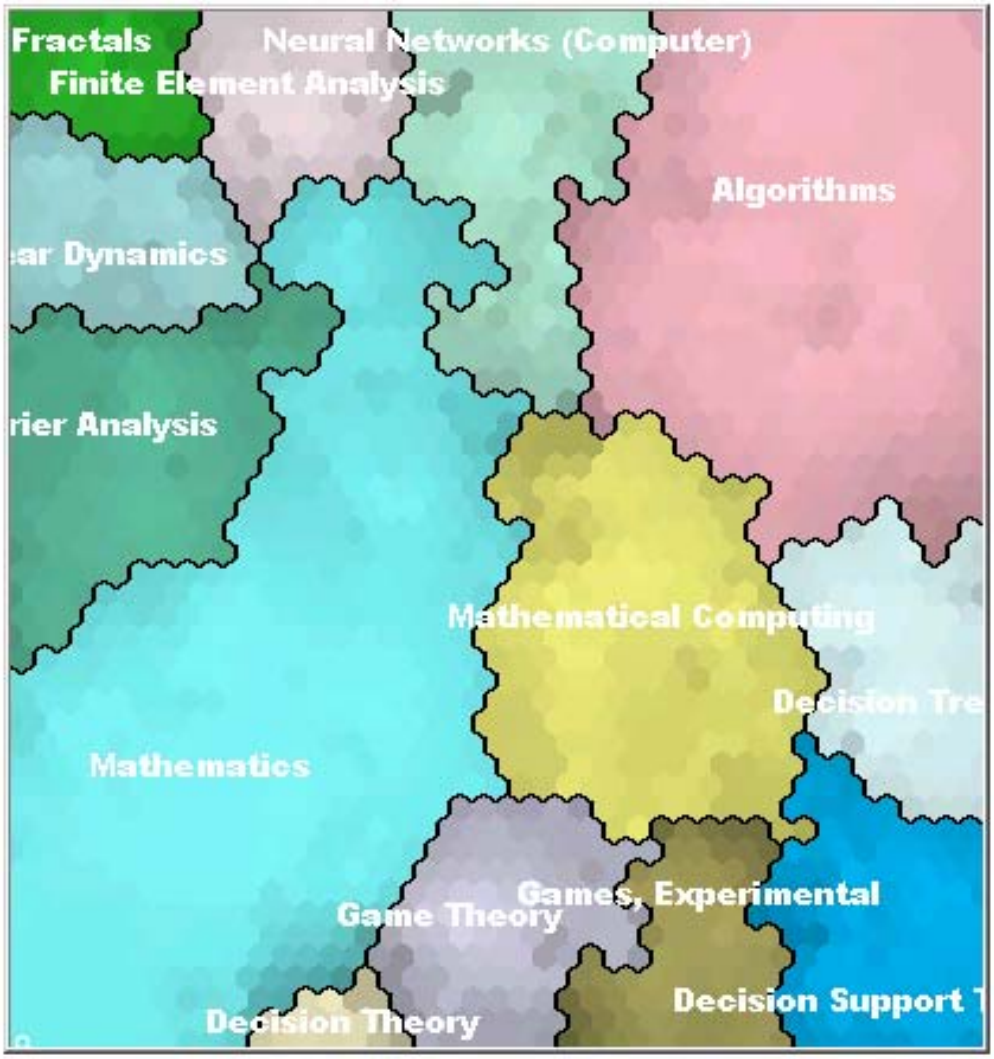


Figura V.0.13: Identificación de regiones bajo el criterio de Courtial

5.3.2 Mapas Auto - Organizantes del al Almacén de Intersección

Solamente, se muestran los mapas auto – organizantes de regiones y de doce componentes para cada criterio.

Datos generales

Los términos biológicos considerados (filas de la matriz de coocurrencia) pertenecen a la Categoría de Ciencias Biológicas *(Biological Sciences Category)*. Esta categoría esta integrada por las siguientes subcategorías.

001 Biological Sciences
002 Health Occupations
003 Environment and Public Health
004 Biological Phenomena, Cell Phenomena, and Immunity
005 Genetic Processes
006 Biochemical Phenomena, Metabolism and Nutrition
007 Physiological Processes
008 Reproductive and Urinary Physiology
009 Circulatory and Respiratory Physiology
010 Digestive, Oral, and Skin Physiology
011 Musculoskeletal, Neural, and Ocular Physiology
012 Chemical and Pharmacologic Phenomena
013 Genetic Phenomena
014 Genetic Structures

En el apéndice D se muestra una porción de esta categoría.

Los términos matemáticos considerados (columnas de la matriz de concurrencia) son los siguientes: *Mathematics; Algorithms; Finite Element Analysis; Fourier Analysis; Fractals; Game Theory; Games, Experimental; Mathematical Computing; Decisión Support Techniques; Decision Theory; Decision Trees; Nomograms; Neural Networks (Computer); Nonlinear Dynamics. Más las "temáticas" pertenecientes a Estadística.* Ver tabla 1.

Matriz de coocurrencia

Población de documentos: 51, 311
Tamaño de la matriz de coocurrencia: 1970 x 57

A) Mapas Auto - Organizantes bajo el Criterio de Jaccard

Esta vez el algoritmo de Kohonen obtuvo 5 regiones. Las regiones están etiquetadas con C1, C2,…., C5.

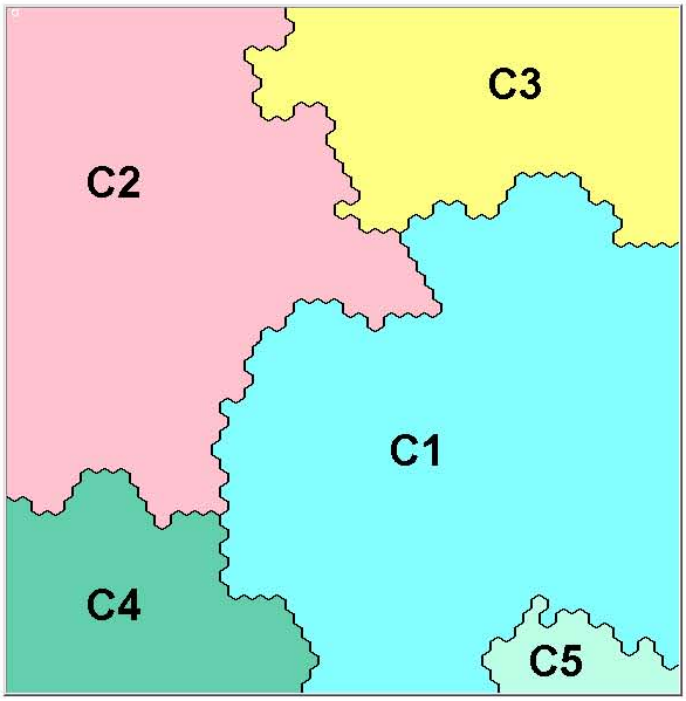


Figura V.0.14: Mapa auto – organizante que representa los conglomerados del conjunto de datos

Solamente, se muestran 12 mapas de componentes. El total de mapas de componentes es 57 que corresponden a los siguientes términos de la subcategoría de "Mathematics": *Mathematics; Algorithms; Finite Element Analysis; Fourier Analysis; Fractals; Game Theory; Games, Experimental; Mathematical Computing; Decisión Support Techniques; Decision Theory; Decision Trees; Nomograms; Neural Networks (Computer); Nonlinear Dynamics* más las "temáticas" de Statisticas. (Vea Tabla V.0.1)

Observe que algunas componentes, por ejemplo, Mathematics, Nonlinear Dynamics, Fourier Analysis, Stochastic Process, etc., influyen muchos sobre las variables mientras que en otros componentes tienen poca influencia.

Además, cada región esta integrada por varias zonas rojas (núcleos). Por ejemplo, la región “C2” contiene las zonas rojas (núcleos) de Mathematics, Fourier Analysis, Nonlinear Dynamics, etc.

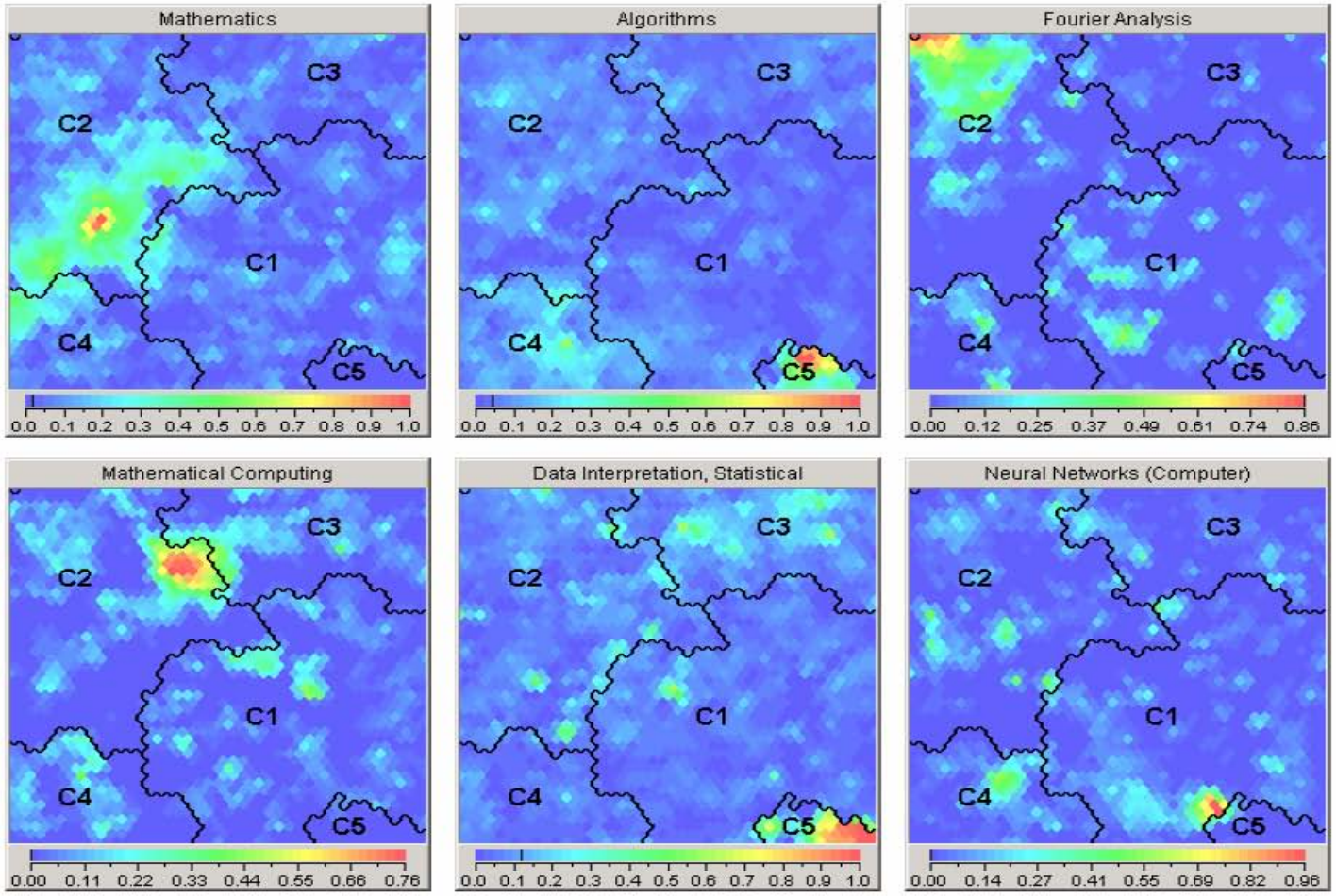


Figura V.0.15: Mapas de componentes bajo el criterio de Jaccard

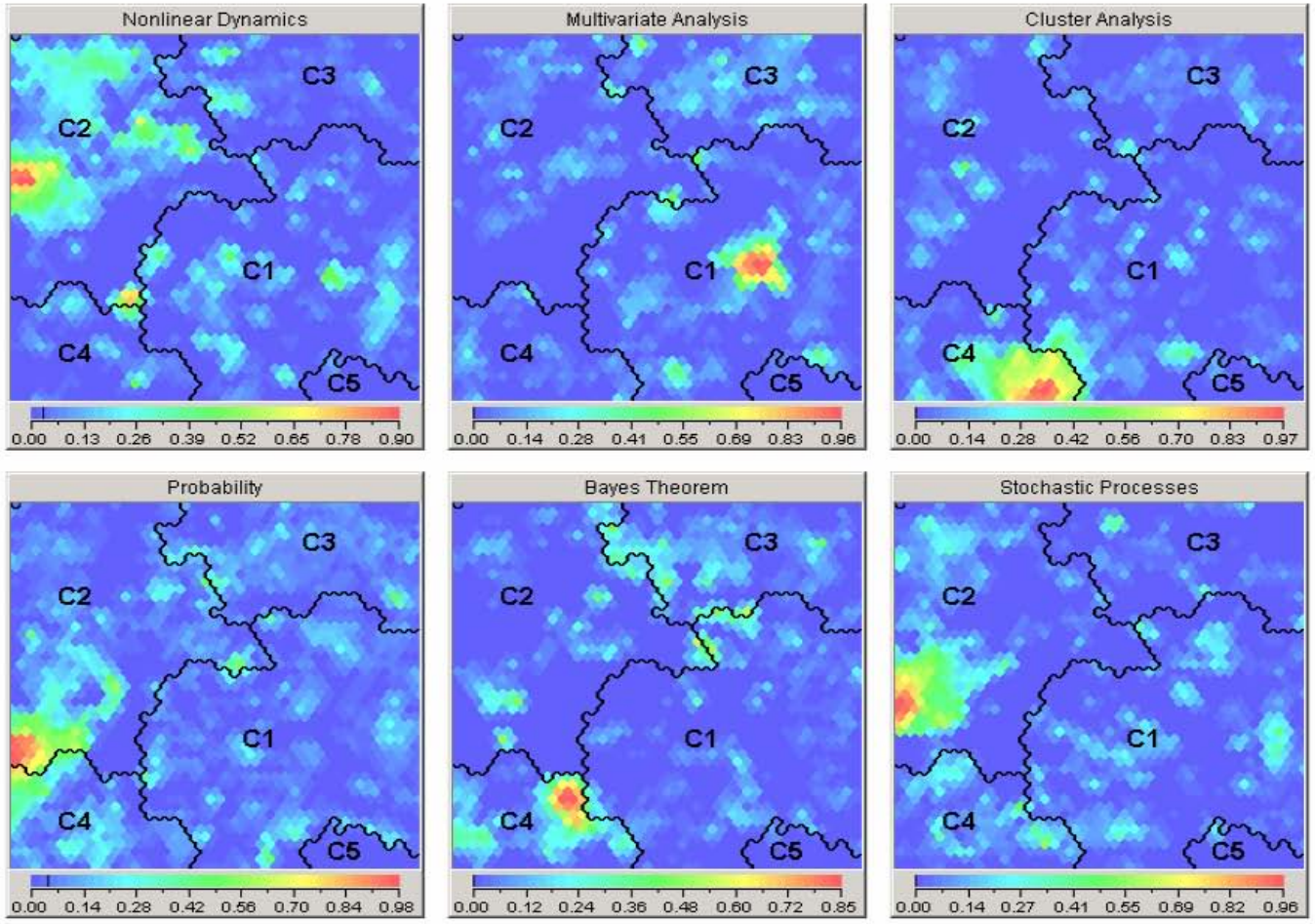


Figura V.0.15: Mapas de componentes bajo el criterio de Jaccard

Observe que cada region en el mapa está determinado por varios componentes.

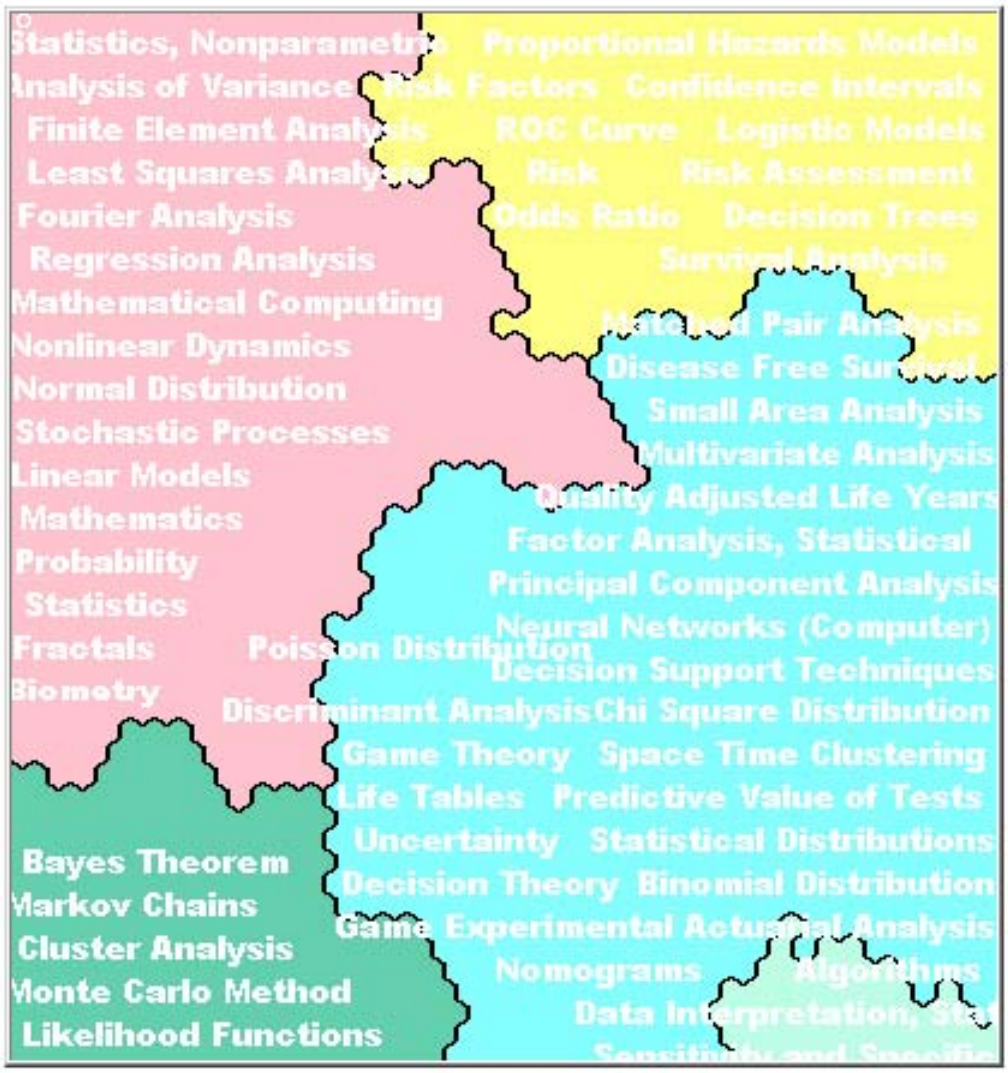


Figura V.0.16: Identificación de regiones bajo el criterio de Jaccard

B) Mapas Auto-Organizantes bajo el Criterio de Courtial

Siguiendo el orden de presentación de mapas que en la sección anterior. Se presentan los mapas auto - organizantes de regiones y componentes para el criterio de Courtial. El algoritmo SOM solamente obtuvo 3 regiones.

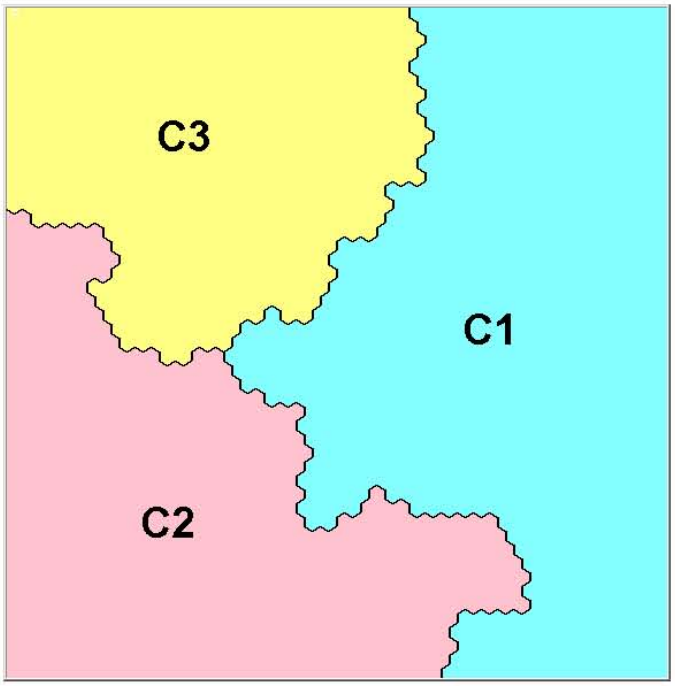


Figura V.0.17: Mapa auto – organizante que representa los conglomerados del conjunto de datos

Observe que el criterio de Courtial resalta más la influencia de las componentes Mathematics, Algorithms, Data Interpretation, Analysis, sobre las variable que el criterio de Jaccard.

