UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

POSGRADO EN CIENCIAS
MATEMATICAS

FACULTAD DE CIENCIAS

DESCUBRIENDO CONOCIMIENTO EN CORPUS DE DOCUMENTOS MEDLINE

T E S I S

QUE PARA OBTENER EL GRADO DE
ACADEMICO DE MAESTRO EN CIENCIAS
P R E S E N T A :

EDGAR VALENCIA ROMERO

DIRECTOR DE TESIS: DOCTOR JOSE LUIS MARTINEZ MORALES

MEXICO, D.F. MARZO 2009

A mi familia.

A la banda de la biblioteca del IIMAS, en especial a Cecilia.

A Karina.

Gracias a todas aquellas personas que contribuyeron a la realización de esta tesis. En especial, al director de tesis y al grupo de sinodales por sus valiosas correcciones, observaciones y comentarios.

- *Dr. José Luis Martínez Morales*
- *Dr. Carlos Díaz Ávalos*
- *Dra. Silvia Ruiz.Velasco Acosta*
- *Dr. José María González Barrios Murguía*
- *M. en C. Crisanto Castillo Castillo*

Arcadio Buendía no logró descifrar el sueño de las casas con paredes de espejos hasta el día en que conoció el hielo. Entonces creyó entender su profundo significado. Pensó que en un fututo próximo podrían fabricarse bloques de hielo en gran escala, a partir de un material tan cotidiano como el agua, y construir con ellos las nuevas casa de la aldea.

Gabriel García Márquez

§

Descubriendo Conocimiento en Corpus de Documentos MedLine

Descubriendo Conocimiento en Corpus de Documentos MedLine

Resumen

Se exhibe brevemente por qué el campo de la Cienciometría ha encontrado en el Descubrimiento de Conocimiento en Bases de Datos, uno de sus modos de realización más prometedores en la actualidad. En especial, el indicador relacional de segunda generación denominado Análisis de Palabras Asociadas es una herramienta potente para descubrir conocimiento en corpus de documentos MedLine.

Tabla de Contenido

Índice de Ilustraciones

Índice de Tablas

Introducción

En la actualidad existe una gran diversidad de datos generados por las tecnologías de información y comunicación. Estos datos son almacenados en una gran variedad de bases de datos. Incluso los datos que se almacenan sobre un mismo tema, son tan diversos en lenguaje, forma, tamaño, etc., que sin herramientas apropiadas es prácticamente imposible analizarlos. Se tiene la certeza de un *potencial conocimiento* en espera de ser descubierto en estos datos almacenados.

Las técnicas de análisis de información que se venían usando para tal fin han sido ampliamente superadas por estos inmensos volúmenes de datos. El campo del Descubrimiento de Conocimiento en Bases de Datos, DCBD, utiliza toda una gama de herramientas provenientes de la Estadística, Inteligencia Artificial, Descubrimiento Científico Automatizado, etc, para encontrar este potencial conocimiento en el contexto de grandes conjuntos de datos.

En la actualidad una gran diversidad de bases de datos científicas y tecnológicas almacenan los resultados de la comunicación científica y técnica. Las bases de datos científicas - tecnologicas poseen herramientas muy sofisticadas de búsqueda y recuperación de información pero no poseen herramientas que permitan a la comunidad cientifica y técnica extraer conocimiento de un corpus de documentos, ya sean científicos o técnicos.

El trabajo tiene como objetivo exhibir brevemente por qué la Cienciometría ha encontrado en el Descubrimiento de Conocimiento en Bases de Datos, DCBD, uno de sus modos de realización más prometedores en la actualidad. La construcción de las redes tecno – científicas requiere alto rendimiento computacional y técnicas de visualización que conjuntamente permitan procesar corpus de documentos científicos - técnicos y visualizar grandes extensiones de la red y sub-redes. Cada una de estas sub-redes representa un *centro de interés*, es decir, zonas de la red muy enlazadas y consistentes, asimilables a polos de atracción de gran intensidad informativa. Representan a los actores temáticos más relevantes, de más significado en el paradigma de la investigación en el período en estudio. Si algo es realmente importante, aparece como centro de interés; si su importancia es pequeña o está difuminada, no se manifiesta.

Capítulo 1. *El Descubrimiento de Conocimiento en Bases de Datos* tiene el objetivo de extraer conocimiento de grandes cantidades de datos por medio de una serie de pasos sistemáticos que van desde la limpieza de los datos hasta la visualización e implementación del conocimiento.

Capítulo 2. *La Cienciométria* ofrece toda una gama de indicadores para realizar análisis cuantitativos de la literatura científica y técnica a varios niveles de especificidad.

Capítulo 3. *PubMed* es una base bibliografica, administrada por National Library of Medicine de Estados Unidos de America. *MedLine* es una base bibliográfica especializada en temas biomédicos que pertenece a PubMed. Los documentos MedLine están indizados con descriptores MeSH, los cuales, son revisados y actualizados anualmente por un equipo de profesionales de National Library of Medicine. Los documentos MedLine son clave en la realización de análisis cienciométricos encaminados al campo biomédico.

Capítulo 4. *Los Mapas Auto - Organizantes* de Teuvo Kohonen han demostrado ser muy útiles en el descubrimiento de conocimiento. La combinación del vector de cuantización y la proyección de datos los hacen una herramienta potente en la visualización y análisis de datos multidimensionales.

Capítulo 5. El indicador relacional de segunda generación denominado Análisis de Palabras Asociadas es una herramienta potente para descubrir conocimiento en corpus de documentos MedLine. Se reconstruyo la red asociada a corpus de documentos MedLine indizados con el MeSH Major Topic "Nonlinear Dynamics" con el fin de conocer las relaciones y contenido del centro de interés "Dinámica no Lineal" en la investigación biomédica en curso.

1 El Descubrimiento de Conocimiento en Bases de Datos

El volumen de las bases de datos de cualquier tema crece año con año gracias a las *tecnologías de información y comunicación*[1]. Estas tecnologías han hecho que las sociedades del siglo XXI dispongan de 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, [Calvelo, 2000]. Algunas instituciones o empresas que utilizan estas tecnologías generan, en periodos breves, una gran cantidad de información:

- El sistema SKICAT [Sky Image Cataloging and Analysis Tool System] explora el cosmos y ha enviado más de 3 terabytes (Vea Apéndice A) de imágenes de estrellas, planetas, galaxias, etc., [Fayyad et al., 1996A].
- Wal–Mart realiza más de 20 millones de transacciones diarias y tiene una base de datos de 11 terabytes.
- Mobil Oil busca almacenar 100 terabytes de datos de exploración petrolera.
- Genbank contiene más de 400 millones de secuencias de DNA.
- Las computadoras más potentes del mundo (en Celera y en Oak Ridge National Laboratory, por ejemplo, con una capacidad de cálculo de 2 teraflops (Vea Apéndice A)) están dedicadas a la investigación biológica, concretamente a la obtención y al análisis de las secuencias de nucleótidos de los genomas conocidos.

Los analistas de información, administradores de conocimiento, científicos, etc., consideran que en estas grandes cantidades de datos se esconde conocimiento en espera de ser descubierto. La extracción del *potencial conocimiento* requiere procesos automatizados que vayan examinando los datos con técnicas de aprendizaje inteligente. El campo del Descubrimiento de Conocimiento en Bases de Datos, DCBD, utiliza toda una gama de herramientas provenientes de la Estadística, Inteligencia Artificial, Descubrimiento Científico Automatizado, etc, con el objetivo de extraer conocimiento de alto-nivel de datos de bajo-nivel en el contexto de grandes conjuntos de datos, [Fayyad et al., 1996B]. El tamaño de las bases de datos crece exponencialmente. De hecho, crecen a la misma tasa que las fuentes computacionales, i.e. según la ley de Moore, se duplican cada 18 meses, [Hegland, 2003]. Los analistas de información, administradores de conocimiento, científicos, etc., mencionan que almacenar datos implica un *costo* y por lo regular no se obtienen beneficios directos de esta práctica. En la tabla 1 se muestra el volumen que llegan a tener algunas colecciones de datos almacenados, [Jorge, 2004].

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
500 terabytes	Todas las bases de datos y bibliotecas científicas en Alemania
1 petabyte	Todos los datos de los viajes espaciales desde sus comienzos
20 petabytes	Casi tres años de televisión no interrumpida
2 exabytes	El volumen de datos digitales generados mundialmente en un año

Tabla 1: El volumen de algunas colecciones de datos.

[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.

1.1.-Reseña Histórica

El campo del Descubrimiento de Conocimiento en Bases de Datos se concibe como la evolución natural de las técnicas de análisis de datos. En la ilustración 1 se muestra la evolución del Descubrimiento de Conocimiento en Bases de Datos.

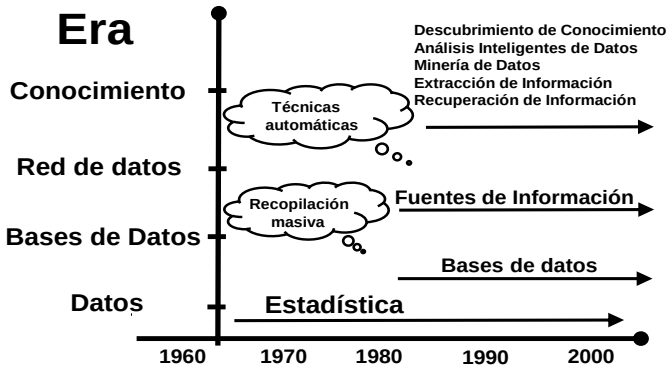


Ilustración 1: Evolución del Descubrimiento de Conocimiento en Bases de Datos.