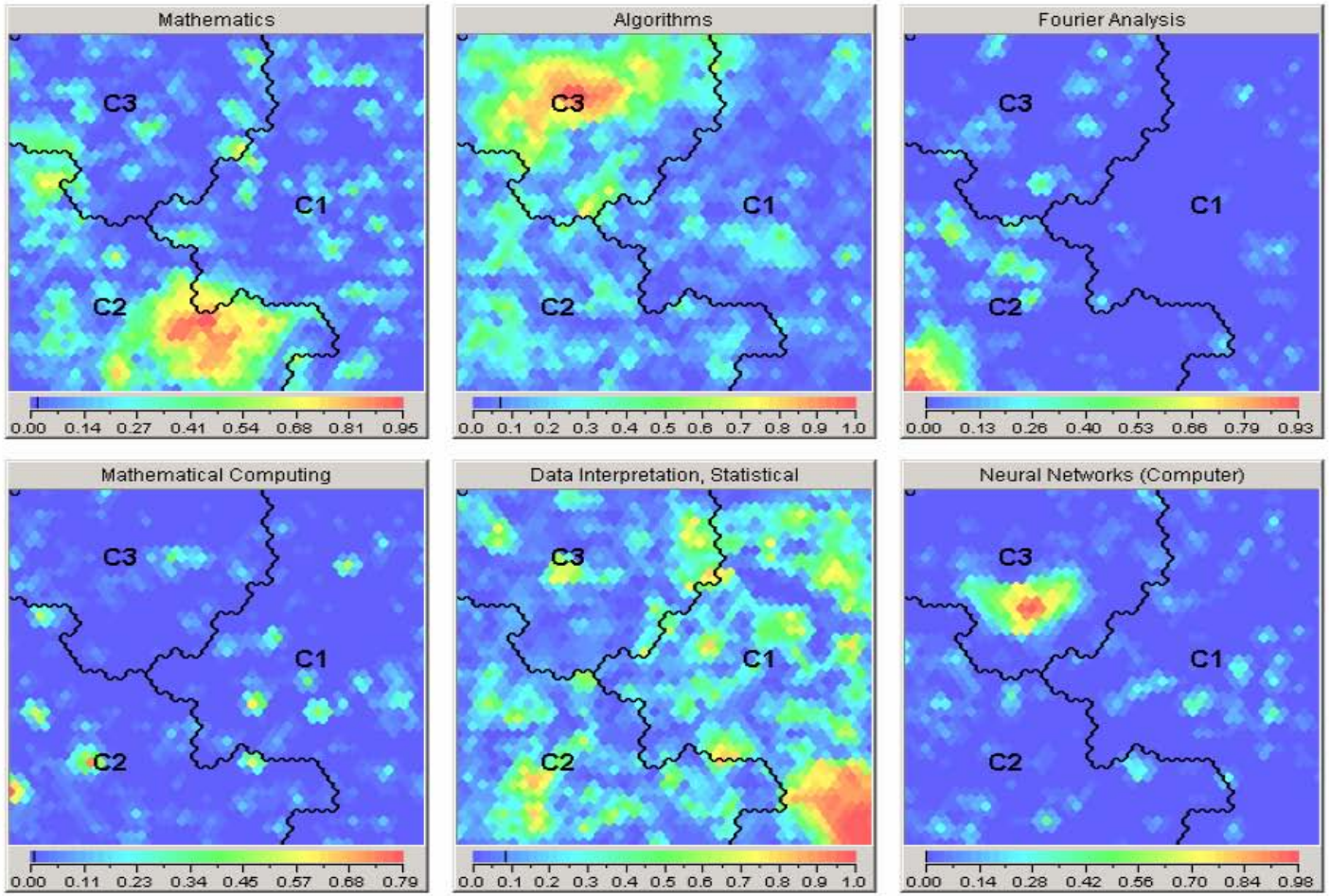


Figura V.0.18: Mapas de componentes bajo el criterio de Courtial

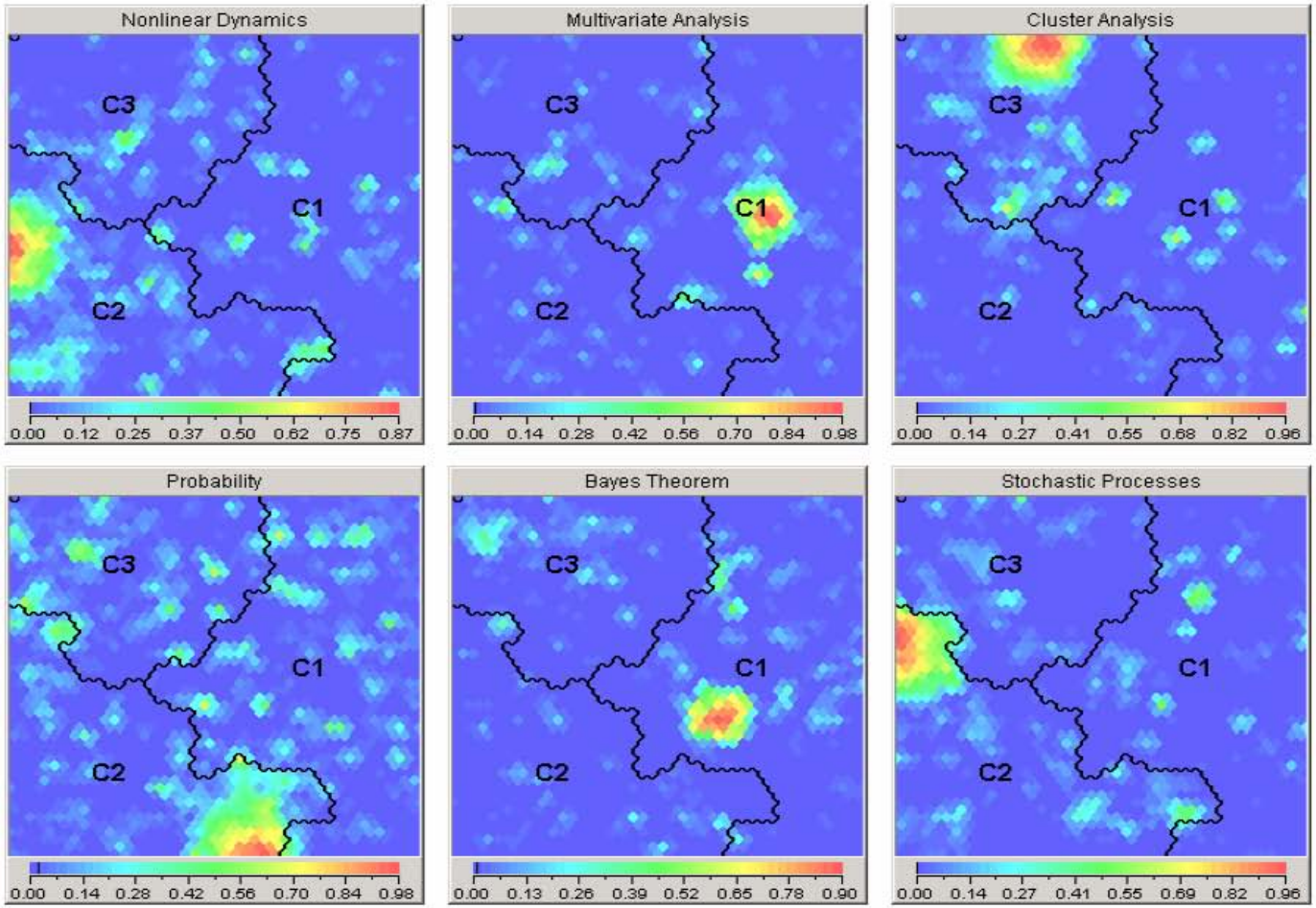


Figura V.0.18: Mapas de componentes bajo el criterio de Courtial

Figura V.0.19: Identificación de regiones bajo el criterio de Courtial

5.3.3 Mapas Auto - Organizantes del Almacén de Unión

Los mapas mostrados anteriormente corresponden a los almacenes de matemáticas e Intersección. A continuación se muestran los mapas auto-organizantes correspondientes a la unión de estos almacenes.

Datos generales

Los términos biológicos considerados (filas de la matriz de coocurrencia) pertenecen a la Categoría de Ciencias Biológicas *(Biological Sciences Category)*. Esta categoría esta integrada por las siguientes subcategorías.

001 Biological Sciences
002 Health Occupations
003 Environment and Public Health
004 Biological Phenomena, Cell Phenomena, and Immunity
005 Genetic Processes
006 Biochemical Phenomena, Metabolism and Nutrition
007 Physiological Processes
008 Reproductive and Urinary Physiology
009 Circulatory and Respiratory Physiology
010 Digestive, Oral, and Skin Physiology
011 Musculoskeletal, Neural, and Ocular Physiology
012 Chemical and Pharmacologic Phenomena
013 Genetic Phenomena
014 Genetic Structures

En el apéndice D se muestra una porción de esta categoría.

Los términos matemáticos considerados (columnas de la matriz de concurrencia) son los siguientes: *Mathematics; Algorithms; Finite Element Analysis; Fourier Analysis; Fractals; Game Theory; Games, Experimental; Mathematical Computing; Decisión Support Techniques; Decision Theory; Decision Trees; Nomograms; Neural Networks (Computer); Nonlinear Dynamics. Más las "temáticas" pertenecientes a Estadística.* Ver tabla 1.

Matriz de coocurrencia

Población de documentos: 167, 923
Tamaño de la matriz de coocurrencia: 1970 x 57

A) Mapas Auto - Organizantes bajo el Criterio de Jaccard

Este mapa muestra las 3 regiones que obtuvo el algoritmo de Kohonen.

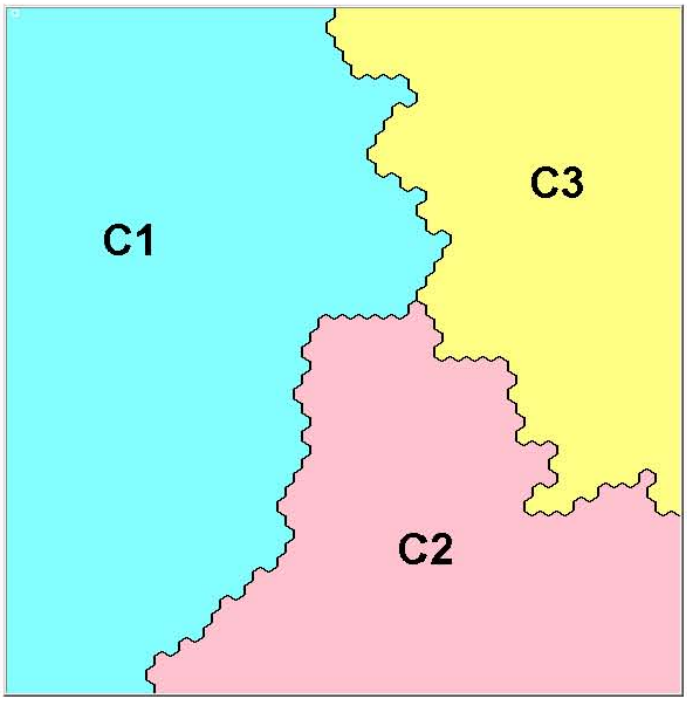


Figura V.0.20: Mapa auto – organizante que representa los conglomerados del conjunto de datos.

A continuación se presentan algunos mapas de componentes.

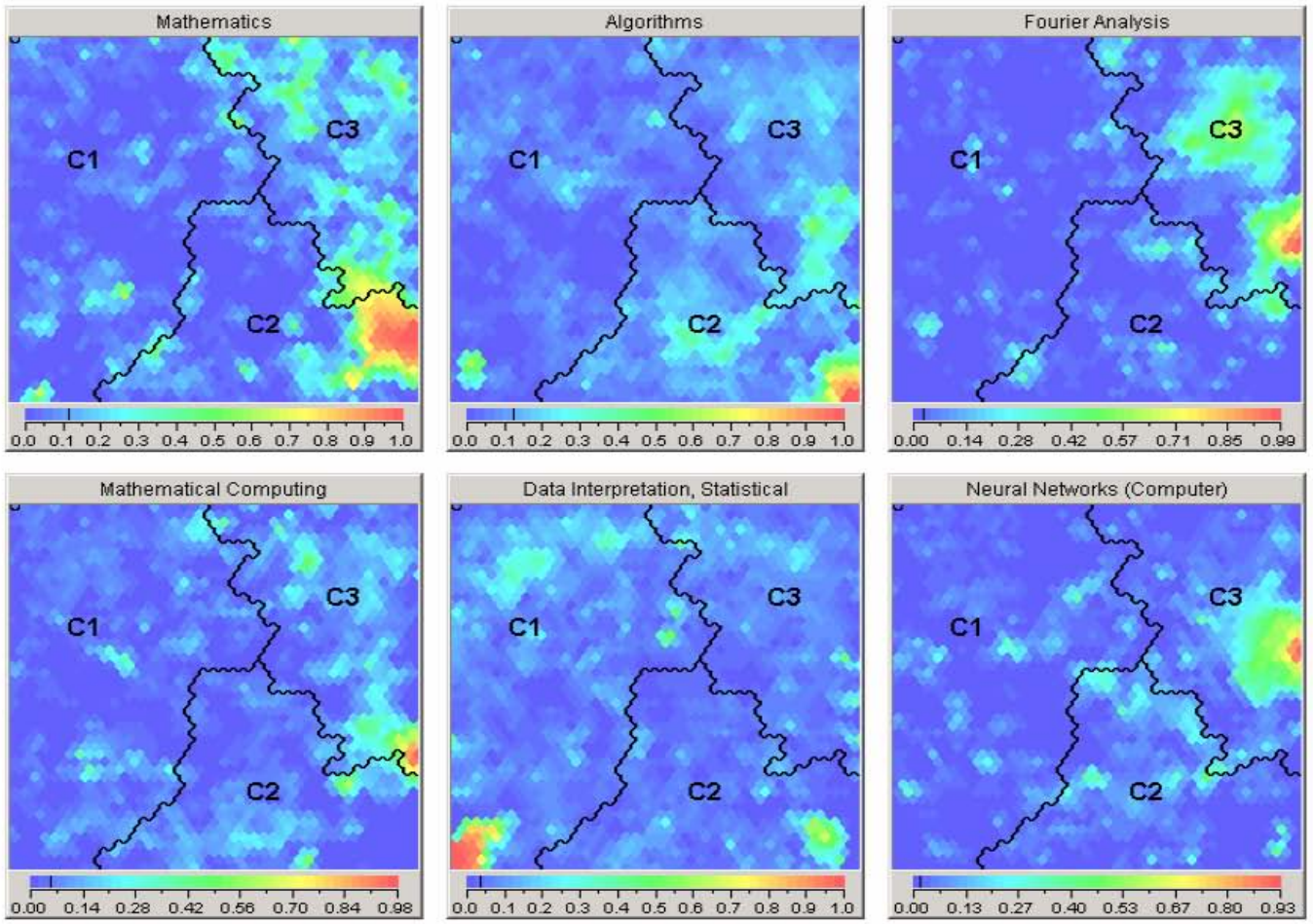


Figura V.0.21: Mapas de componentes bajo el criterio de Jaccard

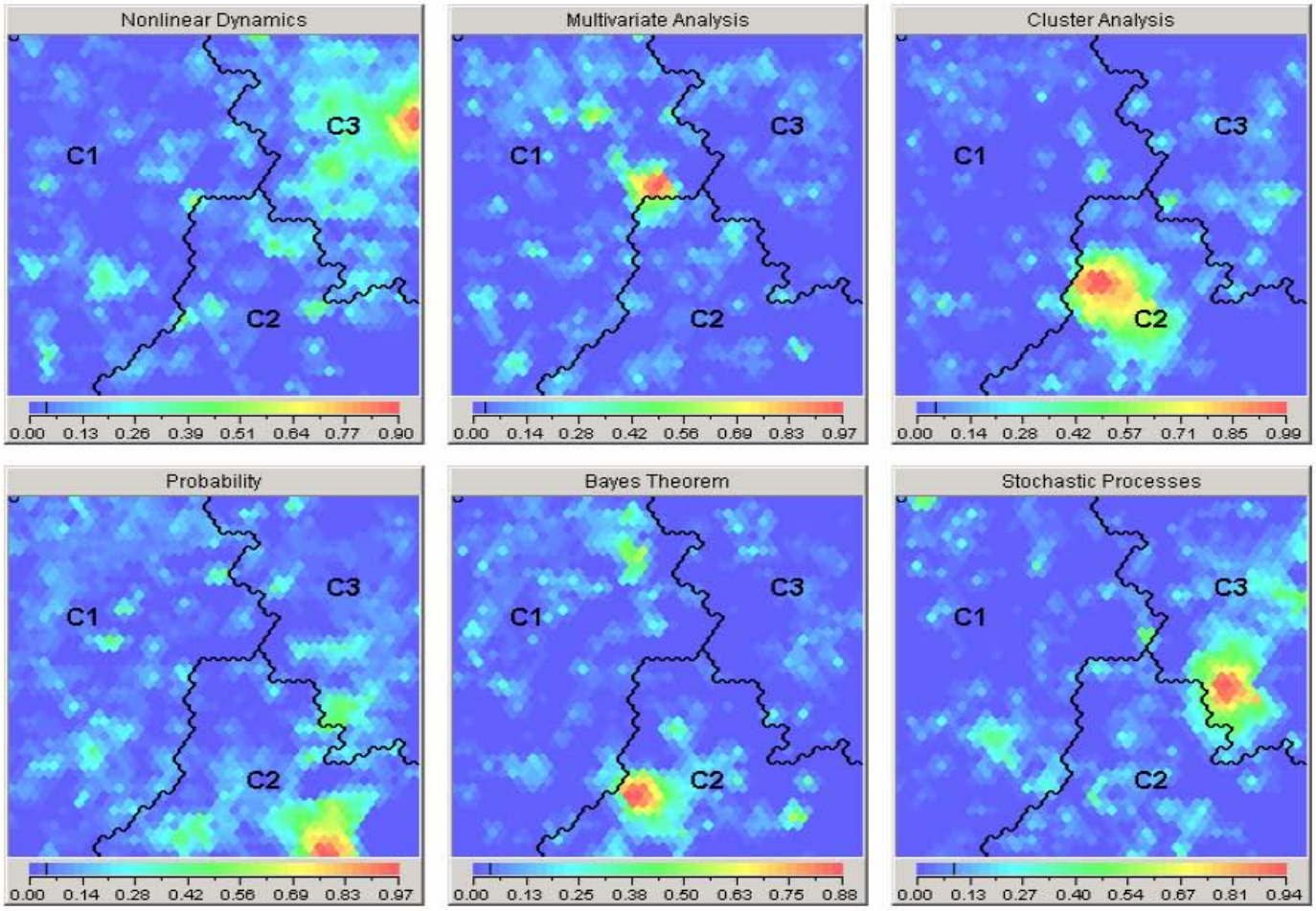


Figura V.0.17: Mapas de componentes bajo el criterio de Jaccard

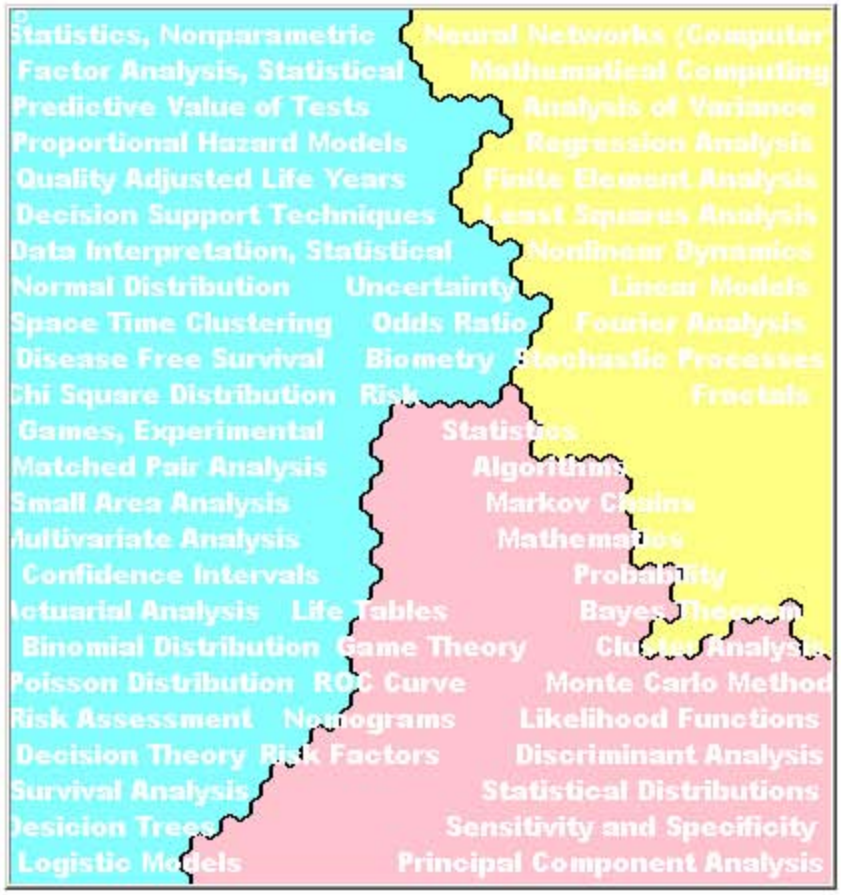


Figura V.0.22: Identificación de regiones bajo el criterio de Jaccard

B) Mapas Auto-Organizantes bajo el Criterio de Courtial

El algoritmo SOM solamente obtuvo 3 regiones.

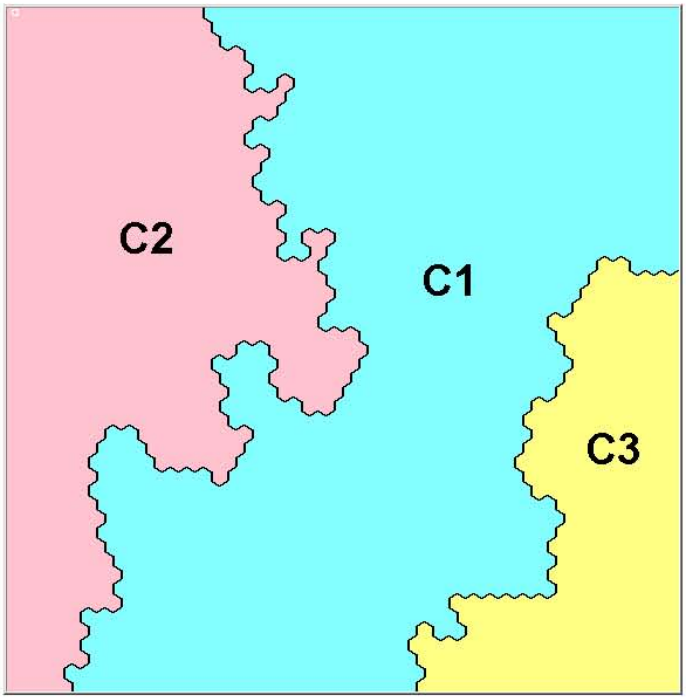


Figura V.0.23: Mapa auto – organizante que representa los conglomerados del conjunto de datos.

Se presentan algunos mapas de componentes. Observe como el criterio de Courtial resalta más las zonas de influencia de algunos componentes que el criterio de Jaccard.

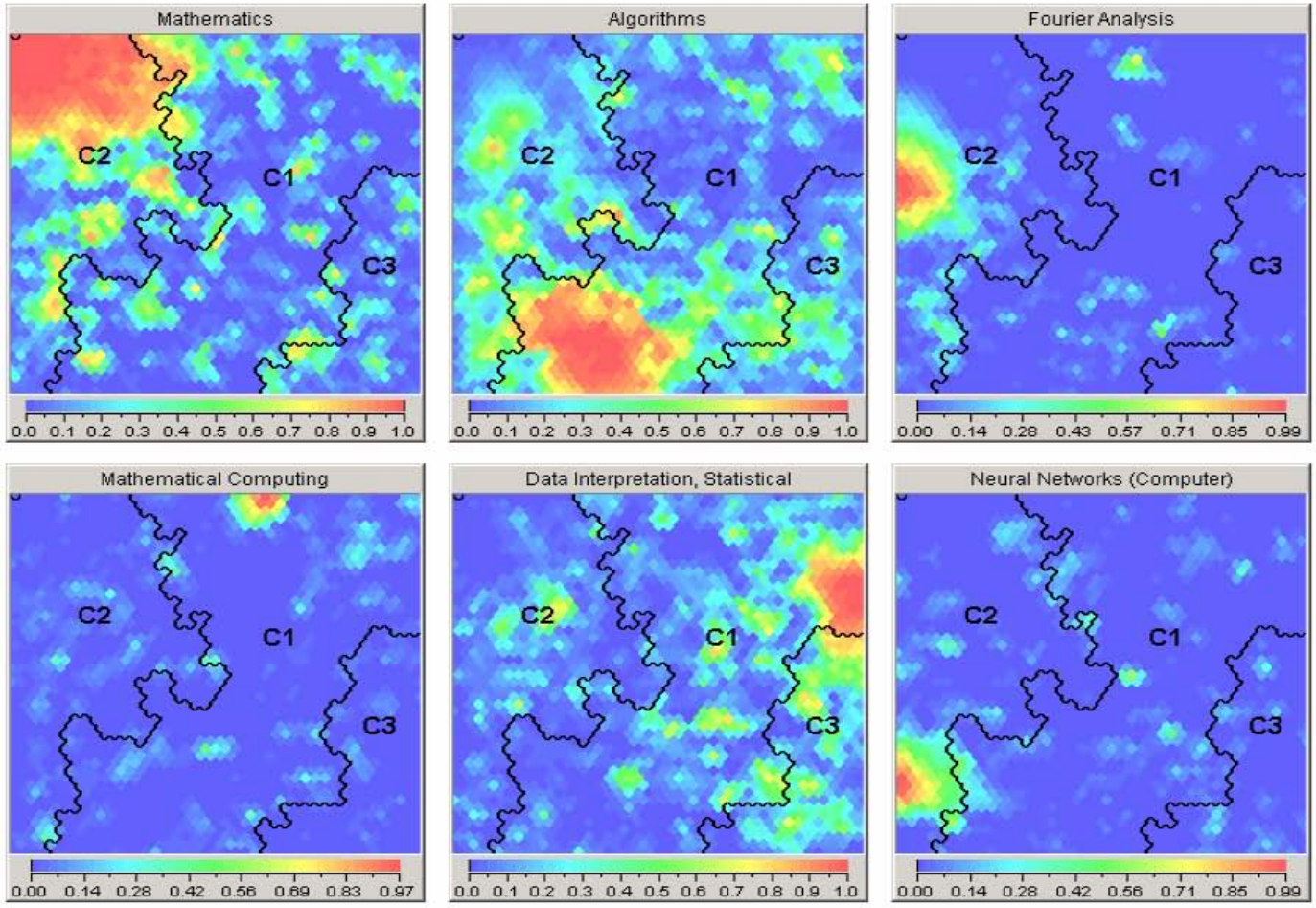


Figura V.0.24: Mapas de componentes bajo el criterio de Courtial

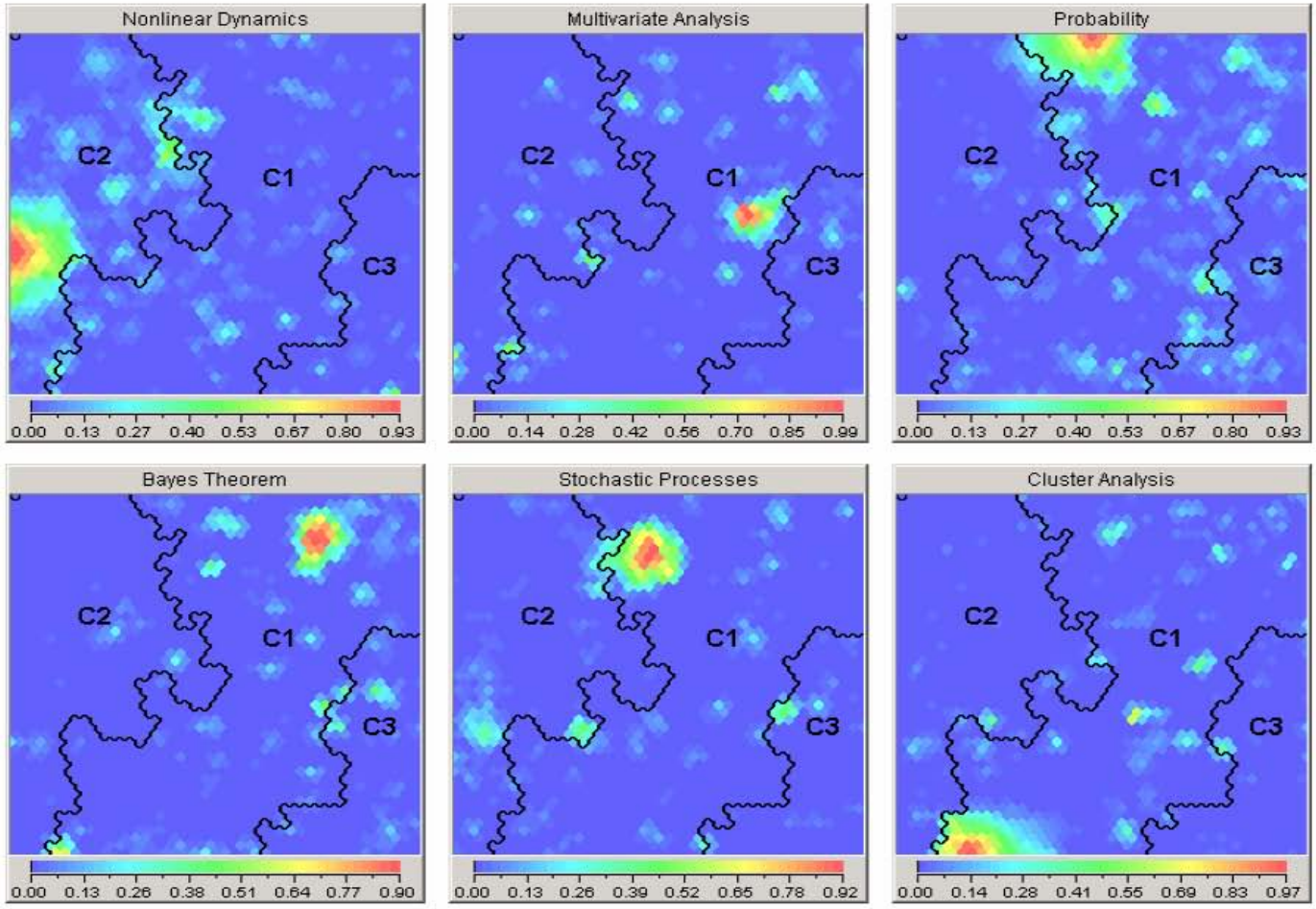


Figura V.0.24: Mapas de componentes bajo el criterio de Courtial

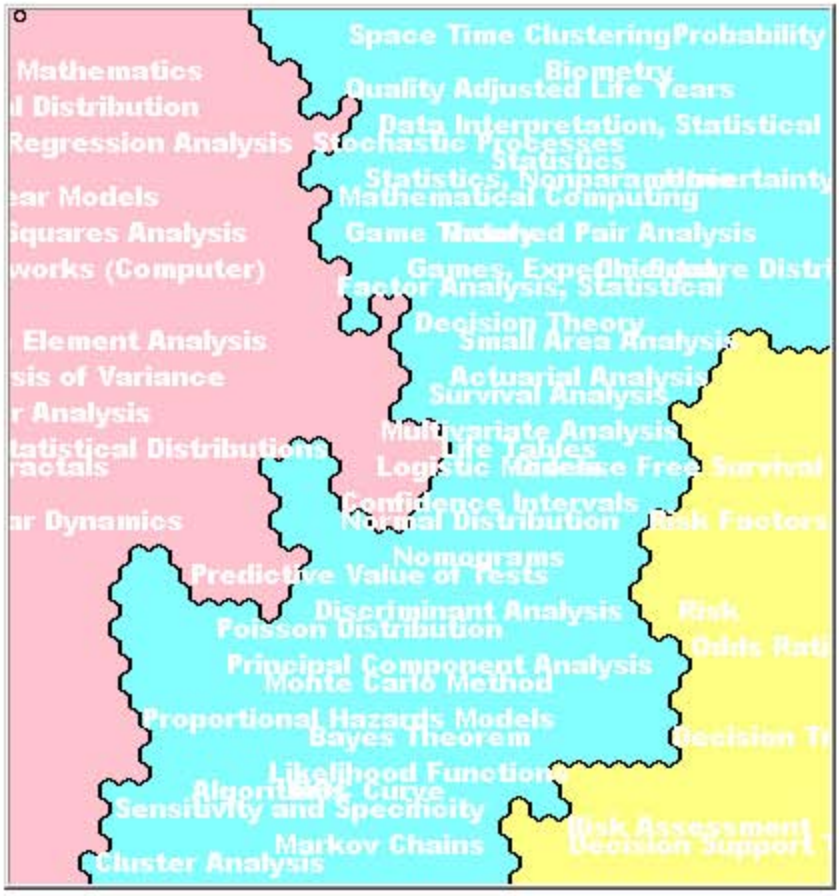


Figura V.0.25: Identificación de regiones bajo el criterio de Courtial

Se aprecia que los mapas auto – organizantes que se obtuvieron con el almacén de intersección y con el almacén de unión (este último es el resultado de haber unidos los almacenes de matemáticas e Intersección) presentan regiones que estan constituidas muy diferentes a las regiones que se obtuvieron con el almacén de matemáticas.

Los mapas de regiones de los almacenes intersección y unión estan integrados por varios núcleos, esto se interpreta así, las regiones representan conglomerados de conglomerados. Mientras que los mapas de regiones del almacén de matemáticas estan constituidos por un núcleo , es decir, cada region representa un conglomerado.

En los mapas de componentes se aprecian los núcleos y dispersiones que generan las variables en cada componente. Es claro que estos núcleos y dispersiones varian de componente en componente pues sencillamente las variables tienen comportamientos distintos según la componente.

Los criterios de Jaccard y de Courtial resaltan la forma y el tamaño -en forma distinta- de los núcleos y las dispersiones favoreciendo el analisis de información. Los nucleos y las dispersiones juegan un papel importante durante el análisis de información. En los núcleos se observa que variables dependen más de sus componentes. Las dispersiones detectan que variables estan correlacionadas a traves de sus componentes.

5.3.4 Propiedades generales

Las propiedades generales de los mapas pueden ser utilizadas durante el proceso de exploración y son significativas en casi cualquier aplicación. Estas propiedades son la preservación de la topología y la distribución de los datos en un despliegue ordenado; basándose en ellas es posible el establecimiento de relaciones entre variables, la visualización de "clusters" y la inspección de relaciones de vecindad entre los nodos en el mapa.

Los tres análisis bibliométricos se llevan a cabo con los mapas auto – organizantes del almacén de matemáticas.

- Distribución de Terminos.
- Recuperación de Información.
- Descubrimiento de Conocimiento.

Distribución de Términos

Para este sencillo ejemplo se emplean los términos MeSH de la subcategoría de Ciencias Biológicas (Biological Sciences, vea el apéndice D). Los términos se distribuyen en las regiones de los mapas auto – organizantes bajo los criterios de Jaccard y Courtial. La distribución se interpreta de la siguiente manera: el mapa clasifica los términos agrupando aquellos términos que tratan de la misma temática en la misma región.

Observe que existen regiones que solamente capturan un término, regiones que no capturan términos y regiones que capturan muchos términos. Además, los términos que se encuentran dentro del cuadro se mantuvieron "cercanos" bajo los dos criterios que aplicamos. Esto indica que existe una relación entre estos términos y ambos criterios lo resaltan.

El mapa de la izquierda muestra la distribución de los términos correspondiente a "Biological Sciences" bajo el criterio de Jaccard. Y mapa de la derecha muestra la distribución de los términos correspondiente a "Biological Sciences" bajo el criterio de Courtial. Además, observe que las regiones C3 y C13 no contienen términos.

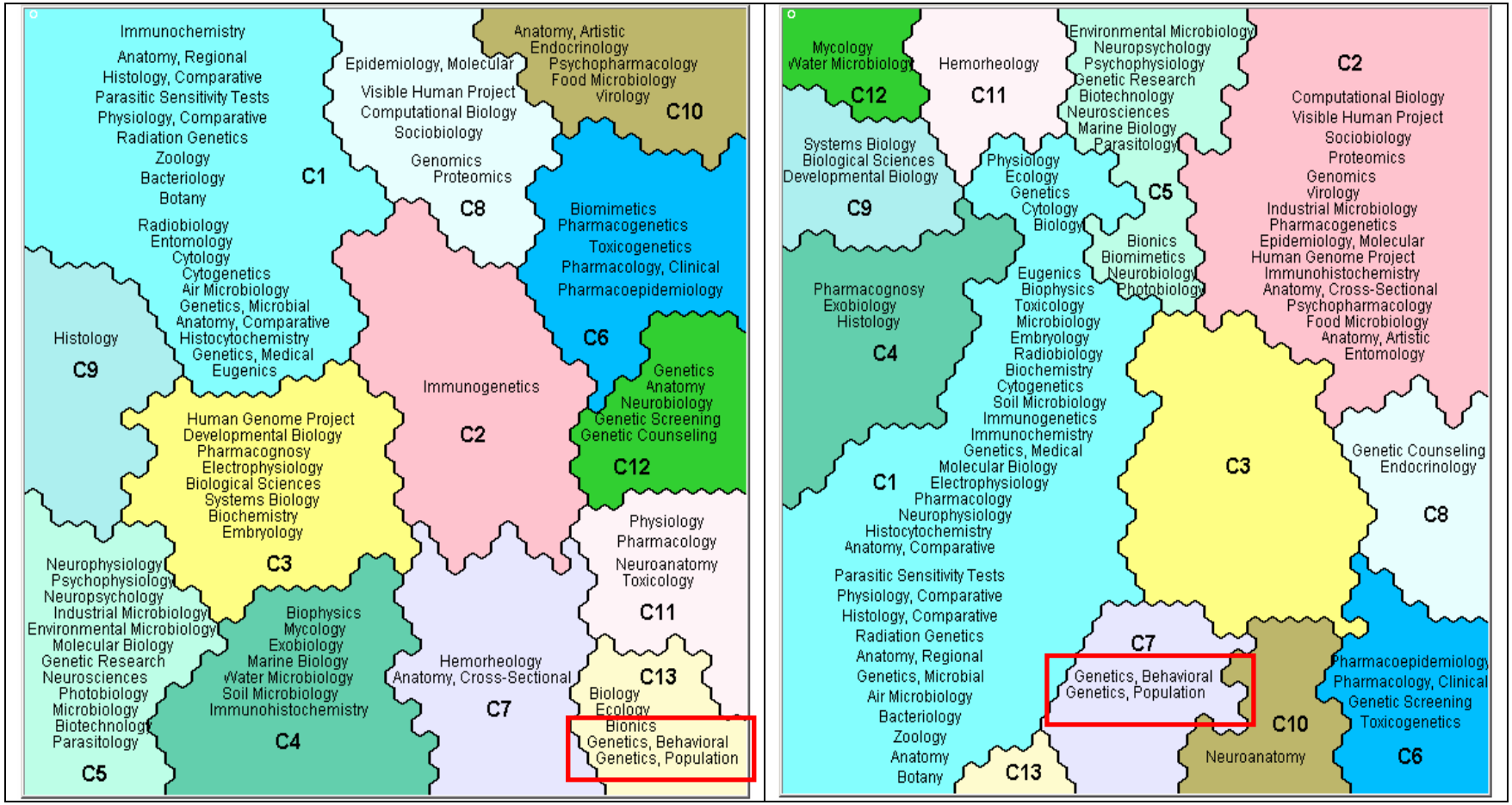


Figura V.0.26: Distribución de Términos

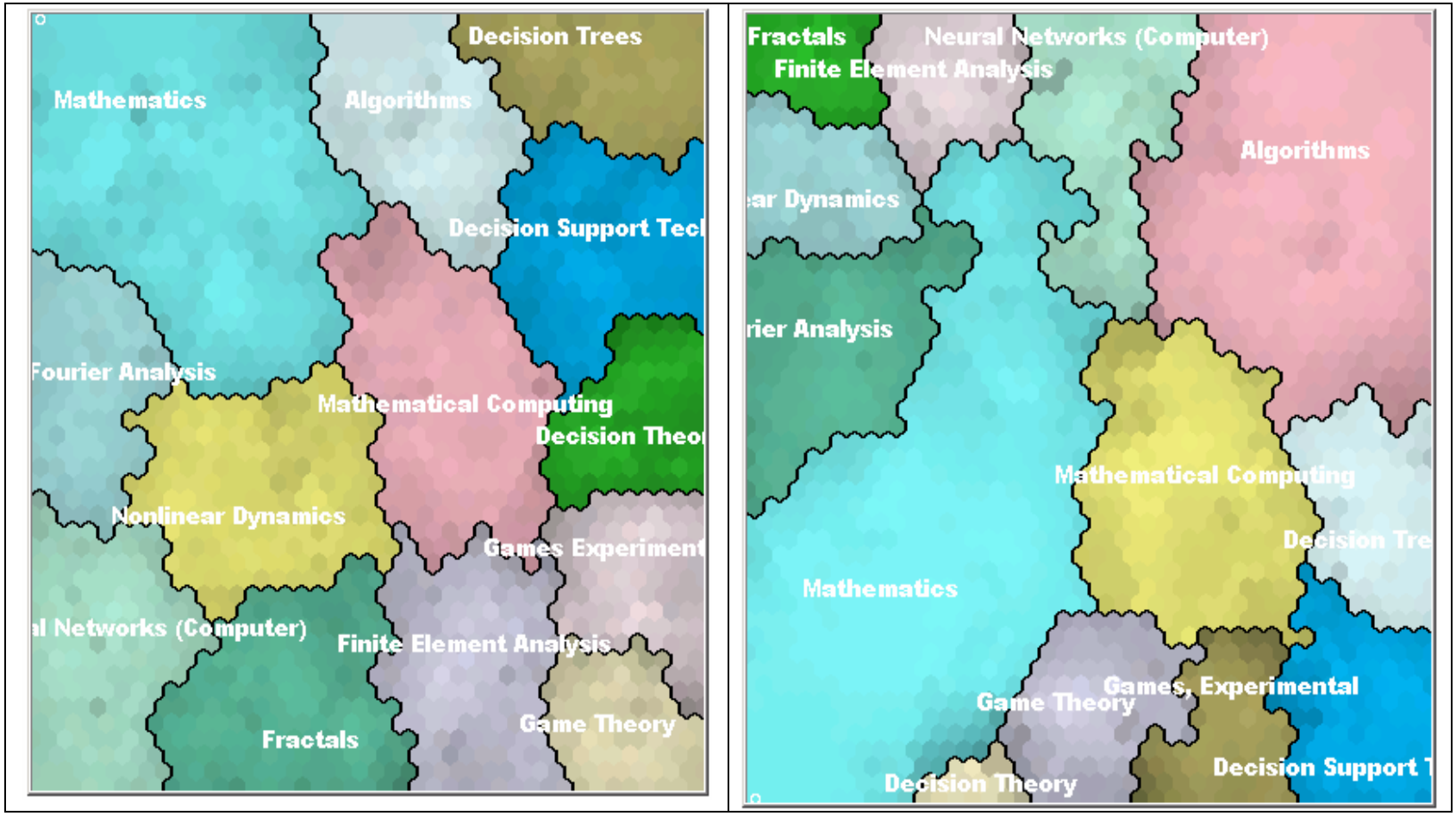


Figura V.0.27: Identificación de los mapas de regiones bajo Jaccard (izquierda) y Courtial (derecha)

Observe, por ejemplo, que la región C8 contiene los terminos Genomics y Proteomics (mapa bajo el criterio Jaccard). Ambos términos desde el punto de vista de la bioinformática se enfocan a estudiar aspectos distintos del genoma. Sin embargo, cuando se identifica que la region C8 corresponde a Algorithms se deduce que ambos términos utilizan los mismos algoritmos para cumplir sus objetivos.