Básicamente, dos factores estimularon dicha evolución: Las mejoras tecnológicas de las bases de datos que hicieron posible la *recopilación masiva* de datos y el desarrollo de técnicas de *aprendizaje inteligentes*[2] en la década de los ochenta. La primer era se denomina *"la era de datos"*, en la cual, los análisis de datos se realizaban sobre algún conjunto de datos utilizando técnicas estadísticas. La segunda era se denomina "*la era de las bases de datos",* en la cual, los análisis de datos empiezan a realizarse sobre cantidades considerables de datos. La característica de esta era es la recopilación masiva de datos. A la par del surgimiento de Internet se desarrolla "*la era de la red de datos".* En esta era se empiezan a desarrollar los *Sistemas de Administración de Bases de Datos*[3] que hacen posible tener acceso a datos provenientes de distintas bases. A través de estos sistemas los análisis de datos disponen de diversas fuentes de datos. La Inteligencia Artificial desarrolla algunas de 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. Las técnicas de análisis de información[4] que se vienen utilizando para tal fin empiezan a tener problemas con el procesamiento de estas grandes cantidades de datos. Para una definición de *Análisis de Información,* vea el Apéndice A.

Fayyad y Piatetsky-Shapiro [Fayyad et al., 1996B] definen al Descubrimiento de Conocimiento en Bases de Datos, DCBD, como: *el proceso no trivial de identificar válidos, novedosos, potencialmente útiles y finalmente entendibles patrones en 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 representa *conocimiento*.

[2] Aprendizajes inductivos y aprendizajes deductivos.
[3] Algunos ejemplos de sistemas son los Sistemas Relacionales, los Sistemas Columna-Orientado, los Sistemas Objecto, los Sistemas Objecto-Relacional, etc.
[4] La Arqueologia de Datos, La Recuperación de Información, La Extracción de Información, El Procesamiento del Lenguaje Natural, El Análisis Inteligente de Datos, La Minería de Datos, etc.

El término *patrón* va más allá de su sentido tradicional incluye modelos y estructuras en datos. En esta definición *datos* comprenden un conjunto de hechos (es decir, casos en una base de datos) y *patrón* es una expresión en algún lenguaje que describe un subconjunto de los datos (o un modelo aplicable a ese subconjunto). Entonces, entendemos que extraer un patrón también designa ajustar un modelo a datos, encontrar estructura de datos, o en general, alguna descripción de alto-nivel de un conjunto de datos, [Fayyad et al., 1996B]. 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. Un patrón que es interesante (de acuerdo a una medida de interés impuesta por el usuario) y cierto (de nuevo de acuerdo al criterio del usuario) se denomina conocimiento. La salida de un programa que monitorea el conjunto de hechos en una base de datos y produce patrones en este sentido es *conocimiento descubierto* [Frawley et al., 1992].

El Descubrimiento de Conocimiento en Bases de Datos se enfoca en el aspecto de encontrar entendibles patrones que pueden ser interpretados como útiles o conocimientos interesantes y pone un fuerte énfasis en trabajar con grandes conjuntos de datos, así las propiedades escalares de los algoritmos para grandes conjuntos de datos son de interés fundamental [Fayyad et al., 1996 B]. Los investigadores Fayyad y Piatetsky-Shapiro resaltan que con el Descubrimiento de Conocimiento en Bases de Datos: *The knowledge, is the end product of a data driven discovery*.

Este proceso requiere utilizar técnicas de otros campos de investigación como la Estadística, Inteligencia Artificial, Descubrimiento Científico Automatizado, Conjutos Fuzzy y Rough, Visualizacion de Información, Reconocimiento de Patrones, Aprendizaje Maquina, Sistemas Expertos, Bases de Datos, Ciencias de la Computación, Computación Evolutiva, etc., para la extracción exitosa de conocimiento. En la ilustración 2 se muestran algunos campos que tienen relación con el Descubrimiento de Conocimiento en Bases de Datos.

Ilustración 2: Algunos campos relacionados con el DCBD.

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

- Las bases de datos no ofrecen técnicas que permitan transformar los *datos brutos*[5] 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 la Estadística y del Aprendizaje Maquina, es el volumen datos que maneja, la complejidad de los datos y los resultados esperados más que los métodos particulares y algoritmos usados.

[5] Los datos que se obtuvieron directamente de la fuente de información.

- Mientras que la Inteligencia Artificial, Aprendizaje Maquina, 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, [Frawley et al., 1992].
- No se debe confundir al descubrimiento de conocimiento con los *buscadores*[6] 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*.

1.2.-El Diseño del Proceso DCBD

El Descubrimiento de Conocimiento en Bases de Datos consiste en una serie de etapas sistemáticas que comúnmente se denomina *Proceso DCBD*[7]. Los elementos claves del proceso son: conocimiento previo, datos, patrones descubiertos y conocimiento, métodos de evaluación para los patrones descubiertos y la colección de operaciones asociadas con las diferentes etapas del proceso, [Williams et al., 1996]. El modelo básico consiste de: *la base de datos D, la representación del conocimiento L, la evaluación de los patrones S y las operaciones E.*

Antes de diseñar un proceso DCBD es recomendable que los analistas entiendan perfectamente el contexto de la aplicación, objetivos y limitaciones del diseño, detectar conocimientos a priori relevantes y lo más importante entender las necesidades de los potenciales clientes. En lo que resta de esta sección, se explica brevemente algunas de las funciones que desempeñan estos elementos en el diseño del proceso DCBD.

La Base de Datos *D.* Una vez que se han seleccionado las *fuentes de información*[8], se necesitan bases de datos para almacenar los conjuntos de datos que se extraigan de éstas. Estos datos son la base para el resto del proceso DCBD. En el mercado existen una enorme variedad de *modelos de bases de datos*[9] para realizar esta tarea, [Dunkel et al., 1997]. Hay que tener en cuenta que cada modelo de base de datos ofrece capacidades distintas de almacenamiento, conectividad a otras bases de datos, precio, sistemas de administración de bases de datos, etc. Aunque las bases de datos pueden contener muchos tipos de datos, algunos de ellos se encuentran protegidos por las leyes de varios países. Por ejemplo en España, los datos personales se encuentran protegidos por la Ley Orgánica de Protección de Datos de Carácter Personal (LOPD), [Wikipedia].

La Representación del Conocimiento *L.* Básicamente, los tipos de conocimiento (i.e., patrón, modelo o estructura) que más se desean descubrir o verificar por los analistas son: clasificación, regresión, conglomerados, resumen, modelado de dependencias, análisis de secuencias, cambio y detección de desviaciones, y la asociación. La representación del conocimiento (i.e. algoritmos o técnicas de la minería de datos) determina la flexibilidad del patrón en representar los datos y la interpretabilidad del patrón en términos humanos. Típicamente, patrones complejos pueden ajustarse mejor a los datos pero pueden ser difíciles de entender. Mientras que los investigadores tienden a defender patrones complejos, los practicantes frecuentemente utilizan en sus aplicaciones patrones simples debido a su robustez e interpretabilidad, [Fayyad et al., 1996C].

---

[6] Por ejemplo, *Google.*

[7] Abreviación de *Proceso para el Descubrimiento de Conocimiento en Bases de Datos*.

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

[9] Por ejemplo, las bases de datos jerárquicas, las base de datos deductivas, las bases de datos relacionales, las bases de datos transaccionales, las bases de datos de objeto-orientado, las bases de datos de objeto-relacional, las bases de datos deductivas, las bases de datos paralelas, etc.

Algunas representaciones como los árboles de decisión se utilizan fundamentalmente para la clasificación y el conglomerado; las reglas se utilizan para la clasificación; las taxonomias (jerarquías) poseen capacidades predictivas que pueden tolerar valores perdidos. En las áreas de Inteligencia Artificial y Aprendizaje Maquina se han desarrollado algoritmos de aprendizaje automático como las redes neuronales artificiales, lógica difusa, algoritmos genéticos y combinaciones entre ellos con aplicaciones en clasificación y conglomerado. Su mayor inconveniente es la mala inteligibilidad de los resultados, aunque algunas nuevas combinaciones y técnicas permiten extraer reglas a partir de los modelos inducidos.

La noción de una "regla" como una representación abstracta de un concepto en la mente humana vino a ser cuestionada por sicólogos y aun no hay una teoría que explique satisfactoriamente como almacenamos conceptos. Esto ha determinado la naturaleza de los algoritmos de aprendizaje en Aprendizaje Maquina, [Michie et al, 1994]. Los algoritmos se dividen en tres clases:

- *Algoritmos de Aprendizaje Basados en Casos.* Los conceptos se aprenden a través del almacenamiento de casos prototipos de conceptos y no se construyen representaciones abstractas.
- *Algoritmos de Aproximación de Función.* Incluyen métodos conexionistas y estadísticos. Estos algoritmos están más cercanos a las nociones matemáticas de aproximación e interpolación y representan conceptos como formulas matemáticas.
- *Algoritmos de Aprendizaje Simbólico*. Los conceptos se aprenden construyendo una simbología, la cual describe una clase de objetos. Consideramos algoritmos que trabajen con representaciones equivalentes a la lógica preposicional o lógica de primer orden.

Cada clase enfrenta el problema de la representación de forma distinta. Es importante considerar que algunas representaciones afectan la velocidad de aprendizaje, la legibilidad de la descripción del concepto, etc., por lo tanto, representaciones incorrectas nos conducen a formas incorrectas de conocimiento.