En forma general, se interpreta que los términos MeSH de la subrama ciencias biologicas se distribuyen en el mapa de regiones de acuerdo a los distintos temas matemáticos que emplean.

Cuando se observa la distribución de los términos en el mapa bajo el criterio de Courtial se nota que difiere de la otra distribución. Esto representa un problema al analista pues sencillamente cual distribución escoge. Lo anterior depende de su experiencia.

Recuperación de Información

En este ejemplo se hace una sencilla recuperación de información. Las variables consideradas son las mismas del ejemplo anterior, es decir, los términos de la subcategoría de Ciencias Biológicas (Biological Sciences). Y únicamente se seleccionaron los mapas auto-organizantes de las componentes "Neural Networks (Computer)" y "Nonlinear Dynamics" obtenidos bajo el criterio de Jaccard.

En ambos mapas los términos resaltados en negro se encuentran en el núcleo, es decir, en la zona en donde los términos (o las variables) dependen más de las componentes. La recuperación conciste en mostrar todas las citas que contengan algún térmno del núcleo.

El ejemplo es similar a lo siguiente. Cuando se busca información sobre algún tema, generalmente se acude a la biblioteca. En la biblioteca se escribe el término o la frase a buscar en la interfaz de búsqueda que proporciona la biblioteca. Entonces, los resultados se muestran en una nueva ventana. Estos resultados contienen citas de libros, revistas, etc. Y para concluir la busquedad, se seleccionan aquellas citas más idóneas del tema.

El mapa de la izquierda muestra la distribución de los términos correspondiente a "Biological Sciences" en la componente "Nonlinear Dynamics" bajo el criterio de Jaccard. Y el mapa de la derecha muestra la distribución de los términos correspondiente a "Biological Sciences" en la componente "Neural Networks (Computer)" bajo el criterio de Jaccard.

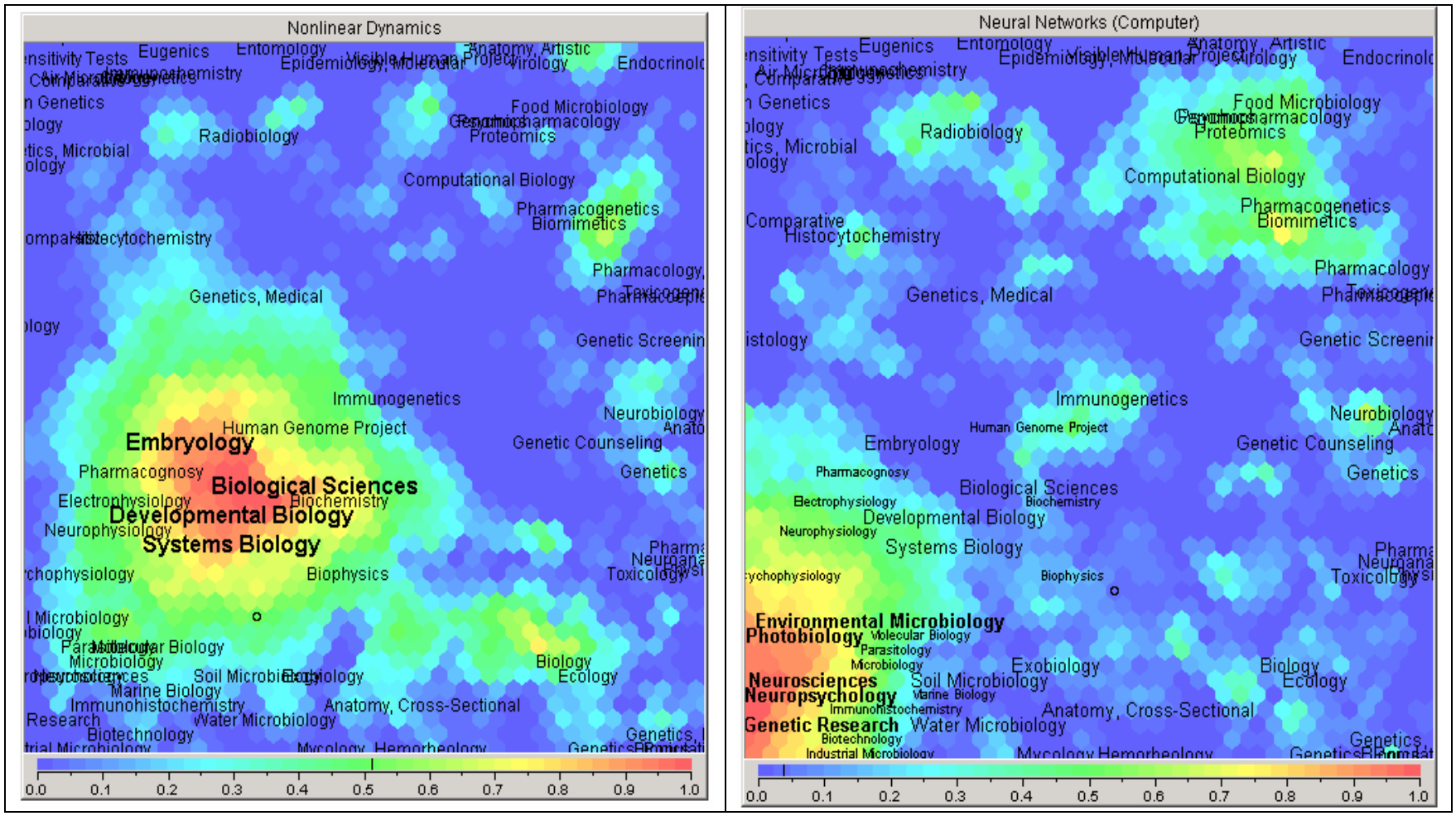


Figura V.0.28: Recuperación de Información

En la siguiente tabla se muestra el número de citas que poseen alguno de los términos que se encuentran en los núcleos correspondientes de “Neural Networks (Computer)” y “Nonlinear Dynamics”. Por ejemplo, hay 25 citas en el núcleo de “Neural Networks (Computer)” que poseen el término *Parasitology*.

Términos	**Número de citas en el núcleo de:**	
	Neural Networks (Computer)	**Nonlinear Dynamics**
Parasitology	25	
Neurosciences	19	
Neuropsychology	12	
Photobiology	3	
Environmental Microbiology	2	
Genetic Research	1	
Embryology		37
Biological Sciences		5
Developmental Biology		5
Systems Biology		5

Tabla V.0.10: Número de citas en los núcleos

A continuación se muestra una cita de un documento que trata sobre *Embryology* y se encuentra en el núcleo de “Nonlinear Dynamics”. Observe el año de publicación de la cita.

Autores: Zhang, Y.; Weng, J.; Hwang, W. S.
Titulo: Auditory learning: a developmental method
Revista: IEEE Trans Neural Netw
Año de publicación: 2005 May

En el apéndice E se presentan las citas de los documentos que tratan de los temas que se encuentran en los núcleos de “Nonlinear Dynamics” y “Neural Networks (Computer)”.

Por lo general, las Redes Neuronales estan íntimamente relacionadas con la Neurociencia mientras que la Biología modela muchos aspectos biológicos a través de Sistemas no Lineales. Sin embargo, en las citas de los documentos que se recuperaron se aprecia que el campo de la Parasitología recurre más a las redes neuronales que la propia Neurociencia. Por otra parte, se aprecia que de todas las áreas de la Biología, la Embriología es la que más recurre a la dinámica no lineal.

Descubrimiento de Conocimiento

Se presenta un ejemplo en donde se analiza el resumen (en el formato MedLine el resumen se identifica con la etiqueta Abstract (43)) de algunos documentos. El análisis tiene como objetivo encontrar información de la cual se derive conocimiento.

Las variables que consideraremos son las mismas de los ejemplos anteriores, es decir, los términos MeSH de la subcategoría de Ciencias Biológicas (Biological Sciences). Nuevamente seleccionamos los mapas auto-organizantes de las componentes "Neural Networks (Computer)" y "Nonlinear Dynamics", obtenidos con el método de conglomerados "Som Ward Clusters" bajo el criterio de Jaccard.

En el ejemplo anterior se identificaron los términos que están en los núcleos de "Neural Networks (Computer)" y "Nonlinear Dynamics". Ahora se van a identificar los términos que se encuentran en la intersección de las dispersiones de "Neural Networks (Computer)" y "Nonlinear Dynamics".

En las figuras siguientes se muestran los términos que se encuentran en la intersección de las dispersiones (letra negra).

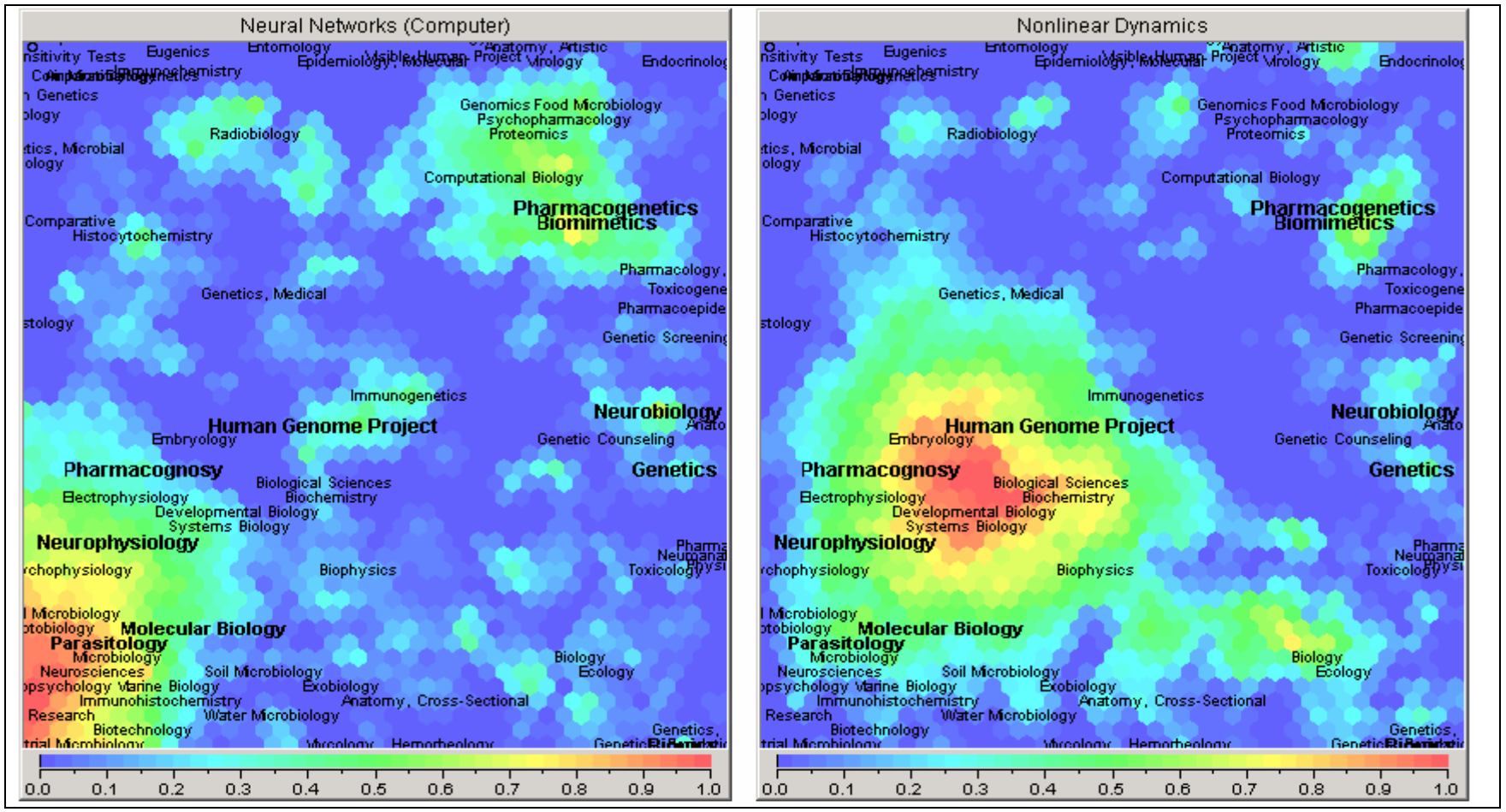


Figura V.0.29: Descubrimiento de Conocimiento

Los términos en la intersección son los siguientes: *Pharmacogenetics, Biomimetics, Human Genome Project, Pharmacognosy, Neurophysiology, Neurobiology, Genetics, Molecular Biology, Parasitology.* Estos términos se localizan en el formato MedLine en la etiqueta Keywords (45). Vea figura V.0.23. En la tabla V.0.11 se muestra el número de documentos que poseen "Neural Networks (Computer)" y alguno de los términos en la intersección. Similarmente, el número de documentos que poseen "Nonlinear Dynamics" y alguno de los términos en la intersección. Por ejemplo, hay un documento que posee "Neural Networks (Computer)" y *Pharmacogenetics* en la etiqueta Keywords (45).

Términos	**Número de documentos en la dispersión de:**	
	Neural Networks (Computer)	**Nonlinear Dynamics**
Pharmacogenetics	1	1
Biomimetics	13	2
Human Genome Project	4	2
Pharmacognosy	2	1
Neurophysiology	10	5
Neurobiology	11	3
Genetics	10	3
Molecular Biology	16	7
Parasitology	3	1

Tabla V.0.11: Número de documentos en las dispersiones

De todos los términos en la intersección solamente se trabaja con *Biomimetics*. El análisis puede hacerse de manera semejante para cualquier otro término. A continuación se presenta una ficha en el formato MedLine[82] que contiene los terminos *Biomimetics* y *Neural Networks (Computer)* en la etiqueta Keywords (45). Observe que se han resaltado con negritas algunos términos o frases del resumen que hacen referencia a la dinámica no lineal.

En el apéndice F se muestran todas las fichas en el formato MedLine. Observe los términos y frases que se han resaltado del resumen.

[82] Las demás citas estan en el apéndice F

Author, Analytic (01): Mtetwa, N. //Smith, L. S.

Article Title (04): Precision constrained stochastic resonance in a feedforward neural network

Medium Designator (05): A_Biomimetics_NNC

Connective Phrase (06):

Journal Title (10): IEEE Trans Neural Netw

Translated Title (11):

Reprint Status (12): **Date:**

Date of Publication (20): 2005 Jan

Volume ID (22): 16

Issue ID (24): 1

Page(s) (25): 250-62

Language (35): eng

Connective Phrase (36):

Address/Availability (37): Department of Computing Science, University of Stirling, Stirling FK9 4LA, UK. nmt@cs.stir.ac.uk

Location/URL (38): 15732404

ISSN (40): 1045-9227

Notes (42):

Abstract (43): Stochastic resonance (SR) is a phenomenon in which the response of a **nonlinear system** to a subthreshold information-bearing signal is optimized by the presence of noise. By considering a nonlinear system (network of leaky integrate-and-fire (LIF) neurons) that captures the **functional dynamics of neuronal firing**, we demonstrate that sensory neurons could, in principle harness SR to optimize the detection and transmission of weak stimuli. We have previously characterized this effect by use of signal-to-noise ratio (SNR). Here in addition to SNR, we apply an entropy-based measure (Fisher information) and compare the two measures of quantifying SR. We also discuss the performance of these two SR measures in a full precision floating point model simulated in Java and in a precision limited integer model simulated on a field programmable gate array (FPGA). We report in this study that stochastic resonance which is mainly associated with floating point implementations is possible in both a single LIF neuron and a network of LIF neurons implemented on lower resolution integer based digital hardware. We also report that such a network can improve the SNR and Fisher information of the output over a single LIF neuron.

Call Number (44):

Keywords (45): Action Potentials: :*physiology/Animals/**Biomimetics: :*methods**/Comparative Study/Computer Simulation/Feedback/Humans/*Models, Neurological/Models, Statistical/*Nerve Net/***Neural Networks (Computer)**/Neurons: :*physiology/Research Support, Non-U.S. Gov't/Stochastic Processes/Synaptic Transmission: :*physiology

Figura V.0.30: Ficha 1 en formato Medline

Se puede argumentar que en estas citas: la biomimetica utiliza a las redes neuronales para el desarrollar de materiales. Los materiales que desarrolla la Biomimetica tienen la característica de utilizar procesos naturales que se encuentran en sistemas biológicos. Los materiales que desarrolla la Biomimetica no son biomateriales, pues estos últimos, son materiales naturales o sintéticos, con excepción de las drogas, que se usan para reparar o reemplazar algunos tejidos del cuerpo.

Como una segunda etapa se analiza el resumen (Abstract). El análisis que se hace consiste en identificar términos o frases que hagan referencia a la dinámica no lineal (se pudo haber escogido algún otro tema pero en este caso nos interesa la dinámica no lineal).

Algunos temas o frases que se identificaron en el resumen de las citas son: **nonlinear dynamics, Lyapunov stability theorem, fixed point, limit cycle, etc**. Todos estos temas son de interés en la dinámica no lineal.

En resumen, se deduce que la biomimetica al simular procesos que se desarrollan en sistemas biológicos se enfrenta a problemas conocidos en los sistemas dinámicos y en los modelos neuronales.

El hecho importante en este ejemplo, consistió en dividir las citas en dos. Las citas que contienen “Biomimetics” y “Neural Networks (Computer) por un lado y las citas que contienen “Biomimetics” y “Nonlinear Dynamics” por otro lado. Esto garantiza lo siguiente: el conjunto de descriptores proveniente de las citas que contienen los términos “Biomimetics” y “Neural Networks (Computer)” generan una matriz de coocurrencia. En esta matriz de coocurrencia el término “Nonlinear Dynamics” no está, es decir, no hay ningún vector en la matriz de coocurrencia que lo represente. Además, se debe considerar que los términos “Biomimetics” y “Neural Networks (Computer)” tienen vectores de coocurrencia similares (por la forma en que se construyo la matriz).

Cuando se realiza el mapeo bibliométrico los términos “Biomimetics” y “Neural Networks (Computer)” deben estar contenidos en la misma región del mapa. Pero ninguna región del mapa contiene el término “Nonlinear Dynamics”. Lo anterior se puede interpretar asi, en las citas seleccionadas la Biomimética hace uso de las redes neuronales para desarrollar materiales biomiméticos, pero sin embargo cuando se revisan los resúmenes de las citas detectamos frases o conceptos desarrollados y estudiados dentro del marco de la dinámica no lineal.

Para complementar el ejemplo, se tendría que hacer el mismo análisis para las citas que contienen “Biomimetics” y “Nonlinear Dynamics” pero ahora identificando términos o frases que hacen referencia a las redes neuronales.

Conclusiones

El ViBlioSOM como metodología para la extracción de conocimiento en bases de datos científicas – tecnológicas promete dar resultados confiables a los analistas de información. Las etapas que integran al ViBlioSOM:

- Adquisición y Selección de Ficheros.
- Preprocesamiento.
- Minería de Datos y Textos a partir de Indicadores Bibliométricos.
- Visualización e Interpretación de los Resultados.

Pretenden ser flexibles pero al mismo tiempo rigurosas durante los análisis bibliométricos. No hay que olvidar que en este tipo de metodologías es recomendable tener un equipo de profesionales -si es posible- en cada una de las etapas. De esta forma los resultados obtenidos seran confiables.