La Evaluación de los Patrones *S.* El proceso de extracción a través de técnicas inteligentes (aprendizajes supervisados o no supervisados entre otros) que van examinando en forma automatizada los datos *arrojan* una enorme cantidad de patrones. La identificación de patrones que realmente proporcionen un conocimiento es una tarea que una persona difícilmente lograría. Algunos investigadores han propuesto medidas que evalúen patrones descubiertos según su utilidad y relevancia para el usuario. Estas se denominan *medidas de interés*. Las medidas de interés son filtros que solamente permiten el paso de patrones que cumplan con las restricciones impuestas por el usuario. De este modo se reduce el número de patrones que son considerados *patrones interesantes* por el usuario. Formalmente, una medida de interés *F* es una función que *mapea* un conjunto de proposiciones expresadas en *L* a un conjunto de valores numéricos (usualmente). Estas medidas se dividen en *objetivas y subjetivas*. Las medidas objetivas dependen solamente de la estructura del patrón y de los datos subyacentes usados en el proceso del descubrimiento. Mientras que las medidas subjetivas dependen del usuario que examina los patrones. Cada usuario diferencia lo que es un *patrón interesante* según su experiencia, conocimiento, intereses, creencias, etc. Es por ello, que las medidas subjetivas se enfocan a obtener dos tipos de patrones: patrones inesperados y patrones procesables, [Silberschatz et al., 1996]. Los patrones inesperados son precisamente aquellos patrones que hacen que el usuario diga *imposible, no lo creo, etc.* Los patrones procesables son los que permiten al usuario obtener algún beneficio. En la tabla 2 se muestran algunos ejemplos de medidas de interés, [Hilderman et al, 1999].

Nombre	Fundamento	Representación	Clase
Función Regla-Interés de Piatetsky-Shapiro	Probabilística	Regla de Clasificación	Objetiva
Medida J de Smyth-Goodman	Probabilística	Regla de Clasificación	Objetiva
Medida Itemset de Agrawal-Srikant	Probabilística	Regla de Asociación	Objetiva
Regla-Plantilla de Klemettinen.	Sintáctica	Regla de Asociación	Subjetiva
Projected Savings de Matheus-Piatetsky-Shapiro	Utilitaria	Resumen	Subjetiva
Interés de Silbershatz-Tuzhilin	Probabilística	Formato – Independiente	Subjetiva

Tabla 2: Algunas medidas de interés para la evaluación de patrones.

La *Función Regla-Interés* propuesta por Piatetsky – Shapiro, por ejemplo, se usa para cuantificar la correlación entre los atributos de una simple regla de clasificación. La regla de clasificación se refiere a la implicación lógica $X \Rightarrow Y$, la cual se interpreta como: a un atributo (lado derecho) le corresponde otro atributo (lado izquierdo). La Función Regla -Interés se define como:

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

Donde N es el número total de atributos. $|X|$ y $|Y|$ son el número de atributos que satisfacen la condición X y Y respectivamente. $|X \cap Y|$ es el número de atributos que satisfacen $X \Rightarrow Y$. Mientras $|X||Y|/N$ es el número de atributos esperados si X y Y son 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 2×2.

Los expertos en el descubrimiento de conocimiento en bases de datos coinciden en lo siguiente:

- Con estas medidas 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; interesantes subjetivamente pero no objetivamente.
- Se pueden obtener patrones inesperados pero que también sean procesables aunque patrones procesables no necesariamente son inesperados.

Como se aprecia, las medidas de interés son bastante complejas y de aplicación específica. En consecuencia, los expertos en la extracción de conocimiento recomiendan considerar conceptos simples como *la validez, lo novedoso, lo potencialmente útil y la comprensibilidad* para detectar patrones interesantes. La *validez* se refiere aquellos patrones que son válidos en nuevos conjuntos de datos o en conjuntos de prueba, bajo algún grado de certeza. Lo *novedoso* se refiere a que los patrones descubiertos, tal vez, no sean los típicos patrones que siempre maneja el usuario. Lo *potencialmente útil* se refiere a los patrones que permiten obtener cierto beneficio al usuario. Por último, la *comprensibilidad* se refiere a que los patrones descubiertos 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, [Fayyad et al., 1996B].

Las Operaciones *E.* Como ya se ha mencionado, el descubrimiento de conocimiento consta de una serie de etapas sistemáticas, que comúnmente reciben el nombre de *Proceso DCBD*. La forma genérica del proceso consta de: *una etapa de preprocesamiento, una etapa de extracción de patrones y una etapa de posprocesamiento.* En cada etapa se realizan operaciones

o tareas que van desde la obtención de los datos hasta la presentación del conocimiento. La ilustración 3 muestra la iteración e interacción que existe en esta serie de etapas sistemáticas. Algunas de las características de esta iteración e 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 DCBD puede llegar a ser compleja dependiendo del contexto de la aplicación.
- Es necesario tener expertos en cada etapa, operación, etc.

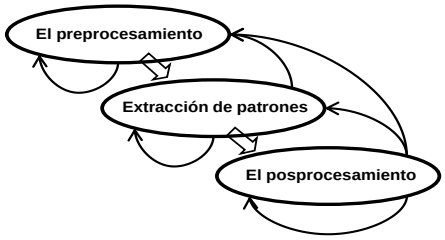


Ilustración 3: El proceso DCBD.

1.- La Etapa de Preprocesamiento. Los datos extraídos de las fuentes de información generalmente no son aptos para aplicarles las técnicas de la minería de datos. El preprocesamiento consiste en la transformación de los *datos brutos* en datos confiables, [Kamber, 2001].

Ilustración 4: Las operaciones en la etapa de preprocesamiento.

Entre las operaciones más comunes que se realizan para este propósito están: la limpieza, la integración, la selección, la transformación y la reducción de los datos. Las técnicas de la minería de datos extraen patrones de estos datos confiables o conjuntos de datos confiables. Además, algunas técnicas de minería requieren datos de entrenamiento que permitan ajustar o calibrar los parámetros antes de extraer patrones [Holsheimer et al, 1991]. En la ilustración 4 se muestran algunas propiedades de los datos confiables. Realizar el preprocesamiento brindará un cierto grado de confiabilidad al conocimiento obtenido. De todas las etapas, esta consume hasta un 50% del tiempo. 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 confiable 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 conjunto confiable o subconjunto de entrenamiento más relevante para la extracción de patrones.
- *Transformación*: Algunos técnicas de la minería de datos que se emplean en la extracción de patrones requieren representaciones apropiadas de los conjuntos confiables o de los subconjuntos de entrenamiento.
- *Reducción*: Algunas veces es posible reducir a formas más compactas al conjunto de datos confiable o al subconjunto de entrenamiento.

Los conjuntos de datos confiables se caracterizan por ser *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.

El Análisis Exploratorio de Datos es clave en esta etapa. La finalidad del análisis exploratorio consiste en 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. En la ilustración 5 se muestran la tipología de las denominadas Técnicas de Preprocesamiento de Datos.

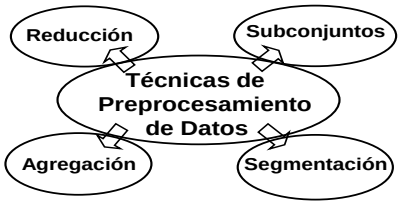


Ilustración 5: Técnicas de preprocesamiento de datos.

- *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 gráficas de histogramas, de pastel, de barras, de líneas, de burbujas, etc.

Las técnicas de preprocesamiento[10] requieren de herramientas matemáticas sofisticadas como el análisis de Fourier, la teoría de ondículas, las curvas principales, el método de Monte Carlo, regresiones lineales y no lineales, etc. Además, es importante emplear gráficas de dispersión, gráficas de barra, histogramas, curvas de Andrews, caras de Chernoff, etc., para detectar tendencias y desviaciones en datos.

En esta etapa es recomendable tener la capacidad de manipular los datos: Se deberá hacer poca o ninguna suposición sobre los procesos estadísticos que generaron los datos. Esto es muy importante porque frecuentemente no tendremos un conocimiento a priori de los datos; Cuando se explore el conjunto o conjuntos de datos por primera vez se debe ser capaz de realizar un amplio rango de análisis que nos permitan ganar rápidamente conocimiento sobre ellos y así determinar el modelo adecuado a utilizar; Siempre es necesario actualizar o recalibrar el modelo cuando se reciban datos adicionales; El analista debe preferir conjuntos pequeño y compacto de datos que le permitan explorarlos manualmente y así pueda hacer inferencias intuitivas sobre asociaciones, clasificaciones, conglomerados, etc.

2. La Etapa de Extracción de Patrones. La extracción[11] se considera como el *motor* del proceso DCBD por los expertos en el descubrimiento de conocimiento. Una vez que se tengan los datos confiables en una base de datos, el analista está en condiciones para decidir qué tipos de conocimiento (i.e., patrón, modelo o estructura) quiere descubrir o verificar. Es importante destacar que la elección del tipo de conocimiento que se desea extraer va a marcar claramente la técnica de minería de datos a utilizar. Los tipos de conocimiento a descubrir o verificar son: [Fayyad et al., 1996B], [Fayyad et al., 1996D].

- *Clasificación:* Se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto predefinido de clases). Por ejemplo, clasificar tendencias en mercados financieros.
- *Regresión:* Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable. Por ejemplo, en el contexto de la investigación de mercados puede utilizarse para predecir el número de ventas de un determinado producto.
- *Conglomerado:* Hace 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. Por ejemplo, descubrir subpoblaciones homogéneas de clientes en bases de datos comerciales.
- *Modelado de dependencias:* Una dependencia (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Uno de los mayores problemas de la búsqueda de dependencias es que suelen existir muchas dependencias nada interesantes y en las que la causalidad es justamente la inversa. Por ejemplo el hecho de que un paciente haya sido ingresado en maternidad determina su sexo.
- *Análisis de secuencias:* Se intenta modelar la evolución temporal de alguna variable, con fines descriptivos o predictivos.
- *Cambios y detección de desviaciones:* Se enfoca en descubrir los cambios más significativos en los datos usando valores normalizados o medidas previas.
- *Asociación:* Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta. Es uno de los patrones con más interés comercial, sobre todo en el análisis de hábitos de los clientes, donde, por ejemplo, en un supermercado se analiza si los pañales y los chupones de bebé se compran conjuntamente.