Por otra parte, la implementación computacional del ViBlioSOM requiere poder de computo del alto nivel (si es posibles clusters de computadoras) pues sencillamente el almacenamiento de unos cuantos formatos de MedLine requieren dos o tres gigabytes de memoria en disco duro y su procesamiento requiere ser eficiente.

Además, realizar análisis bibliométricos usando el MeSH Vocabulary requiere que la implementación computacional del ViBlioSOM sea interactiva e iterativa con el analista, que permita seleccionar varias ramas del MESH, combinarlas, etc.

Otro aspecto importante del ViBlioSOM es el uso de criterios de normalización. Mediante estos criterios -el criterio de Jaccard y el criterio de Courtial- el análisis de palabras asociadas es capaz de discernir qué palabras clave y qué asociaciones son realmente relevantes para el análisis bibliométrico y por su puesto eliminar aquellas que por su baja coocurrencia relativa o su elevada generalidad no lo son.

Los análisis bibliométricos presentados en esta tesis dejan ver las posibiliades que tiene el ViBlioSOM para la extracción de conocimiento:

En la *distribución de terminos*: a traves de este ejercicio se pretende ver al ViBlioSOM como una herramienta que permita construir tesaurus a partir de una colección de términos. El problema de este ejercicio estriba en que

los términos de la colección tienen distintas jerarquias y esto no se ve reflejado en las regiones de los mapas auto – organizantes. Es por ello, que se desea implementar en el ViBlioSOM un SOM jerarquico.

En la *recuperacion de información*: siempre que se va a hacer una nueva investigación sobre algún tema es fundamental realizar una revisión bibliográfica. El ViBlioSOM pretende ser una herramienta de recuperación de información que le proporcione al investigador todas las citas relacionadas con algún tema. Por el momento, las citas serán de documentos de MedLine.

Durante la revisión de las citas se detectó que muchas de estas relacionan a la parasitología con las redes neuronales. Por otra parte, se detectó que de todas las áreas de la biología, la embriología es la que más recurre a la dinámica no lineal. Además, muchas de estas citas tienen fecha de publicación muy reciente.

En el *descubrimiento de conocimiento*: se sabe que las palabras clave en los documentos *"sintetizan"* el contenido de este. El análisis de las palabras clave se puede ver como un proceso de exploración. Sin embargo, el investigador que desea profundizar en el contenido del documento se dirige al resumen. El ViBlioSOM pretende analizar el resumen en busca de información adicional que complemente la información que proporcionan las palabras clave. Analizar el resumen es complicado ya que requiere conocimientos de ingenieria lingüística, tesaurus, etc.

Durante el ejercicio, se dedujo que la biomimetica al simular procesos que se desarrollan en sistemas biológicos se enfrenta a problemas conocidos en los sistemas dinámicos y en los modelos neuronales.

Con estos tres análisis bibliométricos se pretende que el analista observe el comportamiento de los términos MeSH pertenecientes a la subcategoría de ciencias biológicas desde la perspectiva de la subcategoría de Matemáticas. Dicho comportamiento está reflejado en la forma en que algunos de las campos de las matemáticas (sistemas dinámicos, procesos estocásticos, estadística multivariada, etc.) son utilizados y en que forma en la investigación en el área biomédica (genética, farmacología, enfermería clínica, etc.). De esta manera se puede formar un marco de referencia sobre el uso de los terminos de ciencias biologicas:

- ¿Han surgido nuevas áreas de interés dependiendo de las matemáticas en uso? ¿Qué originó estos cambios?

- ¿Quiénes trabajan en el área? Laboratorios, universidades, institutos, compañías, etc.

- ¿Quiénes han abandonado el área? ¿Por qué?

Por su puesto, que la limitante de estos analisis bibliometricos estriba en el volumen de documentos recuperados; se necesitan herramientas que permitan analizar redes de autores, de instituciones; el desarrollo de tesasurus para el analisis del resumen; contar con personal capacitado en el área objeto de estudio; y no olvidar que el conocimiento obtenido representa casos particulares y es valido si y sólo si es entendible.

Apéndice

Apéndice A

Glosario

El Análisis de la Información puede definirse como la aplicación de técnicas de procesamiento automático del lenguaje natural, de clasificación automática y de representación gráfica (cartografía) del contenido cognitivo (conocimientos) y factual (fecha, lengua, tipo de publicación,...) de los datos bibliográficos (o textuales). Esta definición corresponde al análisis asistido por computadora. En general, por análisis de la información se entiende la fase de interpretación que el usuario realiza de una manera directa y manual. Los límites de este tipo de análisis son evidentes desde el momento que se trabaja sobre una cantidad importante de datos y se trata además de incorporar el análisis en un sistema de producción de información elaborada o especializada.

Red Neuronal: tipo de inteligencia artificial que intenta imitar la manera de trabajar del cerebro. Es más que un modelo digital; una red neuronal trabaja creando conexiones entre *elementos de procesamiento* (equivalencia computacional de las neuronas). La organización y los respectivos pesos de las conexiones determinan las salidas (output).

Inteligencia Artificial: rama de las ciencias de la computación dedicada a inventar computadoras que se comporten de manera análoga al cerebro de los seres humanos. El término lo acuñó en 1956 John McCarthy del Instituto de Tecnología de Massachusetts. La inteligencia artificial incluye:

- *Jugar juegos:* programas de computadora para jugar juegos tales como el ajedrez.
- *Sistemas expertos:* programas de computadora para la toma de decisiones en situaciones de la vida real (por ejemplo, algunos sistemas expertos ayudan a los doctores a diagnosticar enfermedades basándose en síntomas).
- *Lenguaje natural:* programas de computadora para entender la naturaleza de los lenguajes humanos.
- *Redes neuronales:* Sistemas que simulan inteligencia reproduciendo los tipos de conexiones físicas que ocurren en el cerebro.
- *Robótica:* programas de computadora para ver oír y reaccionar a otros estímulos sensoriales.

Modelo: Un término general para describir una representación conceptual de algún fenómeno que típicamente se compone de términos simbólicos, factores o estructuras que pueden ofrecerse en lenguaje, imágenes o notación matemática.

La Revisión por parte de Colegas o Pares *(Peer Review):* Formalmente el proceso de revisión por pares del trabajo científico fue iniciado en 1753 por la “Royal Society of London”, sin embargo, este proceso se originó con el surgimiento de las primeras revistas científicas, “The Journal des Scavans” (Francia) y “The Philosophical Transactions of the Royal Society” (Inglaterra), en enero y marzo de 1665, respectivamente. El sistema de evaluación del trabajo científico por los miembros de la comunidad llamado de revisión por pares (“peer

review") o sistema de arbitraje ("referee system"), es un proceso que se inicia cuando un científico somete su trabajo en forma de artículo al editor de una revista para ser publicado; éste selecciona algunos especialistas (árbitros) quienes evalúan la calidad del trabajo y definen si el producto de la investigación realizada por el científico tiene potencial para ese propósito; o si se debe hacer algún trabajo adicional antes de ser publicado o si simplemente no amerita su publicación .

MainStream Sciences (Corriente Principal de la Ciencia): Toda aquella literatura que es generada bajo la Revisión por Pares.

Complejidad Computacional: La Teoría de la Complejidad estudia la manera de clasificar algoritmos como buenos o malos y de clasificar problemas de acuerdo a la dificultad inherente de resolverlos. Algunas preguntas interesantes en este campo son: ¿Para todos los problemas, existe al menos un algoritmo? Si existen varios algoritmos para un problema, ¿Cómo hacer una selección en términos de su eficiencia?, ¿Cómo pueden ser clasificados los problemas mismos, en cuanto a la dificultad inherente de resolverlos?. Podemos entender *algoritmo* como una serie finita de pasos para resolver un problema. Los algoritmos se clasifican básicamente en dos tipos:

- *algoritmo determinístico*: tiene la propiedad de que el resultado de cada operación, se define en forma única.
- *algoritmo no-determinístico*: tiene la propiedad de que el resultado de una o varias operaciones, se determina dentro de un conjunto especificado de posibilidades.

Una forma sencilla de clasificar a un problema es la siguiente:

- *problema polinomial:* existe un algoritmo determinístico, capaz de resolverlo en tiempo polinomial.
- *problema no-polinomial:* no existe un algoritmo determinístico, capaz de resolverlo en tiempo polinomial.

El término tiempo polinomial hace referencia al tiempo que le toma al algoritmo resolver el problema, es decir, el tiempo requerido sigue una trayectoria polinomial.

El proyecto *Sistema de Lenguaje Medico Unificado, (Unified Medical Language System, UMLS)*, es desarrollado por la Biblioteca Nacional de Medicina de Estados Unidos, tiene como objetivo fundamental facilitar la recuperación de información mediante el empleo de frases creadas a partir del MeSH Vocabulary, y mediante el uso de sinóminos y variaciones del léxico del idioma inglés. Este ambicioso proyecto se compone de tres módulos fundamentales, llamados "fuentes de conocimiento".

- *UMLS Methatesaurus:* contiene información semántica sobre conceptos biomédicos, con sus posibles denominaciones y algunas de las relaciones más relevantes entre conceptos. Lo más interesante de este Metatesauro es que supone la integración y unificación de un buen número de tesauros, vocabularios, clasificaciones y sistemas de clasificación. Este Metatesauro está organizado conceptualmente, por lo que permite ver las denominaciones alternativas que diferentes clasificaciones y vocabularios usan para el mismo concepto.
- *UMLS Semantic Network:* una red de tipos semánticos o categorías generales (un total de unas 200) a las que se han asignado los conceptos del Metathesaurus.

- *Specialist Lexicon:* contiene información sintáctica básica sobre los términos biomédicos incluidos en el Metathesaurus. En la actualidad estos tres recursos se están usando en varios proyectos de investigación y desarrollo de aplicaciones, sobre todo en el ámbito de la recuperación de información y la indización automatizada.

Unidades de Almacenamiento: En informática, la unidad básica de almacenamiento es el *Bit*, el nombre se deriva del término *"Binary Digit"*. El *Bit* sólo puede tomar dos valores: el 0 y el 1, por lo cual su capacidad de almacenamiento es nulo. La siguiente unidad de almacenamiento es el Byte, que son 8 bits, con el cual, se puede almacenar una palabra. La tabla 2 muestra las relaciones entre las unidades de almacenamiento más usuales en informática.

UNIDAD	REPRESENTA	Byte (Aproximación.) (potencia)
Byte	8 bits	
Kilobyte	1024 bytes	10(3)
Megabyte	1024 kilobytes	10(6)
Gigabyte	1024 megabytes	10(9)
Terabyte	1024 gigabytes	10(12)
Petabyte	1024 terabytes	10(15)
Exabyte	1024 petabytes	10(18)
Zettabyte	1024 exabytes	10(21)
Yottabyte	1024 zettabytes	10(24)

Tipos de documentos: La Biblioteca Nacional de Medicina de los Estados Unidos indiza los siguientes tipos de documentos.

Clinical Trial
Editorial
Letter
Meta-Analysis
Practice Guideline
Randomized Controlled Trial
Review

- *Ensayo Clínico [Clinical Trial]:* Trabajo que reporta un estudio clínico (planificado con anticipación) sobre seguridad; eficacia; algún programa de dosificación de uno o más diagnósticos; terapéuticos; drogas profilácticas; dispositivos; algunas técnicas en humanos seleccionados de acuerdo a un criterio predeterminado de elegibilidad y evidencia observada a causa de efectos favorables y no favorables.
- *Editorial [Editorial]:* Trabajo que consistente de una manifestación de opiniones, creencias y políticas del redactor o editor del periódico, usualmente sobre temas actuales de medicina o de importancia científica a la comunidad médica o a la sociedad. Las editoriales publicadas por redactores de periódicos representan al órgano oficial de una sociedad u organización.
- *Carta [Letter]:* Trabajos que consisten de escritos o comunicaciones impresas entre individuos o entre personas y representantes jurídicos. La correspondencia puede ser personal o profesional. En las publicaciones medicas y en otras publicaciones

científicas, la carta va usualmente de los autores al editor. Y esta va acompañada de comentarios sobre el artículo.

- *Meta Análisis [Meta-Analysis]:* Trabajos consistentes de estudios que usan un método cuantitativo para la combinación de los resultados de estudios independientes y sintetizando resúmenes y conclusiones los cuales pueden ser usados para evaluar la efectividad terapéutica, estudios de planes nuevos, etc. Frecuentemente son revisiones de ensayos clinicos. Generalmente llamados meta-análisis por el autor o editor y deberán ser diferenciados de las revisiones de la literatura.
- *Normas de Practica [Practice Guideline]:* Trabajos que consisten de un conjunto de principios para ayudar a los profesionales de la salud con decisiones sobre el cuidado de pacientes basadas en diagnostico terapéutico u otros procedimientos clínicos bajo circunstancias clínicas especificas. Estos trabajos pueden ser desarrolladas por agencias gubernamentales, instituciones, organizaciones tales como las sociedades profesionales o paneles de expertos. Estos proporcionan un fundamento para la evaluación de la calidad y efectividad de los cuidados médicos en términos de medir el desarrollo de la salud, reducción de variaciones en servicios o procedimientos, y reducción de variaciones de resultados en los cuidados de la salud.
- *Ensayo Controlado Aleatorizado [Randomized Controlled Trial]:* Trabajo que consiste de un ensayo clínico que involucra al menos un tratamiento de prueba y un tratamiento de control, matriculación concurrente y continuación de pruebas -y control- de grupos, y en los cuales los tratamientos administrados son seleccionados por un proceso aleatorio, tal como el uso de tablas de numero aleatorios. La distribución de los tratamientos se hace por el lanzamiento de monedas, números pares e impares, números del seguro social, días de la semana, registros médicos, o algún otro proceso seudo aleatorio. Algún ensayo que emplea alguno de estos tipos de asignación simplemente se le llama ensayo clínico controlado.
- *Revisión [Review]:* Un artículo o un libro publicado después de la inspección del material publicado sobre un tema. Este puede ser extenso en varios grados y el rango de tiempo del material escrutado puede ser amplio o corto, pero las revisiones son muchas veces repasos de la literatura actual. El material textual examinado puede ser igualmente amplio y puede comprender, en medicina, el material clínico tanto como la investigación experimental o reporte de casos.

Apéndice B

Definiciones de los términos de la subcategoría de Matemáticas

( ) Indica el año en que el termino fue incorporado al MeSH Vocabulary

Algoritmo [Algorithms]: Un procedimiento que consiste de un secuencia de formulas algebraicas y/o pasos lógicos para calcular o determinar una tarea dada. (1987)

Análisis del Elemento Finito [Finite Element Análisis]: Una computadora basada en métodos de simulación o análisis del comportamiento de estructuras o componente. (1999)

Análisis de Fourier [Fourier Analysis]: Análisis basados en una función matemática formulada por Jean Baptiste Joseph Fourier en 1807. La función se conoce como la transformada de Fourier. Y describe el patrón senosoidal de algún patrón fluctuante en el mundo físico en términos de su amplitud y de su fase. Tiene amplias aplicaciones en

biomedicina, i.e., la cristalografía de rayos x analiza datos pivótales para la identificación de la doble hélice del DNA y en el análisis de otras moléculas. (1973)

Fractales [Fractals]: Patrones (reales o matemáticos) que lucen similares en diferentes escalas, por ejemplo, la red de ventilación del pulmón muestra patrones de ramificación similares en amplificaciones progresivas. Los fractales naturales son auto-similares a través de un rango finito de escalas mientras que los fractales matemáticos son los mismos en rangos infinitos. Algunas estructuras biológicas son fractales (o lucen como fractales). Los fractales están relacionados al "caos" (vea NONLINEAR DYNAMICS) en esos procesos caóticos pueden producir estructuras fractales en la naturaleza y representaciones apropiadas de procesos caóticos usualmente revelan auto-similaridad sobre el tiempo. (1994)

Teoría de Juegos [Game Theory]: Construcciones teóricas usadas en las matemáticas aplicadas para analizar ciertas situaciones en las cuales hay interpolación entre las partes que pueden tener intereses similares, opuestos o mixtos. En un juego típico, los jugadores toman decisiones; cada quien tiene sus propios objetivos; tratan de ganar ventaja sobre los demás anticipándose a las decisiones de los demás; el juego finalmente se resuelve como una consecuencia de las decisiones de los jugadores. (1969)

Juegos, Experimental [Games, Experimental]: Juegos diseñados para proporcionar información sobre hipótesis, políticas, procedimiento o estrategias. (1991)

Computación Matemática [Mathematical Computing]: Interpretación asistida por computadora para el análisis de funciones matemáticas relacionadas a un problema en particular. (1987)

Técnicas de Apoyo en las Decisiones [Decision Support Techniques]: Procedimientos matemáticos o estadísticos usados como ayuda en la toma de decisiones. Son frecuentemente usados en la toma de decisiones médicas. (1991)

Dinámica no Lineal [Nonlinear Dynamics]: El estudio de sistemas que responden desproporcionalmente (no lineal) a condiciones iniciales o estímulos. Los sistemas no lineales pueden exhibir caos lo cual se caracteriza como dependencia sensitiva a condiciones iniciales. Sistemas caóticos se distinguen de los sistemas periódicos por ser aleatorios. La conducta en el tiempo se muestra en el "espacio fase"; las restricciones son descritas por "atractores extraños". Las representaciones de los espacios fases de los sistemas caóticos o los atractores extraños usualmente revelan fractales. Algunos sistemas biológicos muestran dinámicas no lineales y caos. (1994)

Estadística [Statistics]: La ciencia y el arte de coleccionar, resumir y analizar datos que están relacionadas a variables aleatorias. El término es también aplicado a los mismos datos y al compendio de datos.

Análisis Actuarial [Actuarial Analysis]: La aplicación de métodos probabilísticos y estadísticos al calculo del riesgo de ocurrencia de algún evento, tales como enfermedades recurrentes, hospitalizaciones, o muerte. Puede incluir el cálculo del costo del evento y la prima necesaria para el pago de tales costos. (1997)

Tablas de Vida [Life Tables]: Técnicas de agregación usadas para describir el patrón de mortalidad y sobrevivencia de la población. Esos métodos pueden ser aplicados al estudio no solamente de la muerte, sino también a la ocurrencia de complicaciones de enfermedades (1990).

Calidad ajustada de vida [Quality-Adjusted Life Years]: Un índice derivado de la modificación de una de tabla de vida a través de procedimientos estándares y diseñado para tomar en cuenta la calidad tanto como la duración de sobrevivencia. Este índice puede ser usado para evaluar el resultado de un procedimiento de salubridad o servicios. (1994)

Análisis de Varianza [Analysis of Variante]: Una técnica estadística que aísla y determina las contribuciones de variables categóricas independientes a la variación en la media de una variable continúa dependiente. (1974)

Análisis Multivariado [Multivariate Análisis]: Un conjunto de técnicas usadas cuando variaciones en muchas variables tienen que ser estudiadas simultáneamente. En estadística, el análisis multivariado es interpretado como algún método analítico que permita simultáneamente estudiar dos o más variables dependientes. (1990)

Biometria [Biometry]: El uso de métodos estadísticos para el análisis de observaciones biológicas y de fenómenos.

Análisis de Conglomerados [Cluster Analysis]: Un conjunto de métodos estadísticos usados para agrupar variables u observaciones en subgrupos fuertemente relacionados. En epidemiología, se usa para analizar series de eventos o casos de enfermedades o de otros fenómenos relacionados con la salud con patrones de distribución bien definidos en relación al tiempo o al lugar de nacimiento. (1990)

Análisis de Áreas - Pequeñas [Small - Area Analysis]: Un método para el análisis de las variaciones en la utilización de los cuidados médicos en área geográficas pequeñas o áreas demográficas. Es frecuentemente usado, por ejemplo, las tasas de uso para un servicio dado o un procedimiento en muchas áreas pequeñas, documentar las variaciones entre las áreas. Comparando áreas de alto y bajo uso, los análisis determinan si hay patrones de tal uso y para identificar variables que están asociadas con en variación y contribuya a la variación (1992)

Conglomerados Espacio – Tiempo [Space-Time Clustering]: Un exceso de significancía estadística de casos de una enfermedad, ocurrida dentro de un límite continúo de espacio tiempo. (1978)

Intervalos de Confianza [Confidence Intervals]: Un rango de valores para una variable de interés i.e. una tasa, construida de tal forma que este rango tiene una probabilidad de incluir el valor verdadero de la variable. (1991)

Interpretación de Datos, Estadísticos [Data Interpretation, Statistical]: Aplicación de procedimientos estadísticos para el análisis de observaciones especificas o asumir hechos de un estudio en particular. (1988)

Análisis de Discriminante [Discriminant Analysis]: Una técnica analítica estadística usada con variables discretas dependientes, concerniente con la separación de conjuntos de los valores observados y asignación de nuevos valores. Es algunas veces usada en vez del análisis de regresión. (1990)

Análisis de Factores, Estadístico [Factor Analysis, Statistical]: Un conjunto de métodos estadísticos para el análisis de correlaciones entre muchas variables para determinar el

número de dimensiones fundamentales que sea la base de los datos observados y describir y medir esas dimensiones. Se usa frecuentemente en el desarrollo de sistemas de notas de escalas de grado y cuestionarios.

Análisis del Mejor – Par [Matched-Pair Analysis]: Un tipo de análisis en el cual temas en un grupo de estudio y un grupo de comparación son comparables con respecto a factores extraños de parejas individuales de temas de estudio con la comparación de los grupos de temas (i.e. controles de edades - emparejadas). (1992)

Método de Monte Carlo [Monte Carlo Method]: En estadística, una técnica de aproximación numérica para la solución de un problema matemático por medio de estudiar las distribuciones de algunas variables aleatorias, frecuentemente generadas por computadora. El nombre se refiere a la característica aleatoria de los juegos de azar de los casinos en Monte Carlo. (1991)

Análisis de Componentes Principales [Principal Component Analysis]: Procedimientos matemáticos que transforman un número de posibles variables correlacionadas en un número pequeño de variables no correlacionadas llamado componentes principales. (2002)

Probabilidad [Probability]: El estudio de los procesos de ocasión o la caracterización de la frecuencia relativa de un proceso de ocasión. (1968)

Teorema de Bayes [Bayes Theorem]: Un teorema de la teoría de la probabilidad debido a Thomas Bayes (1702-1761). En epidemiología, se usa para obtener la probabilidad de enfermedades en un grupo de personas con alguna característica sobre la base de la tasa total de esa enfermedad y de la probabilidad de esa característica de salud y enfermedades individuales. La aplicación más familiar esta en el análisis de decisiones clínicas donde se usa para estimar la probabilidad de un diagnostico particular dada la apariencia de algún síntoma o el resultado de un prueba. (1991)

Funciones de probabilidad [Likelihood Functions]: Funciones construidas a partir de modelos estadísticos y un conjunto de datos observados. Dan la probabilidad de que los parámetros desconocidos del modelo tomen varios valores. Esos valores de los parámetros que maximizan la probabilidad son los estimadores de máxima verosimilitud de los parámetros. (1990)

Cadenas de Markov [Markov Chains]: Un proceso estocástico tal que la distribución condicional de probabilidad para algún estado en algún instante futuro dado el estado actual no es afectado por algún conocimiento adicional del pasado del sistema. (1991)

Cociente de Probabilidades [Odds Ratio]: El cociente de dos probabilidades. El cociente de probabilidades de exposición, en el caso de control de datos, es el cociente de la probabilidad de estar en el caso siendo expuestos y la probabilidad de estar en el caso siendo no expuesto. El cociente de probabilidad de enfermedades, para un cohorte o una sección, es el cociente de la probabilidad de enfermarse de los expuestos y la probabilidad de enfermarse de los no expuestos. El cociente de probabilidades de predominio se refiere a una probabilidad cociente derivada de secciones de cruz de estudios de casos de frecuentes. (1991)

Valores Pronósticos de Pruebas [Predictive Value of Tests]: En la investigación y en el diagnostico de pruebas (i.e. tiene la enfermedad) se refiere al valor pronosticado de una

prueba positiva; Mientras que, el valor pronosticado de una prueba negativa es la probabilidad de que una persona con una prueba negativa no tenga la enfermedad. El valor pronosticado se refiere a la sensibilidad y especificidad de la prueba. (1987)

Modelos de Peligros Proporcionales [Proportional Hazards Models]: Modelos estadísticos usados en el análisis de sobrevivencia que afirma que los efectos de los factores de estudio sobre tasa de peligro en el estudio de poblaciones es multiplicativa y no cambia con el tiempo. (1990)

Riesgo [Risk]: La probabilidad que un evento ocurra. Abarca una variedad de medidas de probabilidad de un resultado generalmente no favorable. (1988)

Modelos Logísticos [Logistic Models]: Modelos estadísticos que describen las relaciones entre un variable cualitativa dependiente (esto es, una que toma solamente ciertos valores discretos, tal como, la ausencia o presencia de una enfermedad) y una variable independiente. Una aplicación común en epidemiología es la estimación de riesgos individuales (probabilidad de una enfermedad) como una función de un factor de riesgo dado. (1990)

Riesgo de Gravamen [Risk Assessment]: La estimación cualitativa o cuantitativa de la probabilidad de aversión al efecto que puede resultar de la exposición a específicos peligros para la salud o de la ausencia de influencias beneficiosas. (1995)

Factores de Riesgo [Risk Factors]: Un aspecto de una conducta personal o estilo de vida, exposiciones ambientales, características innatas o características heredadas, sobre las cuales, basándose en evidencia epidemiológica, se asocian con condiciones relacionadas a la salud consideradas importantes de prevenir. (1988)

Incertidumbre [Uncertainty]: La condición en cual un conocimiento razonable con respecto a riesgos, beneficios o ventajas o el futuro no es disponible. (2003)

Análisis de Regresión [Regression Analysis]: Procedimientos para encontrar la función matemática que describa las relaciones entre una variable dependiente y una o más variables independientes. En regresión lineal las relaciones están restringidas a una línea recta y el análisis de mínimos cuadrados es usado para determinar el mejor ajuste. En regresión logística la variable dependiente es cualitativa más una variable continua y las funciones de máxima verosimilitud son usadas para encontrar la mejor relación. En regresión múltiple la variable dependiente es considerada que depende de muchas variables independiente. (1980)

Análisis de Mínimos Cuadrados [Least-Squares Analysis]: Un principio de estimación en la cual las estimaciones de un conjunto de parámetros en un modelo estadístico son esas que minimizan la suma de los cuadrados de la diferencia entre los valores observados de una variable dependiente y los valores pronosticados por el modelo. (1991)

Modelos Lineales [Linear Models]: Modelos estadísticos en los cuales el valor de un parámetro para un valor dado de un factor es asume igual a $a + bx$ donde a y b son constantes. (1990)

Curva ROC [ROC Curve]: Un medio grafico para determinar la habilidad de una prueba de investigación para discriminar entre personas saludables y enfermas; pueden también ser

usadas en otros estudios, i.e. distinguir estímulos de respuestas, estímulos débiles o respuestas sin estímulos. (1988)

Sensibilidad y Especificidad [Sensitivity and Specificity]: Medidas para determinar el resultado de un diagnostico y de pruebas de investigación. La sensibilidad representa la proporción de personas enfermas en una población que son identificadas por la prueba como enfermas. Es una medida de la probabilidad de diagnosticar correctamente una condición. La especificidad es la proporción de las personas no enfermas quienes son identificadas por la prueba. Es una medida de la probabilidad de identificar correctamente a las personas no enfermas por la prueba. (1991)

Distribuciones Estadísticas [Statistical Distributions]: El resumen completo de las frecuencias de los valores o categorías de una medida sobre un grupo de ítems, una población o alguna otra colección de datos. La distribución dice cuantos o que proporción del grupo fue tenia cada valor (o rango de valores) de todos lo posibles valores que la medida cuantitativa puede tener (1998)

Distribución Binomial [Binomial Distribution]: La distribución de probabilidad asociada con dos resultados mutuamente exclusivos; usada para modelar tasa de incidencia acumulativa y tasas de predominó. La distribución Bernoulli es un caso especial de la distribución binomial. (1990)

Distribución Chi-Cuadrada [Chi-Square Distribution]: Una distribución en la cual una variable es distribuida como la suma de los cuadrados de alguna variable aleatoria independiente, cada una de las cuales tiene distribución normal con media cero y varianza uno. La prueba de la Chi-cuadrada es una prueba estadística basada en la comparación de un estadístico de prueba de una distribución Chi-cuadrada. La más vieja de esas pruebas son usadas para detectar si dos o más distribuciones poblacionales difieren de otra. (1990)

Distribución Normal [Normal Distribution]: Distribución de frecuencia continua de rango infinito. Sus propiedades son: 1, continua, distribución simétrica con ambas colas extendiéndose al infinito; 2, media aritmética, moda y mediana idéntica; y 3, forma determinada completamente por la media y la desviación estándar. (1990)

Distribución Poisson [Poisson Distribution]: Una función de distribución usada para describir la ocurrencia de eventos raros o para describir la distribución maestral de conteos aislados en un continuo de tiempo o espacio. (1990)

Estadística No Parametrica [Statistics, Nonparametric]: Una clase de métodos estadísticos aplicable a grandes colecciones de distribuciones de probabilidad usadas para probar correlación, localización, independencia, etc. En muchas pruebas estadísticas no parametricas, los scores originales u observaciones son reemplazadas por otras variables que contienen menos información. Una clase importante de pruebas parametricas emplea las propiedades ordinales de los datos. Otra clase importante de pruebas usa información sobre si una observación esta arriba o debajo de algún valor fijo tal como la mediana, y una tercera clase se basa en la frecuencia de la ocurrencia de corridas en los datos. (1995)

Procesos Estocásticos [Stochastic Processes]: Procesos que incorporan algunos elementos de aleatoriedad, usados particularmente para referirse a una serie de tiempo de variables aleatorias. (1991, [1975])

Análisis de Sobrevivencia [Survival Analysis]: Una clase de procedimientos estadísticos para estimar la función de sobrevivencia (función de tiempo, iniciando con una población al 100% en un tiempo dado y a condición de que el porcentaje de la población aun tiempo después). El análisis de sobrevivencia es usado para hacer inferencia sobre los efectos de tratamientos, pronósticos de factores, expuestos y entre otros.

Sobrevivientes Libre-Enfermedad [Disease-Free Survival]: Periodo después de un tratamiento exitoso en el cual no hay apariencia de síntomas o efectos de la enfermedad. (1995)

Ciencias Naturales [Natural Sciences]: Las ciencias relacionadas con los procesos observables en la naturaleza. (1998)

Apéndice C

Estadísticas del capítulo 5

Estadísticas correspondientes al comportamiento de la indización por año en MedLine. **(Cantidad Aproximada de Documentos 15, 200, 000 hasta finales del 2004)**

Año	Documentos indizados	Año	Documentos indizados	Año	Documentos indizados
1950	80797	1970	213304	1990	398017
1951	101009	1971	218798	1991	399999
1952	106252	1972	223085	1992	403905
1953	106929	1973	226614	1993	412123
1954	103654	1974	230162	1994	422171
1955	106188	1975	244391	1995	432528
1956	104791	1976	249047	1996	442972
1957	109459	1977	255686	1997	440652
1958	107160	1978	265526	1998	459385
1959	107427	1979	274299	1999	477927
1960	108374	1980	272503	2000	517600
1961	116736	1981	274448	2001	526534
1962	124091	1982	285434	2002	543165
1963	139770	1983	299201	2003	568939
1964	158703	1984	308070	2004	597965
1965	173213	1985	338400		
1966	174710	1986	338400		
1967	186696	1987	356388		
1968	203998	1988	374616		
1969	211200	1989	390738		

Estadísticas correspondientes a la proporción de documentos que han sido indizados con algún término correspondiente a la Categoría de Ciencias Físicas. **(Total de documentos 6, 411, 641 hasta el primer semestre del 2005).**

Physical Sciences Categories (Natural Sciences)	Suma Acumulada de documentos
Physics	1482511
Chemistry	1237820
Science	1227067
Mathematics	1016052
Time	743409
Biological Sciences	516131
Weights and Measures	133027

Electronics	21063
Geography	19114
Geology	6570
Astronomy	4161
Nanotechnology	3606
Evolution, Chemical	513
Evolution, Planetary	401
Nature	196
Suma total	**6,411,641**

Estadísticas correspondientes al comportamiento de los términos matemáticos en el proceso de indización de la literatura biomédica (entre 1950 y 2004).

Año	**Cantidad de documentos indizados de:**								
	Mathematics	Algorithms	Finite Element Analysis	Fourier Analysis	Fractals	Game Theory	Mathematical Computing	Nonlinear Dynamics	Statistics
1950	0	0	0	0	0	0	0	0	71
1951	0	0	0	0	0	0	0	0	53
1952	0	0	0	0	0	0	0	0	42
1953	0	0	0	0	0	0	0	0	52
1954	0	0	0	0	0	0	0	0	73
1955	0	0	0	0	0	0	0	0	40
1956	0	0	0	0	0	0	0	0	36
1957	0	0	0	0	0	0	0	0	39
1958	0	0	0	0	0	0	0	0	30
1959	0	0	0	0	0	0	0	0	46
1960	0	0	0	0	0	0	0	0	55
1961	0	0	0	0	0	0	0	0	58
1962	0	0	0	0	0	0	0	0	100
1963	4	0	0	0	0	0	0	0	2583
1964	24	0	0	0	0	0	0	0	5289
1965	229	0	0	0	0	0	0	0	3153
1966	558	0	0	1	0	0	0	0	1374
1967	858	0	0	0	0	1	0	0	1362
1968	1536	0	0	0	0	18	0	0	1596
1969	1619	1	0	0	0	13	1	0	1893
1970	1731	1	0	0	0	4	0	0	1455
1971	2276	0	0	3	0	12	1	0	1690
1972	2877	3	0	34	0	26	0	0	1358
1973	3038	0	0	126	0	30	1	0	1688
1974	3464	2	0	138	0	21	1	1	1679

1975	2767	0	0	119	0	20	0	0	1027
1976	2466	4	0	143	0	11	32	0	1026
1977	1966	0	0	89	0	21	1	0	1043
1978	1939	4	0	76	0	26	4	0	1218
1979	2208	3	0	109	0	23	8	0	1237
1980	2186	1	0	94	0	26	28	0	1310
1981	1991	0	0	87	0	18	20	0	1229
1982	2078	0	0	116	0	13	26	0	1396
1983	2178	2	0	125	0	18	17	0	1537
1984	2090	12	0	163	0	21	20	1	1548
1985	2517	36	0	180	0	20	48	0	1505
1986	2502	174	0	223	0	17	91	0	1679
1987	2016	543	0	190	0	21	277	1	1871
1988	1642	955	1	248	0	15	649	1	1659
1989	1915	817	0	302	1	23	919	0	1413
1990	1682	934	0	344	0	31	974	1	1132
1991	1814	1153	0	422	2	32	1232	1	1052
1992	1859	1303	0	467	6	26	1396	4	781
1993	1643	1455	0	443	34	25	1529	44	631
1994	1428	1880	0	462	67	36	1870	149	780
1995	1737	1964	0	475	70	39	2319	160	694
1996	1060	1954	1	446	77	45	2526	224	554
1997	885	2470	8	407	76	39	2486	254	524
1998	703	2866	125	442	88	25	3034	290	603
1999	567	3663	150	451	111	40	3030	399	704
2000	620	4129	195	495	82	55	3092	456	938
2001	773	4650	201	462	130	43	3329	468	1866
2002	697	4980	314	477	93	74	3227	477	2395
2003	688	6246	371	463	107	69	4040	592	1997
2004	667	7672	278	404	94	115	4417	662	1812

Estadísticas correspondientes el porcentaje de documentos que han sido indizados con algún término perteneciente a **Statistics (Total de documentos 932,144 hasta el primer semestre del 2005).**

Estadística (Statistics)	**Suma acumulada de documentos indizados**
Probability	433824
Sensitivity and Specificity	156647
Analysis of Variante	118720
Regression Analysis	113515
Biometry	60900
Survival Analysis	50980
Statistical Distributions	26921
Statistics, Nonparametric	20429
Data Interpretation, Statistical	16870
Actuarial Analysis	12437
Confidence Intervals	11957
Factor Analysis, Statistical	9903
Cluster Analysis	8429
ROC Curve	6750
Monte Carlo Method	6503
Stochastic Processes	5975
Discriminant Analysis	3495
Matched-Pair Analysis	2073
Principal Component Analysis	462
Suma Total	**1066790**
	932, 144

Estadísticas correspondientes el porcentaje de documentos que han sido indizados con algún término perteneciente a *Mathematics* **(Total de documentos 1, 037, 883 hasta el primer semestre del 2005)**

Subcategoría de Matemáticas	Suma acumulada de Documentos
Statistics	938708
Mathematics	67628
Algorithms	51945
MathematicalComputing	41829
FourierAnalysis	9332
NonlinearDynamics	4367
FiniteElementAnalysis	1719
GameTheory	1137
Fractals	1074
Suma total	**1,037,883**

Estadísticas correspondientes de los almacenes (entre 1950 y 2004)

Almacén	Suma acumulada de documentos indizados
Almacén de Matematicas	116, 612
Almacén de Estadistica	887, 877
Almacén de Intersección	51, 311
Suma total	1, 055, 800

Apéndice D

Categoría de Ciencias Biologicas (Biological Sciences Category)

Estos términos son las filas (o variables) de las matrices de coocurrencia. Total de términos de esta categoría 1994. Esta categoría comprende todas las divisiones de las ciencias naturales que se enfocan a los aspectos de los fenómenos de la vida y a los procesos vitales. Los conceptos incluyen anatomia, fisiologia, bioquímica, biofisica, biología de animales, plantas y microorganismos.

Biological Sciences Category
- Biochemical Phenomena +
- Metabolism +
- Nutrition +

Biological Phenomena, Cell Phenomena, and Immunity
- Biological Phenomena +
- Cell Physiology +
- Immunity +
- Plant Physiology +
- Species Specificity

Biological Sciences
- Anatomy +
- Biochemistry +
- Biology +
- Biophysics +
- Biotechnology +
- Neurosciences +
- Pharmacology +
- Physiology +

Chemical and Pharmacologic Phenomena
- Biopharmaceutics +
- Cytoprotection
- Depression, Chemical
- Dose-Response Relationship, Drug
- Down-Regulation
- Drug Design
- Drug Interactions +
- Drug Resistance +
- Drug Tolerance +
- Inhibitory Concentration 50
- Kinetics
- Lethal Dose 50

Maximum Tolerated Dose
No-Observed-Adverse-Effect Level
Stimulation, Chemical
Structure-Activity Relationship +
Up-Regulation

Circulatory and Respiratory Physiology

Blood Physiology +
Cardiovascular Physiology +
Respiratory Physiology +

Digestive, Oral, and Skin Physiology

Dental Physiology +
Digestive Physiology +
Skin Physiology +

Environment and Public Health

Environment +
Public Health +
Public Health Dentistry +

Genetic Phenomena

Consanguinity
Founder Effect
Gene Frequency +
Gene Order
Gene Pool
Genetic Load
Genomic Instability +
Genotype +
Hybrid Vigor
Inheritance Patterns +
Linkage (Genetics) +
Phenotype +
Phylogeny
Ploidies +
Sequence Homology +
Sex Ratio
Structural Homology, Protein
Variation (Genetics) +

Genetic Processes

Breeding +
Cell Division +
DNA Damage +
DNA Methylation
DNA Packaging +
DNA Repair +
DNA Replication +
Evolution +
Gene Expression +
Gene Expression Regulation +
Gene Rearrangement +
Heredity
Mutagenesis +
Recombination, Genetic +
Selection (Genetics)
Sex Determination (Genetics)
Virus Integration +

Genetic Structures

Attachment Sites (Microbiology)
Base Sequence +
Chromosome Structures +
Chromosomes +
Gene Library +
Genes +

Genetic Code +
Genetic Vectors +
Genome +
Genome Components +
Histone Code
Plasmids +
Templates, Genetic

Health Occupations

Acupuncture
Allied Health Occupations +
Biomedical Engineering
Chiropractic
Dentistry +
Dietetics
Environmental Health +
Health Services Administration
Hospital Administration
Medical Illustration
Medicine +
Mortuary Practice +
Nursing +
Nursing, Practical
Nutrition
Optometry
Orthoptics
Pharmacology +
Pharmacy
Podiatry
Psychology, Medical
Serology
Sociology, Medical
Specialism
Technology, Pharmaceutical
Veterinary Medicine +

Musculoskeletal, Neural, and Ocular Physiology

Musculoskeletal Physiology +
Nervous System Physiology +
Ocular Physiology +

Physiological Processes

Adaptation, Physiological +
Body Constitution +
Body Temperature +
Cell Physiology +
Chronobiology +
Electrophysiology +
Growth and Development +
Homeostasis +
Somatotypes

Reproductive and Urinary Physiology

Reproduction +
Urinary Tract Physiology +

Apéndice E

Recuperación de información

Se muestran todas las citas de los documentos que tratan de los temas que se encuentran en los núcleos de “Nonlinear Dynamics” y “Neural Networks (Computer)”.