[10] Según el tipo de dato, ya sea, nominal, continuo, intervalos, categórico, etc., seleccionamos la herramienta matemática.
[11] Denominada Minería de Datos.

- *Resumen:* Se obtienen representaciones compactas para subconjuntos de los datos de entrada. Por ejemplo, análisis interactivo de datos, generación automática de informes, visualización de datos, etc.
- *Reglas Generales:* Evidentemente muchos patrones no se ajustan a los tipos anteriores. Recientemente los sistemas DCBD incorporan capacidad para establecer otros patrones más generales.

Afortunadamente, los propios sistemas DCBD se encargan generalmente de elegir la técnica más idónea entre las disponibles para un determinado tipo de patrón a buscar. Según la manera de representar los patrones, podemos distinguir entre *Técnicas de Minería de Datos no Simbólicas* y *Técnicas de Minería de Datos Simbólicas.*

Las técnicas de minería de datos no simbólicas, generalmente son más apropiadas para variables continuas. Entre las técnicas no simbólicas, podemos destacar: algunas técnicas estadísticas se pueden utilizar para confirmar asociaciones y dependencias, y para realizar conglomerados. Técnicas muy importantes son las regresiones lineales, no lineales y redes de regresión para establecer tendencias; Vecinos más próximos y razonamiento por casos se aplican generalmente a clasificación y conglomerado, basándose en medidas de distancias o similitud con el prototipo o los miembros de los grupos; Las redes neuronales artificiales, lógica difusa, algoritmos genéticos y combinaciones entre ellos son técnicas tradicionales de aprendizaje automático con aplicaciones especialmente en clasificación y conglomerado. Su mayor inconveniente es la mala inteligibilidad de los resultados, aunque algunas nuevas combinaciones y técnicas permiten extraer reglas a partir de los modelos inducidos, [Holsheimer et al, 1991]. Además, existen los algoritmos específicos que son algoritmos eficientes para la búsqueda de asociaciones o dependencias. Por ejemplo, en un supermercado con miles de artículos, no se puede evaluar estadísticamente cada uno de los posibles pares o tripletes de combinaciones entre productos. Se han diseñado algoritmos (específicos) que permiten buscar todas las asociaciones significativas existentes eficientemente, por ejemplo, Apriori, AprioriTid, AIS, SETM, etc. [Agrawal et al., 1994].

El mayor inconveniente de las técnicas no simbólicas es su poca (o nula) inteligibilidad. En el caso del razonamiento por casos o las redes neuronales, el resultado del proceso es una caja negra que sirve para predecir o clasificar nuevos casos, pero no se sabe cómo y, por tanto, no se ha obtenido conocimiento. Por el contrario, las técnicas simbólicas generan un modelo "legible" y además aceptan mayor variedad de variables y mayor riqueza en la estructura de los datos. Entre las técnicas de minería de datos simbólicas, podemos destacar: Los árboles de decisión son utilizados fundamentalmente para la clasificación y conglomerado, consisten en una serie de *tests* que van separando el problema, siguiendo la técnica del divide y vencerás, hasta llegar a las hojas del árbol que determinan la clase o grupo a la que pertenece el registro o individuo. Existen muchísimas técnicas para inducir árboles de decisión, siendo el más famoso el algoritmo C4.5 de Quinlan. Los árboles de regresión son similares a los árboles de decisión pero basados en técnicas estadísticas.

La programación inductiva y otras técnicas de inducción simbólica de alto nivel se usan fundamentalmente para obtener patrones de tipo general, que se pueden establecer entre varios individuos o son intrínsecamente estructurales. Aunque existen algunas aproximaciones basadas en reglas simples, es la programación lógica inductiva (ILP) el área que ha experimentado un mayor avance en la década de los noventa. ILP se basa en utilizar la lógica de primer orden para expresar los datos, el conocimiento previo y las hipótesis. Como la mayoría de bases de datos actuales siguen el modelo relacional, ILP puede trabajar directamente con la estructura de la misma, ya que una base de datos relacional se puede ver como una teoría lógica. Aparte de esta naturalidad que puede evitar o simplificar la fase de preprocesado, ILP permite representar hipótesis o patrones relacionales, aprovechando y descubriendo nuevas relaciones entre individuos. Por ejemplo no tiene sentido enviarle propaganda de albercas a una persona si ésta convive con otra que ya instaló una alberca recientemente. Estos patrones son imposibles de

expresar con representaciones clásicas. Nótese que un árbol de decisión siempre se puede convertir fácilmente en un conjunto de reglas pero no viceversa. En definitiva, existen multitud de técnicas, combinaciones y nuevas variantes que aparecen recientemente, debido al interés del campo. Así, los sistemas de DCBD se afanan por incorporar la mayor cantidad de técnicas, así como ciertas heurísticas para determinar o asesorar al usuario sobre qué métodos son mejores para distintos problemas.

La minería de datos involucra ajustar modelos o determinar patrones de los datos. Los modelos ajustados juegan un papel importante en la inferencia de conocimiento. Decidir si los modelos reflejan conocimiento útil es una parte de la interacción total del proceso DCBD para el cual una valoración subjetiva humana es usualmente requerida. La idea de "ajuste" requiere que consideremos: *Un modelo de representación:* es el lenguaje L para describir los patrones descubiertos. El modelo contiene parámetros que han de fijarse a partir de los datos de entrada. Si el modelo de representación es limitado entonces ningún tiempo de entrenamiento o ejemplos producirán modelos exactos para los datos. Por ejemplo, un modelo de clasificación basado en árboles de decisión suele utilizar un *algoritmo greedy* (una búsqueda sin vuelta atrás) y una heurística que favorezca la construcción de árboles de decisión con pocos nodos. *Una evaluación del modelo:* la evaluación usa algún criterio de preferencia que sirva para comparar modelos alternativos. *Los métodos de búsqueda:* consiste en buscar modelos y sus parámetros. En la búsqueda de parámetros, el algoritmo busca los parámetros que optimicen los criterios de preferencia y el modelo de representación estando estos fijos. Mientras que en la búsqueda del modelo se rotan los parámetros, así la representación del modelo está cambiando de tal forma que una familia de modelos es considerada. [Fayyad et al., 1996B].

3. La Etapa de Posprocesamiento. El conjunto de patrones interesantes obtenidos en la etapa anterior deben ser interpretados para su uso. Interpretar los patrones interesantes incluye analizarlos, visualizarlos, remover redundancias e irrelevancias, quizás regresar a las etapas anteriores para detectar posibles errores, así como describir su utilidad en términos comprensibles por los usuarios. El conocimiento descubierto se debe incorporar a un sistema de ejecución, tomar acciones basadas en este conocimiento, o simplemente documentarlo y reportarlo hacia las partes interesadas, así como checar y resolver conflictos potenciales con extracciones anteriores de conocimiento.

Las técnicas de visualización de datos han tomado mucha importancia en el campo del Descubrimiento de Conocimiento en Bases de Datos. Las técnicas de visualización de datos se utilizan fundamentalmente con dos objetivos: en primer lugar aprovechar la gran capacidad humana de extraer patrones a partir de imágenes y en segundo lugar, ayudar al usuario a comprender más rápidamente patrones descubiertos automáticamente por un sistema DCBD. Nótese que la visualización no es sustituta del análisis cuantitativo, [Fayyad et al., 2002].

La visualización 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. La visualización transforma la información original en información más significativa, a partir de la cual el usuario puede ganar en comprensión. El *discernimiento*[12] ganado por la visualización hace que el usuario elabore decisiones y tienda a explicar el comportamiento de los datos. Aunque las técnicas no son excluyentes, estos dos objetivos marcan dos momentos diferentes del uso de la visualización de los datos: visualización previa contenida dentro del análisis exploratorio de datos y visualización posterior al proceso de minería de datos.

La visualización previa se utiliza para entender mejor los datos y sugerir posibles patrones o qué tipo de técnicas de minería de datos utilizar. Se utiliza frecuentemente por analistas para ver tendencias, resúmenes de los datos e irregularidades que investigar. Un

[12] Distinción.

ejemplo típico de estas visualizaciones previas es el conglomerado mediante funciones de densidad, generalmente representadas tridimensionalmente, donde los seres humanos ven claramente los conglomerados que aparecen con distintos parámetros. También es muy común para encontrar asociaciones el uso de gráficos bidimensionales donde en las abscisas y ordenadas están todos los factores y en la intersección se muestran las frecuencias (utilizando color, puntos gordos o una tercera dimensión) de cada par de factores. Para detectar valores extraños y anomalías (Outliers) se utilizan gráficos especializados, algunos de ellos tradicionales en Estadística son los diagramas de dispersión. Sin embargo, el mayor problema de la visualización de datos es que la información en almacenes de datos suele ser multidimensional, siendo además el número de dimensiones mucho mayor que 3. El objetivo es por tanto conseguir proyectar las dimensiones en una representación en 2 (ó 3 simuladas) dimensiones, que son las únicas que pueden ser representadas en la pantalla de una computadora o en papel.