Citas recuperadas para Nonlinear Dynamics

Embryology

No.	Autores	Titulo	Revista	Año de Publicación
1	Zhang, Y. //Weng, J. //Hwang, W. S.	Auditory learning: a developmental method	IEEE Trans Neural Netw	2005 May
2	Connor, R. M. //Allen, C. L. //Devine, C. A. //Claxton, C. //Key, B.	BOC, brother of CDO, is a dorsoventral axon-guidance molecule in the embryonic vertebrate brain	J Comp Neurol	2005 Apr 25
3	Xu, H. //Whelan, P. J. //Wenner, P.	Development of an inhibitory interneuronal circuit in the embryonic spinal cord	J Neurophysiol	2005 May
4	Koutmani, Y. //Hurel, C. //Patsavoudi, E. //Hack, M. //Gotz, M. //Thomaidou, D. //Matsas, R.	BM88 is an early marker of proliferating precursor cells that will differentiate into the neuronal lineage	Eur J Neurosci	2004 Nov
5	Honda, H. //Kobayashi, T.	Large-scale micropropagation system of plant cells	Adv Biochem Eng Biotechnol	2004
6	Oyen, M. L. //Calvin, S. E. //Cook, R. F.	Uniaxial stress-relaxation and stress-strain responses of human amnion	J Mater Sci Mater Med	2004 May
7	Hansen, M. J. //Dallal, G. E. //Flanagan, J. G.	Retinal axon response to ephrin-as shows a graded, concentration-dependent transition from growth promotion to inhibition	Neuron	2004 Jun 10
8	Dryer, S. E. //Lhuillier, L. //Cameron, J. S. //Martin-Caraballo, M.	Expression of K(Ca) channels in identified populations of developing vertebrate neurons: role of neurotrophic factors and activity	J Physiol Paris	2003 Jan
9	Glover, J. C.	The development of vestibulo-ocular circuitry in the chicken embryo	J Physiol Paris	2003 Jan
10	Chrobok, V. //Meloun, M. //Simakova, E.	Descriptive growth model of the height of stapes in the fetus: a histopathological study of the temporal bone	Eur Arch Otorhinolaryngol	2004 Jan
11	Pribyl, M. //Muratov, C. B. //Shvartsman, S. Y.	Discrete models of autocrine cell communication in epithelial layers	Biophys J	2003 Jun
12	Tchuraev, R. N. //Galimzyanov, A. V.	Parametric stability evaluation in computer experiments on the mathematical model of Drosophila control gene subnetwork	In Silico Biol	2003
13	Gurrin, L. C. //Moss, T. J. //Sloboda, D. M. //Hazelton, M. L. //Challis, J. R. //Newnham, J. P.	Uising WinBUGS to fit nonlinear mixed models with an application to pharmacokinetic modelling of insulin response to glucose challenge in sheep exposed antenatally to glucocorticoids	J Biopharm Stat	2003 Feb
14	Konner, M.	Weaving life's pattern	Nature	2002 Jul 18
15	Melnick, M. //Jaskoll, T.	Mouse submandibular gland morphogenesis: a paradigm for embryonic signal processing	Crit Rev Oral Biol Med	2000

16	White, T. //Andreasen, N. C. //Nopoulos, P.	Brain volumes and surface morphology in monozygotic twins	Cereb Cortex	2002 May
17	Aizenberg, I. //Myasnikova, E. //Samsonova, M. //Reinitz, J.	Temporal classification of Drosophila segmentation gene expression patterns by the multi-valued neural recognition method	Math Biosci	2002 Mar
18	Waliszewski, P. //Konarski, J.	Neuronal differentiation and synapse formation occur in space and time with fractal dimension	Synapse	2002 Mar 15
19	Clarke, B. C. //Hobbs, M. //Skylas, D. //Appels, R.	Genes active in developing wheat endosperm	Funct Integr Genomics	2000 May
20	Bem, T. //Le Feuvre, Y. //Simmers, J. //Meyrand, P.	Electrical coupling can prevent expression of adult-like properties in an embryonic neural circuit	J Neurophysiol	2002 Jan
21	Honda, H. //Liu, C. //Kobayashi, T.	Large-scale plant micropropagation	Adv Biochem Eng Biotechnol	2001
22	Beloussov, L. V.	Morphogenetic fields: outlining the alternatives and enlarging the context	Riv Biol	2001 May-Aug
23	Lapeer, R. J. //Prager, R. W.	Fetal head moulding: finite element analysis of a fetal skull subjected to uterine pressures during the first stage of labour	J Biomech	2001 Sep
24	Reinhard, I. //Wellek, S.	Age-related reference regions for longitudinal measurements of growth characteristics	Methods Inf Med	2001 May
25	Schwab, M. //Schmidt, K. //Roedel, M. //Mueller, T. //Schubert, H. //Anwar, M. A. //Nathaniels, P. W.	Non-linear changes of electrocortical activity after antenatal betamethasone treatment in fetal sheep	J Physiol	2001 Mar 1
26	Akiyama, R. //Nagashima, T. //Tazawa, H.	Dynamical systems analysis of arterial blood pressure signals in relation to heart rate fluctuations in chick embryos	Comp Biochem Physiol A Mol Integr Physiol	1999 Dec
27	van Heijst, J. J. //Touwen, B. C. //Vos, J. E.	Implications of a neural network model of early sensori-motor development for the field of developmental neurology	Early Hum Dev	1999 May
28	Brunner, K. //Kussinger, M. //Stetter, M. //Lang, E. W.	A neural network model for the emergence of grating cells	Biol Cybern	1998 May
29	Vrba, E. S.	Multiphasic growth models and the evolution of prolonged growth exemplified by human brain evolution	J Theor Biol	1998 Feb 7
30	Gurgen, F. //Onal, E. //Varol, F. G.	IUGR detection by ultrasonographic examinations using neural networks	IEEE Eng Med Biol Mag	1997 May-Jun
31	Bassukas, I. D.	Use of the recursion formula of the Gompertz survival function to evaluate life-table data	Mech Ageing Dev	1996 Aug 29
32	Dambricourt Malasse, A.	[Unpublished aspects of hominization. Fundamental ontogeny, chaotic aspects, harmonic aspects]	Acta Biotheor	1995 Jun
33	Burstein, Z.	A network model of developmental gene hierarchy	J Theor Biol	1995 May 7
34	Beksac, M. S. //Durak, B. //Ozkan, O. //Cakar, A.	An artificial intelligent diagnostic system with neural	Eur J Obstet Gynecol Reprod Biol	1995 Apr

	N. //Balci, S. //Karakas, U. //Laleli, Y.	networks to determine genetical disorders and fetal health by using maternal serum markers		
35	Lin, I. E. //Taber, L. A.	Mechanical effects of looping in the embryonic chick heart	J Biomech	1994 Mar
36	Kowtha, V. C. //Kunysz, A. //Clay, J. R. //Glass, L. //Shrier, A.	Ionic mechanisms and nonlinear dynamics of embryonic chick heart cell aggregates	Prog Biophys Mol Biol	1994
37	Mjolsness, E. //Sharp, D. H. //Reinitz, J.	A connectionist model of development	J Theor Biol	1991 Oct 21

Biological Sciences

No.	Autores	Titulo	Revista	Año de Publicación
1	Robertson, R.	The case of the missing third	Nonlinear Dynamics Psychol Life Sci	2005 Jan
2	May, R. M.	Uses and abuses of mathematics in biology	Science	2004 Feb 6
3	Zipfel, W. R. //Williams, R. M. //Webb, W. W.	Nonlinear magic: multiphoton microscopy in the biosciences	Nat Biotechnol	2003 Nov
4	Patterson, P. E.	Development of a learning module using a virtual environment to demonstrate EMG and telerobotic control principles	Biomed Sci Instrum	2002
5	Theodoropoulos, G. //Loumos, V.	Parasitology tutoring system: a hypermedia computer-based application	Comput Methods Programs Biomed	1994 Feb 14

Developmental Biology

No.	Autores	Titulo	Revista	Año de Publicación
1	Fitzgerald, M.	The development of nociceptive circuits	Nat Rev Neurosci	2005 Jul
2	Konner, M.	Weaving life's pattern	Nature	2002 Jul 18
3	Robinson, S. R.	Dynamical systems: a grammar for discussing process in ontogeny?	Dev Psychobiol	1995 Dec
4	Hofer, M. A.	Notes of a chaos watcher	Dev Psychobiol	1995 Dec
5	Ermentrout, G. B. //Edelstein-Keshet, L.	Cellular automata approaches to biological modeling	J Theor Biol	1993 Jan 7

Systems Biology

No.	Autores	Titulo	Revista	Año de Publicación
1	Mellor, J. C. //Wu, J. //Delisi, C.	Constructing networks with correlation maximization methods	Genome Inform Ser Workshop Genome Inform	2004
2	Yang, Z. R.	Biological applications of support vector machines	Brief Bioinform	2004 Dec
3	Galvanauskas, V. //Simutis, R. //Lubbert, A.	Hybrid process models for process optimisation, monitoring and control	Bioprocess Biosyst Eng	2004 Dec
4	Morgan, J. J. //Surovtsev, I. V. //Lindahl, P. A.	A framework for whole-cell mathematical modeling	J Theor Biol	2004 Dec 21
5	Priami, C. //Quaglia, P.	Modelling the dynamics of biosystems	Brief Bioinform	2004 Sep

Citas recuperadas para Neural Networks (Computer)

Parasitology

No.	Autores	Titulo	Revista	Año de Publicación
1	Sanchez-Ramos, I. //Castanera, P.	Effect of temperature on reproductive parameters and longevity of Tyrophagus putrescentiae (Acari: Acaridae)	Exp Appl Acarol	2005
2	Malkin, E. M. //Durbin, A. P. //Diemert, D. J. //Sattabongkot, J. //Wu, Y. //Miura, K. //Long, C. A. //Lambert, L. //Miles, A. P. //Wang, J. //Stowers, A. //Miller, L. H. //Saul, A.	Phase 1 vaccine trial of Pvs25H: a transmission blocking vaccine for Plasmodium vivax malaria	Vaccine	2005 May 2
3	Bejon, P. //Andrews, L. //Andersen, R. F. //Dunachie, S. //Webster, D. //Walther, M. //Gilbert, S. C. //Peto, T. //Hill, A. V.	Calculation of liver-to-blood inocula, parasite growth rates, and preerythrocytic vaccine efficacy, from serial quantitative polymerase chain reaction studies of volunteers challenged with malaria sporozoites	J Infect Dis	2005 Feb 15
4	Widmer, K. W. //Srikumar, D. //Pillai, S. D.	Use of artificial neural networks to accurately identify Cryptosporidium oocyst and Giardia cyst images	Appl Environ Microbiol	2005 Jan
5	Peleg, M. //Tu, S. //Manindroo, A. //Altman, R. B.	Modeling and analyzing biomedical processes using workflow/Petri Net models and tools	Medinfo	2004
6	Brockman, A. //Singlam, S. //Phiaphun, L. //Looareesuwan, S. //White, N. J.	Field evaluation of a novel colorimetric method--double-site enzyme-linked lactate dehydrogenase	Antimicrob Agents Chemother	2004 Apr

	//Nosten, F.	immunodetection assay--to determine drug susceptibilities of Plasmodium falciparum clinical isolates from northwestern Thailand		
7	Xu, X. //Ma, F. //Zou, Y. //Cheng, X.	[Chaotic diagnosis of Nilaparvata lugens occurrence system]	Ying Yong Sheng Tai Xue Bao	2003 Aug
8	Davidson, G. //Phelps, K. //Sunderland, K. D. //Pell, J. K. //Ball, B. V. //Shaw, K. E. //Chandler, D.	Study of temperature-growth interactions of entomopathogenic fungi with potential for control of Varroa destructor (Acari: Mesostigmata) using a nonlinear model of poikilotherm development	J Appl Microbiol	2003
9	Bjornstad, O. N. //Peltonen, M. //Liebhold, A. M. //Baltensweiler, W.	Waves of larch budmoth outbreaks in the European alps	Science	2002 Nov 1
10	Ranta, E. //Lundberg, P. //Kaitala, V. //Stenseth, N. C.	Ecology. On the crest of a population wave	Science	2002 Nov 1
11	Widmer, K. W. //Oshima, K. H. //Pillai, S. D.	Identification of Cryptosporidium parvum oocysts by an artificial neural network approach	Appl Environ Microbiol	2002 Mar
12	Solano, P. //Guegan, J. F. //Reifenberg, J. M. //Thomas, F.	Trying to predict and explain the presence of African trypanosomes in tsetse flies	J Parasitol	2001 Oct
13	Yang, Y. S. //Park, D. K. //Kim, H. C. //Choi, M. H. //Chai, J. Y.	Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network	IEEE Trans Biomed Eng	2001 Jun
14	Wood, S. N.	Minimizing model fitting objectives that contain spurious local minima by bootstrap restarting	Biometrics	2001 Mar
15	Gubbins, S. //Gilligan, C. A. //Kleczkowski, A.	Population dynamics of plant-parasite interactions: thresholds for invasion	Theor Popul Biol	2000 May
16	Theodoropoulos, G. //Loumos, V. //Anagnostopoulos, C. //Kayafas, E. //Martinez-Gonzales, B.	A digital image analysis and neural network based system for identification of third-stage parasitic strongyle larvae from domestic animals	Comput Methods Programs Biomed	2000 Jun
17	Sun, P. //Yang, X. B.	Dynamic behaviors of the Ricker population model under a set of randomized perturbations	Math Biosci	2000 Apr
18	Kay, J. W. //Shinn, A. P. //Sommerville, C.	Towards an automated system for the identification of notifiable pathogens: using as an example	Parasitol Today	1999 May
19	Marion, G. //Renshaw, E. //Gibson, G.	Stochastic effects in a model of nematode infection in ruminants	IMA J Math Appl Med Biol	1998 Jun
20	Gupta, S. //Ferguson, N. //Anderson, R.	Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents	Science	1998 May 8
21	Subramanian, S. //Krishnamoorthy, K. //Ramaiah, K. D. //Habbema, J. D. //Das, P. K. //Plaisier, A. P.	The relationship between microfilarial load in the human host and uptake and development of Wuchereria bancrofti microfilariae by Culex quinquefasciatus: a study under natural conditions	Parasitology	1998 Mar
22	Basco, L. K. //Le Bras, J.	In vitro sensitivity of Plasmodium falciparum to anti-folinic agents (trimethoprim, pyrimethamine, cycloguanil): a study of 29 African strains]	Bull Soc Pathol Exot	1997

23	Cheng, G. //Liu, X. //Wu, J. X. //Jones, B.	Establishing a reliable visual function test and applying it to screening optic nerve disease in onchocercal communities	Int J Biomed Comput	1996 Mar
24	Cavalieri, L. F. //Kocak, H.	Chaos: a potential problem in the biological control of insect pests	Math Biosci	1995 May
25	Theodoropoulos, G. //Loumos, V.	Parasitology tutoring system: a hypermedia computer-based application	Comput Methods Programs Biomed	1994 Feb 14

Neurosciences

No.	Autores	Titulo	Revista	Año de Publicació
1	Robertson, R.	The case of the missing third	Nonlinear Dynamics Psychol Life Sci	2005 Jan
2	Sandri, G.	Does computation provide a model for creativity? An epistemological perspective in neuroscience	J Endocrinol Invest	2004
3	Dalgleish, T.	The emotional brain	Nat Rev Neurosci	2004 Jul
4	Bug, W. //Nissanov, J.	A guide to building image-centric databases	Neuroinformatics	2003
5	Gutkin, B. //Pinto, D. //Ermentrout, B.	Mathematical neuroscience: from neurons to circuits to systems	J Physiol Paris	2003 Mar-May
6	Grant, S. G.	Systems biology in neuroscience: bridging genes to cognition	Curr Opin Neurobiol	2003 Oct
7	Truccolo, W. A. //Rangarajan, G. //Chen, Y. //Ding, M.	Analyzing stability of equilibrium points in neural networks: a general approach	Neural Netw	2003 Dec
8	Eng, K. //Klein, D. //Babler, A. //Bernardet, U. //Blanchard, M. //Costa, M. //Delbruck, T. //Douglas, R. J. //Hepp, K. //Manzolli, J. //Mintz, M. //Roth, F. //Rutishauser, U. //Wassermann, K. //Whatley, A. M. //Wittmann, A. //Wyss, R. //Verschure, P. F.	Design for a brain revisited: the neuromorphic design and functionality of the interactive space 'Ada'	Rev Neurosci	2003
9	Kenward, M. //Morris, R. //Tarassenko, L.	Neural connections that compute	Trends Neurosci	2003 Aug
10	Bob, P.	Dissociation and neuroscience: history and new perspectives	Int J Neurosci	2003 Jul
11	Brinkley, J. F. //Rosse, C.	Imaging and the Human Brain Project: a review	Methods Inf Med	2002
12	Westen, D. //Gabbard, G. O.	Developments in cognitive neuroscience: II. Implications for theories of transference	J Am Psychoanal Assoc	2002 Winter
13	Westen, D. //Gabbard, G. O.	Developments in cognitive neuroscience: I. Conflict, compromise, and connectionism	J Am Psychoanal Assoc	2002 Winter
14	Beedle, A. S.	A philosopher looks at neuroscience	J Neurosci Res	1999 Jan 15

15	Shepherd, G. M. //Mirsky, J. S. //Healy, M. D. //Singer, M. S. //Skoufos, E. //Hines, M. S. //Nadkarni, P. M. //Miller, P. L.	The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data	Trends Neurosci	1998 Nov
16	Henningsen, P.	[Self-recognition in the mirror of another? On the significance of cognitive neuroscience for psychoanalysis]	Psychother Psychosom Med Psychol	1998 Mar-Apr
17	Bower, J. M.	What will save neuroscience?	Neuroimage	1996 Dec
18	Barto, A. G.	Reinforcement learning control	Curr Opin Neurobiol	1994 Dec
19	Milner, P. M.	The mind and Donald O. Hebb	Sci Am	1993 Jan

Neuropsychology

No.	Autores	Titulo	Revista	Año de Publicación
1	Chamberlain, S. R. //Blackwell, A. D. //Fineberg, N. A. //Robbins, T. W. //Sahakian, B. J.	The neuropsychology of obsessive compulsive disorder: the importance of failures in cognitive and behavioural inhibition as candidate endophenotypic markers	Neurosci Biobehav Rev	2005 May
2	Parsons, T. D. //Rizzo, A. A. //Buckwalter, J. G.	Backpropagation and regression: comparative utility for neuropsychologists	J Clin Exp Neuropsychol	2004 Feb
3	Bullinaria, J. A.	Dissociation in connectionist systems	Cortex	2003 Feb
4	Plaut, D. C.	Interpreting double dissociations in connectionist networks	Cortex	2003 Feb
5	Kello, C. T.	The emergence of a double dissociation in the modulation of a single control parameter in a nonlinear dynamical system	Cortex	2003 Feb
6	Deco, G. //Pollatos, O. //Zihl, J.	The time course of selective visual attention: theory and experiments	Vision Res	2002 Dec
7	Deco, G. //Zihl, J.	A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system	J Comput Neurosci	2001 May-Jun
8	Kennepohl, S.	Toward a cultural neuropsychology: An alternative view and a preliminary model	Brain Cogn	1999 Dec
9	Emrich, H. M.	[Cognition and feelings--on the neurobiology of the perception-emotion connection]	Geburtshilfe Frauenheilkd	1996 Jan
10	Wu, F. Y. //Slater, J. D. //Ramsay, R. E.	Neural network approach in multichannel auditory event-related potential analysis	Int J Biomed Comput	1994 Apr
11	Popper, K. R. //Lindahl, B. I. //Arhem, P.	A discussion of the mind-brain problem	Theor Med	1993 Jun
12	Rossler, O. E. //Rossler, R.	Is the mind-body interface microscopic?	Theor Med	1993 Jun

Photobiology

No.	Autores	Titulo	Revista	Año de Publicación
1	Di Paolo, E. A.	Evolving spike-timing-dependent plasticity for single-trial learning in robots	Philos Transact A Math Phys Eng Sci	2003 Oct 15
2	Jung, S. K. //Lee, S. B.	Image analysis of light distribution in a photobioreactor	Biotechnol Bioeng	2003 Nov 5
3	Cornet, J. F. //Favier, L. //Dussap, C. G.	Modeling stability of photoheterotrophic continuous cultures in photobioreactors	Biotechnol Prog	2003 Jul-Aug

Environmental Microbiology

No.	Autores	Titulo	Revista	Año de Publicación
1	Schlager, K. J.	Status and progress in on-line spectrometric monitoring and control of plant nutrient solutions	Adv Space Res	1996
2	Giacomini, M. //Ruggiero, C. //Calegari, L. //Bertone, S.	Artificial neural network based identification of environmental bacteria by gas-chromatographic and electrophoretic data	J Microbiol Methods	2000 Dec 1

Environmental Microbiology

No.	Autores	Titulo	Revista	Año de Publicación
1	van Someren, E. P. //Wessels, L. F. //Backer, E. //Reinders, M. J.	Genetic network modeling	Pharmacogenomics	2002 Jul

Apendice F

Descubrimiento de Conocimiento

En el apéndice F se muestran todas las fichas en el formato MedLine. Observe los términos y frases que se han resaltado del resumen.

Author, Analytic (01): Mtetwa, N. //Smith, L. S.

Article Title (04): Precision constrained stochastic resonance in a feedforward neural network

Medium Designator (05): A_Biomimetics_NNC

Connective Phrase (06):

Journal Title (10): IEEE Trans Neural Netw

Translated Title (11):

Reprint Status (12): **Date:**

Date of Publication (20): 2005 Jan

Volume ID (22): 16

Issue ID (24): 1

Page(s) (25): 250-62

Language (35): eng

Connective Phrase (36):

Address/Availability (37): Department of Computing Science, University of Stirling, Stirling FK9 4LA, UK. nmt@cs.stir.ac.uk

Location/URL (38): 15732404

ISSN (40): 1045-9227

Notes (42):

Abstract (43): Stochastic resonance (SR) is a phenomenon in which the response of a **nonlinear system** to a subthreshold information-bearing signal is optimized by the presence of noise. By considering a nonlinear system (network of leaky integrate-and-fire (LIF) neurons) that captures the **functional dynamics of neuronal firing**, we demonstrate that sensory neurons could, in principle harness SR to optimize the detection and transmission of weak stimuli. We have previously characterized this effect by use of signal-to-noise ratio (SNR). Here in addition to SNR, we apply an entropy-based measure (Fisher information) and compare the two measures of quantifying SR. We also discuss the performance of these two SR measures in a full precision floating point model simulated in Java and in a precision limited integer model simulated on a field programmable gate array (FPGA). We report in this study that stochastic resonance which is mainly associated with floating point implementations is possible in both a single LIF neuron and a network of LIF neurons implemented on lower resolution integer based digital hardware. We also report that such a network can improve the SNR and Fisher information of the output over a single LIF neuron.