La proyección geométrica es muy popular. La idea consiste en mapear el espacio *k*-dimensional en dos dimensiones mediante el uso de *k* ejes de ordenadas (escalados linealmente) por uno de abscisas. Cada punto en el espacio *k*-dimensional se hace corresponder con una línea poligonal (polígono abierto), donde cada vértice de la línea poligonal interseca los *k* ejes en el valor para la dimensión. Cuando hay pocos datos cada línea se dibuja de un color. Cuando hay muchos datos se utiliza una tercera dimensión para los casos. El mayor problema de estas representaciones (y de otras muchas) es que no acomodan bien las variables discretas. En este sentido, existen otro tipo de técnicas que sí permiten combinar atributos continuos y discretos, mediante el uso de transformaciones menos estándar y el uso de iconos. Se utilizan rasgos compatibles y diferenciados para distintas dimensiones, como son círculos, estrellas, puntos, etc., con la ventaja de que se pueden combinar más convenientemente valores discretos y continuos. Otras aproximaciones más sofisticadas se basan en estructuras jerárquicas, por ejemplo, como los árboles cono.

Por otro lado, *la visualización posterior* se utiliza para mostrar los patrones y entenderlos mejor. Un árbol de decisión es un ejemplo de visualización posterior. Otros gráficos de visualización posterior de patrones son los que muestran un determinado conglomerado de los datos, una asociación, una determinada clasificación, utilizando para ello gráficos de visualización previa en los que además se señala el patrón.

1.3.-El Sistema DCBD

El *sistema DCBD*[13] permite a los usuarios interactuar con los datos en todas las etapas del proceso DCBD. Un sistema DCBD comprende una colección de componentes que juntos pueden identificar eficientemente y extraer interesantes y útiles patrones nuevos de los datos almacenados en bases de datos. Los sistemas DCBD deben basarse en los siguientes principios: simplicidad, autonomía, confiabilidad, reusabilidad, disponibilidad y seguridad. Además, el sistema DCBD no debe imponer conocimiento a los usuarios, sino que debe guiarlos a través del proceso DCBD para que ellos elaboren sus hipótesis, ideas, etc.

Las técnicas de la minería de datos empleadas por el sistema DCBD se pueden clasificar en dos grandes grupos: técnicas de verificación (en las que el sistema se limita a comprobar hipótesis suministradas por el usuario) y técnicas de descubrimiento (en los que el sistema encuentra patrones potencialmente interesantes de forma automática). El resultado obtenido con las técnicas de descubrimiento puede ser de carácter descriptivo o predictivo. Las predicciones nos sirven para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción nos puede ayudar a su comprensión. De hecho, los modelos predictivos pueden ser

[13] Abreviación de *Sistema para el Descubrimiento de Conocimiento en Bases de Datos.* El sistema DCBD se refiere al sistema computacional de hardware y al software que corre en una computadora o computadoras.

descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones [Fayyad et al., 1996B].

Entre las características deseables del sistema DCBD, se encuentran: [Dunkel et al., 1997], [Matheus et al 1993].

- *Manipulación de grandes volúmenes de datos.* Se requiere de grandes cantidades de datos que proporcionen información suficiente para obtener un conocimiento adicional.
- El sistema DCBD 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.
- *Eficiencia.* Debido al volumen de datos que fluye entre cada una de las etapas del proceso DCBD, es esencial, la *eficiencia*. Un proceso DCBD mal planificado implicará que su respectivo sistema DCBD tenga problemas con la escalabilidad.
- *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. Entre estos encontramos el aprendizaje supervisado y el no supervisado entre otros.
- *Interactividad e iteración.* Debido a esta característica el sistema DCBD puede ser interpretado como: *Human - Assisted Computer Discovery* o quizás como *Computer - Assisted Human Discovery.*

Algunas técnicas visuales sofisticadas que pueden ser implementadas en el sistema DCBD, son las técnicas de proyección geométricas, las técnicas basadas en íconos, las técnicas orientadas en píxeles, las técnicas jerárquicas y basadas en gráficas, entre otras. Hay muchas clases de operaciones de interacción que pueden y deben ser integradas al proceso de visualización en los sistemas DCBD. Estas incluyen operaciones de selección de datos (los conjuntos de datos deben ponerse en alguna resolución en particular), operaciones de manipulación de datos (suavidad, filtrado, interpolado, etc.), operaciones de representación (modificar atributos específicos tales como el color), orientación de las vistas (fondos, zooms, ventanas) e interacciones en la visualización (navegar o realizar selecciones vía manipulaciones directas de los elementos mostrados en pantalla). Una clase muy potente de operaciones de interacción son las *"probes"* que permiten al usuario aislar una sección de los datos a ser visualizados y presentarlos en una visualización secundaria.

Los expertos en el descubrimiento de conocimiento recomiendan que antes de diseñar algún proceso y sistema DCBD se tomen en cuenta los siguientes criterios [Fayyad et al., 1996B]: *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 previo[14], etc. Además, algunos problemas que pueden surgir durante el desarrollo del sistema DCBD son: entrenamiento insuficiente de los métodos de aprendizaje inteligente; herramientas de soporte inadecuadas; abundancia de patrones; cambios rápidos de los datos en el tiempo; datos complejos (espaciales, imágenes, texto, audio, video, etc.).

1.4.-Tareas del Sistema DCBD

Los profesionales de los más diversos campos del conocimiento utilizan y desarrollan todo tipo de sistemas DCBD. Estos sistemas son usados por las compañías para mejorar sus

[14] Domain Knowledge.

negocios y redefinir sus estrategias, [Barry, 2001], [Delmater et al., 2001]. 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 qué tipo de mercancía convertirá, por ejemplo, a los clientes que compran sólo novedades en clientes que compran todo tipo de mercancía.*

En el mercado, existen una gran variedad de sistemas DCBD, cuyo objetivo es dar soporte a cada etapa del proceso DCBD. Entre los sistemas DCBD más utilizados para realizar algunas tareas son: 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 DCBD denominado CRISP–DM. Utiliza una gama enorme de técnicas de minería de datos para la extracción de patrones en forma automática e inteligente. *MineSet* es desarrollado por Silicon Graphics Inc. Proporciona múltiples técnicas de minería de datos para la clasificación y asociación. La característica distintiva de *MineSet* es su variedad de herramientas de visualización, en todas las etapas del proceso DCBD. *IBM Intelligent Miner* proporciona técnicas de minería de datos para la asociación, clasificación, modelado de dependencias, análisis de secuencias, conglomerados, etc. La característica distintiva de *IBM Intelligent Miner* es su integración al sistema de bases de datos DB2 (desarrollado también por IBM) y su excelente escalabilidad de las técnicas de minería de datos. Algunas tareas de estos sistemas DCBD son mostradas en la tabla 3.

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) 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 3: Algunas tareas de los Sistemas DCBD.

2 Cienciometría

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 información y comunicación. Felizmente, también estas tecnologías complementan a otras técnicas y metodologías, brindando la oportunidad de hacerle frente a tales desafíos.

La Cienciometría se inserta en este marco como una disciplina relativamente nueva. Aunque su gestación se inició en el primer cuarto del siglo XX, su auge es reciente. La Cienciometría ha capitalizado la esencia de la cuantificación de la actividad de investigación científica y técnica. En su desarrollo recientemente la Cienciometría 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 Cienciometría ha encontrado en el Descubrimiento de Conocimiento en Bases de Datos 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 Cienciometría 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 Cienciometría como parte de la metodología de investigación, como medio para hacer a la investigación científica y al desarrollo tecnológico más productivos y eficientes.

2.1.-Historia

El término Cienciometría se utiliza para designar un conjunto de trabajos iniciados hace unos treinta años y que están todos, por distintos conceptos, consagrados al análisis cuantitativo de la actividad de investigación científica y técnica. La Cienciometría debe estudiar, por consiguiente, tanto los recursos y los resultados como las formas de organización en la producción de los conocimientos y técnicas. Sin embargo, hasta una fecha reciente se ha ocupado casi exclusivamente del análisis de los documentos redactados por los investigadores y tecnólogos. La originalidad de los resultados obtenidos de esta forma demuestra la pertinencia de la elección. Nos limitaremos, pues, en lo esencial, a esta concepción restringida de la Cienciometría, [Callon et al., 1995].

La Cienciometría como se le conoce actualmente, [Gurjeva, 1992] es el resultado de la convergencia de dos movimientos. En la antigua URSS se desarrolló el movimiento *Naukovodemia*[15], con el objetivo de estudiar científicamente la 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*, [Gurjeva, 1992] término acuñado por Nalimov y Dobrov[16]. Mientras que en 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 sus obras tituladas

[15] Nalimov dirige la escuela de Moscú mientras que Dobrov dirige la escuela de Kiev. Ambos tienen posturas distintas sobre de lo que posteriormente se llamará la Ciencia de la Ciencia.
[16] Se cree que Nalimov y Dobrov fueron influenciados por el trabajo de Solla Price.

"Science since Babylon" y "Little Science, Big Science" describe la evolución de movimiento científico desde sus orígenes y propone un marco teórico para estudiar a la ciencia.

La Cienciometría 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. La Cienciometría 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, [Vanti, 2000]. La Bibliometría 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. Tradicionalmente se le atribuye a Alan Pritchard la paternidad del término Bibliometría a partir de su trabajo de 1969 llamado *¿Bibliografía Estadística o Bibliometria?* Sin embargo, fue Otlet, 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, [Vanti, 2000].

El término Informetría comenzó a emplearse en el campo de las Ciencias de la Información a partir de la década del ochenta. En 1979, Otto Nacke el director del "Institut fur Informetrie" de Alemania propuso el término, [Vanti, 2000]. Al principio sólo se le reconoció como un campo general de estudio que incluía elementos de la Bibliometría y la Cienciometría, surgidas con anterioridad. La Informétria es una disciplina instrumental de las Ciencias de la Información, su objeto de estudio son los datos (información), la información social, que se obtiene y utiliza en todos los campos de la actividad del hombre, los procesos del pensamiento creador para la generación y utilización de la información social, los procesos de presentación, registro, procesamiento, conservación, búsqueda, diseminación y percepción de la información, el papel y el lugar de las fuentes de información (documentales y no documentales) en la sociedad, el desarrollo humano y el nivel de informatividad del hombre en la sociedad, los procesos socio-tecnológicos de informatización de la sociedad y la orientación humanista de la informatización.

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, [Macias-Chapula, 1998]. En la tabla 4 se muestra la tipología de estudio de la Bibliometria, la Cienciometria y la Informetria.

Las revistas que publican trabajos sobre Bibliometría, Cienciometría e Informetría son: 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 Française de Bibliométrie, Cahiers de la Societé Française de Bibliométrie Appliquée; Science & Public Policy, Scientometrics, Social Studies of Science.

Tipología	Bibliometría	Cienciometría	Informetría
Objeto de Estudio	Libros, documentos, revistas, artículos, autores y usuarios.	Disciplinas, temas, campos científicos y tecnológicos.	Palabras, documentos, bases de datos.
Variables	Números en circulación, citas, frecuencia de aparición de palabras, longitud de las oraciones, etc.	Aspectos que diferencian a las disciplinas y las subdisciplinas. Revistas, autores, trabajos, formas 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 rector-espacio, modelos boléanos de recuperación, modelos probabilísticos, lenguaje de procesamiento, enfoques basados en el conocimiento, tesauros.
Objetivos	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 4: Tipología de estudio de la Bibliometría, la Cienciometría y la Informetría.

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)* 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 y análisis cienciométricos 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. 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. 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. El estudio de la ciencia por medio de análisis cuantitativos, requiere de modelos teóricos que explique su dinámica dentro de la comunidad científica y fuera de ella, es decir, en la sociedad. En la sección siguiente se exhibe en forma general, El Modelo Cienciométrico de la Ciencia y posteriormente se exhibe el modelo teórico denominado Actividad Científica que nos permitirá comprender el funcionamiento de la ciencia en la sociedad.

2.2.-El Modelo Cienciométrico

J. D. de Solla Price al abogar por una *Ciencia de la Ciencia*[17], amplia considerablemente la perspectiva de la Bibliometría (aplicaciones relacionadas con la gestión y uso de la información en bibliotecas). La Ciencia de la Ciencia va más lejos en la elaboración y en la aplicación de instrumentos estadísticos. Su finalidad es identificar las leyes y las regularidades que rigen la actividad científica considerada en su globalidad. Solla Price confiesa que se inspira en los modelos de la termodinámica. Trata a la ciencia como si fuera un gas del que estudia sucesivamente el volumen global (el número de investigadores y su producción), la

[17] 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 Cienciometría.

distribución de las moléculas que lo componen (los científicos) en función de su velocidad (fecundidad o productividad) y los modelos de interacción de las moléculas (las formas de organización). Su investigación [Callon et al., 1995] estadística le lleva a deducir cuatro leyes:

- A largo plazo, el volumen global de la actividad científica crece de forma regular (Existe un estado inicial, i.e. hubo precursores): el número de los investigadores y de sus publicaciones se duplica aproximadamente cada veinte años.
- Este crecimiento exponencial tiene necesariamente sus límites. Siguiendo una ley general de la naturaleza según la cual a periodos de rápido desarrollo suceden invariablemente fases de estabilización, aquel alcanzará progresivamente una nivelación (curva logística). Esta disminución de crecimiento se debe en particular a los propios límites del poder de análisis de los instrumentos empleados. El estado de crecimiento exponencial se modela con:

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

  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. Mientras que el estado de saturación se modela con:

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

  En donde *g(t)* representa el tamaño del corpus al tiempo *t* y *k* representa el límite superior.
- La comunidad científica se divide en una elite que publica la mayor parte de los artículos y en una masa de investigadores poco productivos.
- Los científicos dado que no pueden tratar mas que una cantidad limitada de informaciones, se agrupan en "colegios invisibles" que apenas cuentan con un centenar de miembros en constante interacción (estado de saturación).

Partiendo de estas observaciones. Price no duda en deducir todo un conjunto de recomendaciones destinadas a inspirar las políticas científicas dirigidas por los poderes públicos. Así, esta Ciencia de la Ciencia, desde sus origines no se limita a una pura y simple constatación, sino que penetra en el terreno de la política y de la gestión para no volver a salir de él.

La Cienciometría esta basada en un conjunto de leyes empíricas de dos tipos: distribuciones biliométricas (Zipf, Lotka y Bradford), de carácter estructural; y de leyes de envejecimiento y crecimiento dinámico (Price, Brookes y Avramescu). La ciencia puede ser estudiada de tres puntos de vista: indicadores de actividad no relacionales, indicadores relacionales de primera generación e indicadores relacionales de segunda generación. Detrás de la Cienciometría yacen varios modelos. Históricamente, los modelos más viejos son El Modelo Fractal de Mandelbrot y El Principio de Ventaja Acumulativa. Referencias anteriores a la estructura de la ciencia, la cual se asemeja a estructuras auto-similares tales como esas que también aparecen en lo social y en la naturaleza. La fractalidad envuelve una geometría que es generada por fenómenos caóticos (teoría del caos, efecto mariposa) y fenómenos complejos (teoría de la complejidad). De otro modo, lo más reciente no se refiere a estructuras pero más bien refleja un concepto generalmente aceptado la acumulación a priori ayuda conferir ventaja en contra de competidores (Efecto Matthew o Ventaja de la Elite y La Hipótesis de Ortega o Ventaja del Mediocre). También, ampliamente aceptado el Modelo de Kuhn que postula que la ciencia va a través de periodos de revolución, en los cuales nuevos paradigmas son creados y periodos alternantes, en los cuales la ciencia sigue paradigmas establecidos. Con el advenimiento de lo relacional, la Teoría Actor - Red y la Teoría de Traducción son los modelos que revelan la estructura de la ciencia y su dinámica. Esas redes científicas y sus traducciones pueden ser manifestadas fundamentalmente por el Análisis de Palabras Asociadas.

La Cienciometría esta basada en muchas teorías, por lo cual, es necesario un modelo unificado. Bailón-Moreno y sus colegas en una serie de articulos han propuesto El Modelo Cienciométrico Unificado, el cual consta de siete princios: [Bailón-Moreno et al., 2005].

Principio 1. El Principio Actor – Red. La ciencia y la tecnología (Tecnociencia) constan de redes de autores según la Teoría Actor – Red de Callon y Courtial. El principio implica:

- Esas entidades que crean y modifican la Tecnociencia son denominados Actores.
- Los actores pueden estar delimitados por palabras tomadas no como unidades lingüísticas sólo como producciones verbales o escritas asociadas a la acción, en otras palabras como una clase de acción elemental (Definición Verbal).
- Los actores pueden estar delimitados según su posición relativa en la red (Definición Estratégica).
- Los actores pueden ser humanos o no humanos (investigadores, laboratorios, países, revistas científicas, temas de investigación, documentos, dispositivos de medidas, financiación, etc.). no hay distinción entre ellos, todos son igualmente necesarios para construir la Tecnociencia.
- La Tecnociencia consta de una red de actores interconectados.
- Una red no es homogénea pero si tiene áreas grandes de interconexiones denominadas Centros de Interés. Los Centros de Interés son actores.

Principio 2. El Principio de Traducción. La dinámica de la red tecnocientífica esta gobernada por la Teoría de Traducción de Latour. El principio implica entre otras cosas, lo siguiente:

- La definición verbal de los actores evoluciona con el tiempo por la imposición de modalidades positivas o negativas (Significado lingüístico de la traducción).
- La posición de los actores dentro de la red y su posición estratégica también cambia con el tiempo (Significado geométrico de la traducción).
- La traducción involucra el equilibrio entre interacciones naturales y sociales y la estabilización de los actores. La traducción depende de los intereses particulares o colectivos de los actores, sobre su ventaja estratégica y sobre su fuerza intrínseca.

Principio 3. El Principio Espacial. La traducción implica la existencia de un espacio, con componentes temporal y geométrico del tipo Hausdorff-Besicovith, el espacio dimensional es fraccionario. Como opuesto al Modelo de Mandelbrot, las condiciones de fractalidad, no son necesariamente impuestas, aunque pueden ser un caso particular.

Principio 4. El Principio de Traducción Cuantitativa. Según Ruiz-Baños, la traducción *T*, es igual a la variación en la calidad o atributos de los actores, *Q*, medido como una cantidad de acuerdo a como se mueven en el espacio de traducción. Esto es, la traducción es la derivada o gradiente de la función de calidad con respecto a las coordenadas del espacio de traslación:

$$T(x) = \frac{dQ(x)}{dx}$$

Donde *x* es una coordenada espacial (geométrica o temporal). Como la coordenada geométrica, el rango, *r*, puede ser elegido y, como la coordenada temporal, tiempo o edad, *t*. Los cuatro principios implican dos situaciones fundamentales:

a) Traducción por cambio de posición estratégica.

$$T(r) = \frac{dQ(r)}{dr}$$

b) Traducción por evolución temporal.

$$T(t) = \frac{dQ(t)}{dx}$$

Principio 5. El Principio de Traducción Composición. Alguna traducción puede ser considerada la composición de traducciones elementales asociadas en series, en paralelo o en combinación con las traslaciones. El principio es análogo al principio de composición del movimiento de Galileo o al mecanismo de las reacciones químicas.