Call Number (44):

Keywords (45): Action Potentials: :*physiology/Animals/**Biomimetics: :*methods**/Comparative Study/Computer Simulation/Feedback/Humans/*Models, Neurological/Models, Statistical/*Nerve Net/***Neural Networks (Computer)**/Neurons: :*physiology/Research Support, Non-U.S. Gov't/Stochastic Processes/Synaptic Transmission: :*physiology

Ficha 1 en formato MedLine

Author, Analytic (01): Voutsas, K. //Langner, G. //Adamy, J. //Ochse, M.

Article Title (04): A brain-like neural network for periodicity analysis

Medium Designator (05): A_Biomimetics_NNC

Connective Phrase (06):

Journal Title (10): IEEE Trans Syst Man Cybern B Cybern

Translated Title (11):

Reprint Status (12): **Date:**

Date of Publication (20): 2005 Feb

Volume ID (22): 35

Issue ID (24): 1

Page(s) (25): 12-22

Language (35): eng

Connective Phrase (36):

Address/Availability (37): Control Theory and Robotics Laboratory, Technical University Darmstadt, 64283 Darmstadt, Germany.

Location/URL (38): 15719929

ISSN (40): 1083-4419

Notes (42):

Abstract (43): This paper introduces a brain-like neural model for sound processing. The periodicity analyzing network (PAN) is a bio-inspired neural network of spiking neurons. The PAN consists of complex models of neurons, which can be used for understanding the dynamics of individual neurons and neuronal networks. On a technical level, the PAN is able to compute the ratio of modulation and carrier frequency of harmonic sound signals. The PAN model may, therefore, be used in audio signal processing applications, such as sound source separation, **periodicity analysis**, and the cocktail party problem.

Call Number (44):

Keywords (45): Algorithms/Animals/Auditory Cortex: :*physiology/Auditory Perception: :*physiology/Biological Clocks: :*physiology/**Biomimetics: :*methods**/Computer Simulation/Hearing: :physiology/Humans/ *Models, Neurological/Nerve Net: :*physiology/***Neural Networks (Computer)**/Periodicity/ Research Support, Non-U.S. Gov't/Sound Spectrography: :*methods

Ficha 2 en formato MedLine

Author, Analytic (01): Lin, C. M. //Peng, Y. F.

Article Title (04): Missile guidance law design using adaptive cerebellar model articulation controller

Medium Designator (05): A_Biomimetics_NNC

Connective Phrase (06):

Journal Title (10): IEEE Trans Neural Netw

Translated Title (11):

Reprint Status (12): **Date:**

Date of Publication (20): 2005 May

Volume ID (22): 16

Issue ID (24): 3

Page(s) (25): 636-44

Language (35): eng

Connective Phrase (36):

Address/Availability (37): Department of Electrical Engineering, Yuan-Ze University, Chung-Li 320 Taiwan, ROC. cml@saturn.yzu.edu.tw

Location/URL (38): 15940993

ISSN (40): 1045-9227

Notes (42):

Abstract (43): An adaptive cerebellar model articulation controller (CMAC) is proposed for command to line-of-sight (CLOS) missile guidance law design. In this design, the three-dimensional (3-D) CLOS guidance problem is formulated as a tracking problem of a **time-varying nonlinear system**. The adaptive CMAC control system is comprised of a CMAC and a compensation controller. The CMAC control is used to imitate a feedback linearization control law and the compensation controller is utilized to compensate the difference between the feedback linearization control law and the CMAC control. The online adaptive law is derived based on the **Lyapunov stability theorem** to learn the weights of receptive-field basis functions in CMAC control. In addition, in order to relax the requirement of approximation error bound, an estimation law is derived to estimate the error bound. Then the adaptive CMAC control system is designed to achieve satisfactory tracking performance. Simulation results for different engagement scenarios illustrate the validity of the proposed adaptive CMAC-based guidance law.

Call Number (44):

Keywords (45): *Algorithms/Animals/**Biomimetics: :*methods**/Cerebellum: :*physiology/Computer Simulation/ Feedback: :physiology/Humans/*Models, Neurological/Movement: :*physiology/Nerve Net: :*physiology/***Neural Networks (Computer)**/Pattern Recognition, Automated: :methods/Research Support, Non-U.S. Gov't/Signal Processing, Computer-Assisted/War

Ficha 3 en formato MedLine

Author, Analytic (01): Harter, D. //Kozma, R.

Article Title (04): Chaotic neurodynamics for autonomous agents

Medium Designator (05): A_Biomimetics_NNC

Connective Phrase (06):

Journal Title (10): IEEE Trans Neural Netw

Translated Title (11):

Reprint Status (12): **Date:**

Date of Publication (20): 2005 May

Volume ID (22): 16

Issue ID (24): 3

Page(s) (25): 565-79

Language (35): eng

Connective Phrase (36):

Address/Availability (37): Division of Computer Science, University of Memphis, TN 38152, USA.

Location/URL (38): 15940987

ISSN (40): 1045-9227

Notes (42):

Abstract (43): Mesoscopic level neurodynamics study the collective dynamical behavior of neural populations. Such models are becoming increasingly important in understanding large-scale brain processes. Brains exhibit **aperiodic oscillations** with a much more rich dynamical behavior than **fixed-point and limit-cycle** approximation allow. Here we present a discretized model inspired by Freeman's K-set mesoscopic level population model. We show that this version is capable of replicating the important principles of aperiodic/chaotic neurodynamics while being fast enough for use in real-time autonomous agent applications. This simplification of the K model provides many advantages not only in terms of efficiency but in simplicity and its ability to be analyzed in terms of its dynamical properties. We study the discrete version using a multilayer, highly recurrent model of the neural architecture of perceptual brain areas. We use this architecture to develop example action selection mechanisms in an autonomous agent.

Call Number (44):

Keywords (45): Action Potentials: :physiology/Animals/Artificial Intelligence/Biological Clocks: :*physiology/**Biomimetics: :*methods**/Brain: :*physiology/Computer Simulation/Humans/*Models, Neurological/Nerve Net: :*physiology/Neural Inhibition: :*physiology/***Neural Networks (Computer)**/Research Support, U.S. Gov't, Non-P.H.S./Robotics: :methods/Synaptic Transmission: :*physiology

Ficha 4 en formato MedLine

Referencias

[1] Calvelo-Ríos, M. “El Papel de las Tecnologías de Información y Comunicación en el Desarrollo Rural y la Seguridad Alimentaría”. Disponible en http://www.fao.org/sd/CDdirect/CDre0055e.htm, Acceso 3/01/06

[2] Fayyad, U. Haussler, D. “KDD for Science Data Analysis: Issues and Examples”. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Oregon, USA, 1996.

[3] Klemettinen, M. Mannila, H. “A Data Methodology and Its Application to Semi-Automatic Knowledge Acquisition”. Proceedings of the 8th International Workshop on Database and Expert Systems Applications, USA, 1997.

[4] Una nota publicada por: Redacción, La Flecha 10/03/2004; “El pc no distingue: solo ve datos, el aumento de los datos digitales”. Disponible en http://weblogs.cfired.org.ar/blog/archives/000226.php , Acceso 3/01/06

[5] Porter, A. Zhu, D. “Technology Opportunities Analysis: Illustrated for the Case of Knowledge Discovery in Databases and Data Mining”. TPAC report, 1998.*

[6] Piatetsky-Shapiro. Matheus, C. “Knowledge Discovery in Databases: An Overview”, AI Magazine, 1992.

[7] Fayyad, U. Piatetsky-Shapiro. “Knowledge Discovery and Data Mining: Towards a Unifying Framework”. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Oregon, USA, 1996.

[8] Fayyad, U. Piatetsky-Shapiro. “From Data Mining to Knowledge Discovery: An Overview”. Advantedge in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.

[9] Fayyad, U. Piatetsky-Shapiro. “The KDD Process for Extracting Useful Knowledge from Volumenes of Data”. Communications of the ACM, 39[11]: 27-34, 1996.

[10] Williams, G. Huang, Z. “Modelling the KDD Process: A Four Stage Process and Four Element Model”. CSIRO Division of Information Technology, Data Mining Portfolio – TR DM 96013, 1996.

[11] Michie, D. J. Spiegelhalter, C. C. Taylor. “Machine Learning, Neural and Statistical Classification”. Ellis Horwood, 1994. Disponible en: http://citeseer.ist.psu.edu/, Acceso 3/01/06

[12] Holsheimer, M. Siebes, A. “Data Mining: The Search for Knowledge in Databases”, Report CS-R9406, ISSN 0169-118X, Amersterdam, The Netherlands 1991. Disponible en: http://citeseer.ist.psu.edu/holsheimer91data.html.

[13] Silberschatz, A. Tuzhilin, A. “What Makes Patterns Interesting in Knowledge Discovery Systems”. IEEE Transactions on Knowledge and Data Engineering, Volume 8, Issue 6, Pages: 970 – 974, 1996.

[14] Hilderman R.J., Hamilton H.J. “Knowledge discovery and interestingness measures: A survey”. Technical Report CS 99-04, Department of Computer Science, University of Regina, October 1999. Disponible en: http://citeseer.ist.psu.edu/hilderman99knowledge.html, Acceso 3/01/06

[15] Kamber, M. “Data Mining: Concepts and Techiques”. Morgan Kaufmman Publishers, 2001.

[16] Glymour, C. Madigan, D. “Statistical Themes and Lessons for Data Mining”, Kluwer Academic Publishers, Data Mining and Knowledge Discovery 1, 11-28, 1997.

[17] Keim A. Kriegel H. “Visualization Techniques for Mining Large Databases: A Comparison”. In Transactions on Knowledge and Data Engineering TKDE'96), Special Issue on Data Mining, Vol. 8, No. 6, 1996, pp. 923-938.

[18] Abbas, H. Sarker, R. “Data Mining: A Heuristic Approach”, Idea Group Publishing, 2002.

[19] Bigus J., “Data Mining with neural networks”, Mc GrawHill, USA, 1996.

[20] Markus H. “Data Mining: Challenges, Models, Methods and Algorithms”. 2003 Disponible en: http://citeseer.ist.psu.edu/hegland03data.html., Acceso 3/01/07

[21] Martín S. Castro, S. “Interacción en Visualización de Información”. Laboratorio de Investigación en Visualización y Computación Gráfica (VyGLab). Disponible en: http://www.educ.ar/educar/servlet/Downloads/S_BD_AREA6/INTERACCION_EN_VISUALIZ.PS.PDF, Acceso 3/01/06

[22] Grupo de Investigación en Visualización, “Visualización”. Disponible en: http://www.educ.ar/educar/servlet/Downloads/S_BD_WICK2000/TRA7_3.PDF, Acceso 3/01/06

[23] Barry de Ville, “Microsoft Data Mining: Integrated Bussines Intelligence for e-Commerce and Knowledge Management”, Digital Press, 2001.

[24] Delmater, R. Hancock, M. “Data mining Explained: A Manager´s Guide to Customer-Centric Business Intelligence”. Digital Press, 2001.

[25] Guzmán M. V. “Seminario-Taller”. Instituto Finlay. 2001.

[26] Soparkar, N. Uthurusamy, R. “System for KDD: From Concepts to Practice”. Future Generation Computer Systems, Volume 13, Issue 2-3, Special double issue on data mining, Pages: 231 – 242, 1997.

[27] Piatetsky - Shapiro G. Chan, P.;"Systems for Knowledge Discovery in Databases"; to appear in the IEEE TKDE special issue on Learning & Discovery in Knowledge - Based Databases; 1993.

[28] Peggy W. "El Descubrimiento del Conocimiento en las Bases de Datos: Herramientas y Técnicas". Disponible en: www.acm.org/crossroads/espanol/xrds5-2/kdd.html , Acceso 3/01/06

[29] Disponible en: http://www.crisp-dm.org/ , Acceso 3/01/06

[30] Disponible en: http://www.sas.com/ , Acceso 3/01/06

[31] Carrillo Calvet H. Guzmán Sánchez, MV. ''ViBlioSOM: Aplicaciones en MEDLINE, Generación Automática de Mapas de Conocimiento con Redes Neuronales para el Descubrimiento de Conocimiento en Bases de Información Biomédica", Versión 1.0, Laboratorio de Dinámica no Lineal de la Facultad de Ciencias, UNAM, Mexico, D.F. 2005

[32] Vanti, N. "Métodos Cuantitativos de Evaluación de la Ciencia: Bibliometría, Cienciometría e Informetría". Investigación Bibliotecológica, v14, No. 29, julio/diciembre de 2000.

[33] Tague-Sutcliffe J. "An introduction to Informetrics". Information Processing & Management. 1992; 28(1): 1-3. Versión condensada, Lic. José Antonio López Espinosa ACIMED 3(2):26-35, septiembre-diciembre, 1994

[34] Dorothy H. Hertzel, Bibliometrics History, Case Western Reserve University, Cleveland. Ohio, USA.

[35] Jiménez-Contreras E. "Los métodos bibliométricos: Estado de la cuestión y aplicaciones". Dpto. de B. y Documentación. Universidad de Granada, Primer Congreso Universitario de Ciencias de la Documentación. Disponible en: wotan.liu.edu/dois/data/Papers/juljuljut3484.html, Acceso 3/01/06

[36] Carrizo Sainero, G. "Hacia un concepto de Bibliometría", Universidad Carlos III, de Madrid. Disponible en: wotan.liu.edu/dois/data/Articles/juljuljary:2000:v:1:i:2:p:8528.html, Acceso 3/01/06

[37] Sotolongo-Aguilar G. "Sistema de Información Bibliométrica". Tesis presentada en opción al grado científico de Doctor en Ciencias de la Información, Instituto Finlay, Cuba, 2002.

[38] Gurjeva G. "Early Soviet Scientometrics and Scientometricians", Universiteit van Amsterdam. Thesis for the degree of MSc in Science Dynamics. 1992. Disponible en: http://www.chstm.man.ac.uk/people/gurjeva.htm, Acceso 3/01/06

[39] Ruiz A. Arencibia R. "Informetría, Bibliometría y Cienciometría: aspectos teórico-prácticos", ACIMED 04 2002.

[40] Van Raan, A. F. J. Scientometrics: state of art. Scientometrics, Ámsterdam, v. 38. n.1. p. 205 – 218, 1997.

[41] Disponible en: http://www.inria.fr/recherche/equipes/cortex.en.html, Acceso 3/01/06

[42] Sierra Flores, MM. "Identificación y Estudio de los Principales Grupos de Investigación en el Campo de la Física de la UNAM a través de Indicadores Bibliométricos". UNAM, Facultad de Filosofía y Letras, División de Estudio de Posgrado. Tesis Maestría en Bibliotecología (Maestro en Bibliotecología). 126h, 2005.

[43] Disponible en: http://www.dynamics.unam.edu/ptid/redinformatica/, Acceso 3/01/06

[44] Macias-Chapula, C. "Papel de la Informetría y de la Cienciometría y su perspectiva nacional e internacional". Ciencias de la Información, Brasilia, v. 27, n. 2, p. 134-140, mayo-agosto de 1998.

[45] Polanco. X. ″Aux Sources de la Scientométrie″. SOLARIS n¡ 2, Presses Universitaires de Rennes, p. 13-79. 1995. Disponible en: http://www.info.unicaen/bnum/jelec/Solaris. Acceso 3/01/06

[46] Noyer, J. ″Scientométrie, Informétrie : Pourquoi Nous Intéresent – Elles ?″ . SOLARIS n¡ 2, Presses Universitaires de Rennes, p. 1-23. Disponible en: http://www.info.unicaen/bnum/jelec/Solaris. Acceso 3/01/06

[47] Gorbea Portal. S.; ″Modelacion matematica de la Actividad Bibliotecaria: Una Revision″. Investigación Bibliotecologica v. 12. No. 24 enero/junio de 1998. Disponible en http://www.ejournal.unam.mx/iibiblio/iib_v12-24.html. Acceso 3/01/06

[48] Bookstein, A., "Robustness Properties of the Bibliometric Distributions". Disponible en: http://citeseer.ist.psu.edu/137755.html, Acceso 3/01/06

[49] Ungern-Sternberg, S. "Bradford´s Law in the Context of Information Provision", Scientometrics, V.49. No. 1 (2000) 161-186.

[50] Urbizagástegui. A., "La Ley de Lotka y la Literatura de Bibliometría". Investigación Bibliotecológica, V.13, NO. 27, 1999.

[51] Guzmán M. Sotolongo G. "Seminario-Taller: Sistema de Gestión de Información Bibliométrico – Métodos y herramientas". Instituto Finlay. 2001.

[52] Guzmán M. "Patentometría: Herramienta para el Análisis de Oportunidades Tecnológicas". Master gerencia de información en las organizaciones Cátedra UNESCO.

[53] Bordons. M., Zulueta, M. "Evaluación de la Actividad Científica a través de Indicadores Bibliométricos", Rev Esp Cardiol 1999; 52: 790-800, ISSN: 1579-2242.

[54] Price, D J. S., “Little Science, Big Science” New York, Columbia University Press.1963

[55] Sancho, R. “Indicadores Bibliométricos utilizados en la Evaluación de la Ciencia y la Tecnología”. Revisión bibliográfica; Rev. Esp. Doc. Cienti.13, 3-4, 1990.

[56] Vanegas Arbeláez, N. “Inventario Breve de Índices e Indicadores de Ciencia y Tecnología”, Universidad de Antioquia, 2003. Disponible en: http://purace.unicauca.edu.co/DeInteres/ContratoSocialCyT/mesa_9.html, Acceso 3/01/06

[57] Spinak, E. “Indicadores Cienciométricos”, ACIMED v.9 n.s supl.4, 2001.

[58] Russell J. “El Uso de las Bases de Datos Bibliogràficas en la Definicion de Politicas en Ciencia y Tecnologia en America Latina”. La informacion en el inicio de la era electronica; organizacion del conocimiento y sistemas de informacion. Mexico, CUIB, UNAM, 1998, vol. 1.

[59] Kodratoff Y. "About Knowledge Discovery in Texts: A Definition and an Example". Proc. ISMIS’99, Warsaw, June 1999. Disponible en: http://citeseer.ist.psu.edu/kodratoff00about.html., Acceso 3/01/06

[60] Kodratoff Y. “Knowledge Discovery in Texts: A Definition and Applications”; in Foundation of Intelligent Systems, Ras & Skowron (Eds.) LNAI 1609, Springer 1999.

[61] Oscar J. “La difusión de la actividad científica mediante publicaciones”. Instituto de Historia de la Ciencia y Documentación. Disponible en: www.imedea.uib.es/public/cursoid/_html/textos/Tema%2012.4%20JO%20txt.pdf, Acceso 3/01/06

[62] Council of biology Editors. Proposed definition of a primary publication. Newsletter Bethesda, CBE 1968

[63] Sánchez-Paus. L. “Fuentes de Información Bibliográfica y Estadística en Economía”. Jefe de Información Bibliográfica Biblioteca de la Facultad de Ciencias Económicas y Empresariales. UCM. Disponible en: http://www.ucm.es/BUCM/cee/010001.htm, Acceso 3/01/06

[64] Russell J. “La Bibliométria en la Evaluación de la Ciencia”. Curso impartido en el Centro Universitario de Investigaciones Bibliotecológicas, CUIB, UNAM. Del 14 al 18 de junio de 2004.

[65] Castells E. “La Vigilancia Tecnológica, Requisito Imprescindible para la Innovación”. Disponible en: www.iale.es. Acceso 3/01/06

[66] Disponible en: http://www.ncbi.nlm.nih.gov, Acceso 3/01/06

[67] Jiménez M. “Acceso a MEDLINE y LILACS mediante el MeSH y el DeCS”. ACIMED v.6 n.3, 1998.

[68] Excelente manual en donde se explican aspectos técnicos del la indización. Disponible en la siguiente dirección de Pubmed: www.chapter_19_Qualifiers_Subheadings.htm, Acceso 3/01/06

[69] Villaseñor E. “Análisis Inteligente de Datos con Redes Neuronales Artificiales”. UNAM, Facultad de Ciencias, Tesis de Licenciatura. 2004.

[70] Carreras A. “Realism: Neural Networks and Representation”. Disponible en http://155.210.60.15/Dpto_filo/Alberto/texsantiago96.htm, Acceso 3/01/06

[71] Kohonen T., “The Self-Organizing Map”, 3ra. Edición -Verlag, Springer, 2001.

[72] Kohonen T, The Self-Organizing Map, Neurocomputing 21 (1998) 1Ð6

[73] Kaski S., “Data Exploration Using Self-Organizing Maps”, Ph. D. Thesis, Helsinki University of Technology, Finland, 1997.

[74] Disponible en http://www.ugr.es/~rruizb/cognosfera/index.htm, Acceso 3/01/06