Principio 6. El Principio Centro - Periférico. El espacio de traducción es el campo generado por un punto, que puede ser denominado centro o núcleo, en el cual los actores buscan aproximarse a su mejor posición estratégica. El concepto de centro implica la existencia de un periférico de acuerdo al modelo de Hongzhou-Gouhua y Jiménez-Contreras, i.e., la existencia de un núcleo y sucesivas zonas Bradford, (vea Apéndice B). En combinación con el Principio Espacial establecido anteriormente, se deduce que:

- Con coordenadas geométricas, el espacio es una línea, circulo, esfera o híper esfera (según las dimensiones 1, 2, 3, etc.) o formas de dimensiones fraccionarias de eso, dependiendo de la complejidad de la red de actores.
- Con coordenadas temporales, cuando la dimensión es igual a 1, todos los actores evolucionan a lo largo de un línea temporal única; cuando la dimensión es mayor estricta a 1, los actores evolucionan a la largo de múltiples líneas temporales.

Principio 7. Principio Unificado de Ventaja Acumulativa. La traducción *T*, es proporcional al producto de la ventaja estratégica, *s*, (función de lo espacial, coordenadas geométricas o temporales) por la ventaja intrínseca, *q*, (función de variable cuantitativa de la calidad o atributos del actor o actores). Matemáticamente, lo anterior se define con la expresión denominada Ecuación Fundamental del Modelo Cienciométrico Unificado:

$$T = ksq$$

Donde *k* es la constante de proporcionalidad relaciona a la dimensión del espacio de traducción, *s* es la ventaja estratégica, y *q* es la ventaja intrínseca. En otras palabras, un actor (o grupo de actores) tienen una gran capacidad de traducción dependiendo de capacidades intrínsecas como tan bién a más o a menos posiciones ventajosas dentro de la red.

En lo esencial, los estudiosos de la Cienciometría comparten tres convicciones inamovibles que garantizan la coherencia necesaria a esta disciplina. La primera es que el estudio de las ciencias y de las técnicas pasa necesariamente por el análisis sistemático de las producciones literarias de los investigadores y de los ingenieros: ciertamente, la Cienciometría no se limita exclusivamente a este objeto, pero le concede un lugar esencial. La segunda es que los estudios cuantitativos, siempre que no se constituyan un fin en si, enriquecen la comprensión y la descripción de la dinámica de la tecnociencia. La tercera es la prioridad absoluta y casi obsesiva que conceden a la concepción de instrumentos de análisis sólidos y fiables.

Teoría Actor – Red: La construcción de esta teoría ha sido llevada a cabo por un equipo multidisciplinar y multiinstitucional de investigadores procedentes, básicamente, de varios países europeos y que desde finales de los años setenta hasta la actualidad ha sido adoptada por un buen número de laboratorios y centros de investigación de toda Europa y de Estados Unidos, [Jurado-Alameda et al., 2002].

La Teoría Actor-Red considera la ciencia y la técnica como una red sociocognitiva en la que los aspectos sociales y cognitivos del conocimiento o la técnica se entremezclan

íntimamente. Ciencia y tecnología no se estudian sólo en sí mismos, sino que son consecuencia de la relación interactiva social y cognitiva de un conjunto de actores. Los pilares básicos de la teoría actor – red son: la existencia de un conjunto de actores; la asociación de estos actores en un entramado o red; y la continua transformación de los actores y de la red: proceso denominado *Traducción*.

Los actores se caracterizan porqué pueden ser descritos mediante textos escritos, artículos científicos, libros, patentes etc., y consecuentemente mediante palabras. Estos actores así descritos no poseen una definición constante, sino cambiante en función de las palabras que en cada momento los describan. Un aforismo del tipo "Dime qué escribes y te diré quién eres" podría muy bien representar este modelo. Además, como la ciencia y la técnica producen textos escritos constantemente modificados y evolucionados, los actores que los producen y las relaciones que los unen, deben cambiar de igual forma. Según Callon *un actor se define como alguien o algo creador de asociaciones: un científico, un texto, un aparato, un concepto, etc.* Hay que subrayar que no se hace distinción alguna entre actores humanos y no humanos, ya que son definidos por la red sociocognitiva que crean y por el cambio a lo largo del tiempo de las palabras que los forman. "Un científico que escribe un texto es un actor que cambia constantemente: él nunca permanece lo mismo que anteriormente, ya que investiga continuamente con eficiencia e incrementa el conocimiento". El concepto de actor se puede abordar desde dos puntos de vista complementarios, como las dos caras de la misma moneda, actor-entorno y actor-red. Entendemos por actor-entorno el conjunto de entes que forman ese actor. Todas estas entidades no están jerarquizadas, de tal forma que todas son de igual importancia. Si faltara alguna de ellas, es decir, que se encontrara fuera del mundo o entorno del actor, el actor no podría existir. Supongamos un actor dedicado a la investigación de detergentes formado, básicamente, por un conjunto de investigadores, unos laboratorios, proyectos de investigación, publicaciones y la financiación necesaria, ilustración 6.

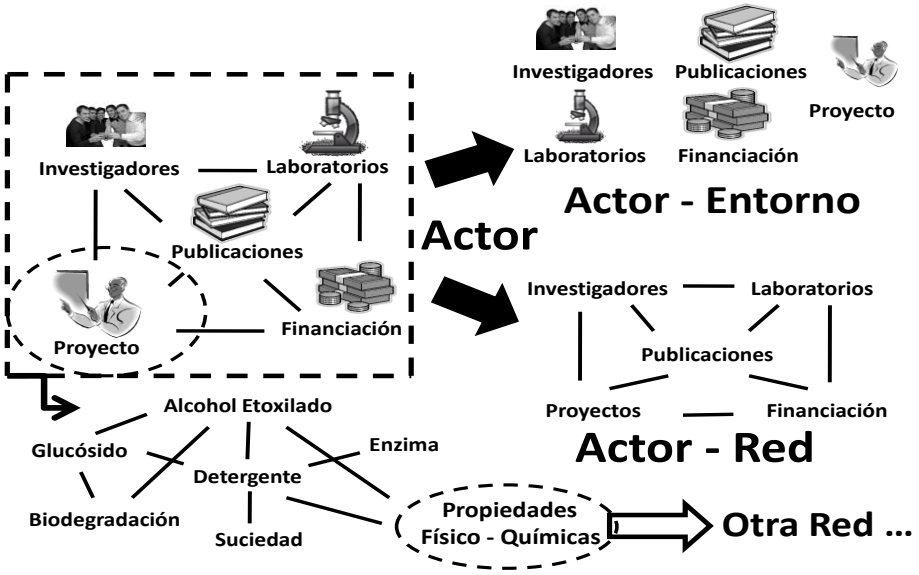


Ilustración 6: Actores y redes.

El actor-entorno estaría formado por cada una de las entidades citadas, la forma en que éstas han sido reclutadas y la distribución de los roles o funciones que deban cumplir cada una de ellas. Indudablemente, si faltara alguna de estas entidades, por ejemplo, los investigadores, no existiría el actor y sería imposible llevar a cabo los proyectos prefijados. Esto es lo que

induce a mantener al mismo nivel cualquier actor, sea humano o no humano. Tan importantes son los investigadores como la financiación, los proyectos o los medios instrumentales: todos son imprescindibles. El actor-red representa, sin embargo, la estructura de relaciones entre los entes que componen el actor. Esta estructura es susceptible de cambiar, es más, la esencia del progreso científico consiste en el cambio continuo y en la redefinición permanente de los actores de la red sociocognitiva. Los entes que forman un actor son a su vez actores, con su entorno y su red. En el ejemplo de la ilustración 6 se observa que el ente "proyectos" es un actor temático cuya red se especifica debajo. Los entes de este actor son las palabras clave glucósido, alcohol etoxilado, suciedad, propiedades físico-químicas, que a su vez son actores temáticos. Cada uno de ellos posee su propia red y su propio entorno, expresable con nuevas palabras.

Teoría de la Traducción: El cambio continuo de la estructura de las redes científicas y técnicas, en definitiva de los actores-red y los actores-entorno, es debido a lo que el antropólogo de la ciencia Bruno Latour llama *traducción*. La traducción representa dos ideas: por un lado, consiste en dar nuevas interpretaciones de los intereses de los actores (significado lingüístico) y, por otro, canalizar a los actores hacia otras direcciones (significado geométrico). Es decir, la traducción de los actores implica un cambio en las palabras que definen el actor y un cambio en la posición estratégica dentro de la red. Supongamos un grupo de investigación dedicado al estudio de las enzimas para su aplicación en la industria alimentaria y cuyo interés es abrir nuevas líneas de investigación en el campo de los productos lácteos. Sea también una empresa que se dedica a comercializar productos lácteos y que necesita desarrollar un nuevo producto que le proporcione ventajas económicas en el mercado. Si ambos actores se acercan y entran a colaborar en un proyecto común, se habrá producido una traducción del tipo "aproximación por convergencia". Se generará un nuevo actor, heredero de los anteriores, cuyas palabras claves serán la conjunción de las palabras clave de los actores precedentes. Este es el significado lingüístico de traducción. A su vez, ambos actores han cambiado su posición dentro de la red acercándose uno al otro. Este es el significado geométrico. En la ilustración 7 se hace un esquema de lo ocurrido. En el transfondo de la traducción ha habido connivencia entre los intereses de ambos actores.

Ilustración 7: Ejemplo de traducción de aproximación por convergencia entre un grupo de investigación y una empresa.

Como se ve, en esta concepción de la dinámica de la ciencia y la técnica, los factores sociales son los que la hacen progresar. Además, se rompe el antiguo mito de que la investigación consiste en "descubrir las leyes ocultas de la naturaleza", como si se tratara de un tesoro oculto. Simplemente consiste en hacer uso de ella para generar nuevos conceptos, nuevos productos, nuevas entidades. En definitiva, los actores formados por redes simplemente extienden sus redes cada vez un poco más allá con el objetivo de crear nuevos actores.

Según el concepto de la traducción, Newton "no descubrió" la Ley de la Gravitación Universal, sino que "creo" una teoría que intentaba explicar el porqué caen los objetos. El modelo creado fue simplemente una construcción matemática de carácter puramente humano, una abstracción semántica que puede ofrecer una explicación plausible de lo observado. Como los términos y relaciones creados por Newton fueron convincentes, quedó como "un hecho

incontrovertible" que la Ley de la Gravitación Universal era una ley natural, oculta y preexistente, dada a la luz por Newton. No obstante, ya en los primeros tiempos el mundo científico planteó preguntas que Newton jamás pudo contestar de forma contundente y que ponían seriamente en entredicho su teoría:

- -¿Cuál era la naturaleza íntima de la interacción entre las masas? ¿Partículas, ondas,…?
- -¿A qué velocidad se transmite la interacción entre las masas? ¿Es finita o infinita?, ¿Cuánto tiempo tarda en llegar la atracción de la tierra a la luna o de la tierra a la manzana que cae de un árbol?
- -La atracción debería "viajar" a través del universo. ¿Para ello era necesario apoyarse en un soporte material, como una especie de éter de propiedades extraordinarias, o bien viajaba sin soporte a través del vacío?
- -¿Esa interacción duraba eternamente o se gastaba? Si no se gastaba nunca, ¿qué la compensaba eternamente? ¿Cuándo se le acabará la gravedad al universo?

A pesar de estas objeciones, la capacidad explicativa de la Teoría de la Gravedad Universal era, y es hoy en día tan grande, que podían pasar inadvertidas y lo más conveniente consiste en aceptarla plenamente como un "hecho". Para la teoría de la traducción un "hecho" es aquello que tras un periodo de controversias es aceptado mayoritariamente por la comunidad científica y deja de discutirse. No obstante, todos los hechos, aun los más asentados como el de la teoría de la Gravitación, pueden en cualquier momento ser revisados, entrar en un nuevo periodo de controversias y ser modificados. Es decir, pueden traducirse de nuevo.

En nuestro ejemplo, el hecho permaneció incontrovertible durante varios siglos hasta que una nueva traducción se cernió sobre el edificio construido por Newton: la Teoría de la Relatividad de Einstein. Para este científico, la Ley de la Gravitación es una red de conceptos mal construida y propone una nueva red que cambia términos como "fuerza", "acción a distancia" o la ecuación $F = ma$, por "velocidad de la luz", "relatividad del tiempo", "espacio geométrico" y $E = mc^2$. No obstante, permanecen muchos otros términos como "masa" o "campo gravitatorio" que nos indican que la red temática de Einstein es la sucesora de la red temática de Newton. Lo que preexiste en la naturaleza es, por ejemplo, la caída de los objetos, en cambio las leyes que explican el fenómeno son tan sólo redes de conceptos que satisfacen necesidades humanas y sociales. En el caso de la "gravitación / espacio geométrico" (depende de la visión explicativa que se emplee), la investigación responde a la necesidad teleológicas de explicar la naturaleza última del Universo. Los intereses pueden ser muy variados y siempre consisten en satisfacer necesidades sociales. En definitiva, Newton y Einstein no "descubrieron" un conjunto de leyes físicas sino que las "inventaron". Es el mismo fenómeno de Henry Ford que "no descubrió" el automóvil con motor de gasolina, sino que lo "invento" y su objetivo era poder disponer de un medio de transporte más eficaz que los anteriores. La ciencia y la tecnología, tal como se pone de manifiesto en el ejemplo de la ilustración 7 de la empresa láctea y el grupo de investigación, no se dirigen a "descubrir" sino a "inventar" conceptos, objetos o artilugios. Ciencia y tecnología son una construcción social e indisociables de la propia sociedad.

Las traducciones que subyacen en la actividad científico-técnica son muy diversas y se han descrito multitud de ellas. Ruiz-Baños clasifica las traducciones según sean de evolución, de bifurcación o de aproximación. Incluye en su clasificación la influencia de la fortaleza de los actores, así como la duración de la traducción, ilustración 8. [Jurado-Alameda et al., 2002].

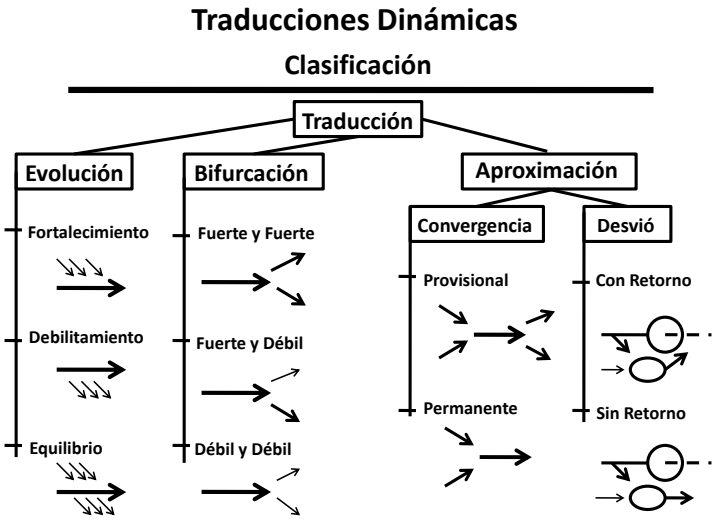


Ilustración 8: Clasificación de las traducciones.

2.3.-Los Indicadores Cienciométricos

Los análisis cuantitativos de la Cienciometría requieren modelos que permitan aplicar técnicas matemáticas al estudio de la ciencia. Los investigadores consideran *al artículo científico como un indicador de la producción de la investigación científica*. Desde esta concepción, la literatura científica se presta para su conteo, su clasificación y su representación bajo la forma de series temporales. 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. Una vez establecido un modelo cienciométrico de la ciencia, se esta en condiciones para desarrollar indicadores que analicen en forma cuantitativa a la ciencia, [Sotolongo, 2002], [Guzmán, 2001]. Para ello considere los postulados 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.

También considere las medidas siguientes: *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; *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; *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; *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 relacionada con la frecuencia de la coocurrencia.

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 aceptados. De los postulados y medidas anteriores se derivan todos los indicadores que se pueden construir en Cienciometría. 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 cienciométricos 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 cienciométricos se clasifican en generaciones según el nivel de complejidad que vayan alcanzando y la evolución (según su surgimiento) en el tiempo en Indicadores de Actividad e Indicadores Relacionales, [Callon et al., 1995].

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 (Datos acerca del volumen y del impacto de las actividades de investigación). Mientras que Indicadores Relacionales están orientados con los aspectos estructurales de la ciencia. Pretenden delimitar las fronteras continuamente cambiantes de las disciplinas, conocer cómo están construidas, su evolución y su solapamiento con otras áreas científicas. 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: *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.*

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. Los Indicadores Relacionales se clasifican en dos clases: Indicadores Relacionales de Primera Generación y de Segunda Generación. Los Indicadores Relacionales de Primera Generación 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*. Los Indicadores Relacionales de Segunda Generación 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. Los indicadores

más importantes son el *Análisis de Palabras Asociadas (Co-Word Analysis)* y el *Análisis de Citas Asociadas (Co-citation Analysis).*

2.4.-El Análisis de Palabras Asociadas

El análisis fue desarrollado por el Centre de Sociologie de l'Innovation, CSI y el Service d'Etude et de Réalisation de Produits d'Information Avancés, SERPIA a mediados de la década de los ochentas para estudiar la evolución de los campos científicos. Este análisis visualiza la estructura de las redes científicas, según la teoría actor - red que concibe la ciencia como una red que entreteje intereses entre actores. Un actor es cualquier ente que participa en esta red y es capaz de generar nuevas redes. De esta forma se revela y se visualiza la evolución de los campos científicos a través de la construcción de conglomerados y diagramas estratégicos.

Las bases teórico–metodológicas han sido establecidas por J. P. Courtial, [Courtial, 1986] que analiza las ventajas del método frente a un análisis escalar multidimensional, MDS, la estabilidad de los esqueletos de las redes en función del umbral y de la indización y otras cuestiones relativas a la pertinencia del análisis de asociaciones. Para la puesta en marcha de este método se ha desarrollado un conjunto de programas informáticos denominados Leximappe.

El Inicio del Análisis de Palabras Asociadas: El análisis se aplica a documentos indizados debidamente indizados a partir del resumen o del contenido textual con palabras claves, en especial, artículos científicos, técnicos y patentes. El Análisis de Palabras Asociadas considera que el contenido de una publicación científica esta descrito parcialmente por las palabras claves o descriptores. Se parte, por tanto, de una matriz de datos *"documentos x descriptor"*, denominada *Matriz de Ocurrencias*, que representa el contenido conceptual del campo científico en estudio, [Courtial, 1990]. Sin embargo, a diferencia de datos habituales, la lista de descriptores puede ser muy extensa por lo que las dimensiones de esta matriz de ocurrencias son extraordinarias. Si partimos de, por ejemplo, una base de datos con 500 documentos y manejamos un vocabulario de 400 descriptores, el número de celdas que contendrá será de 200,000. El programa Leximappe trabaja habitualmente sobre un total de 1500 descriptores. Ya que, según la Ley de Zipf, [Vea Apéndice B] 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. De esta forma, la matriz de ocurrencias se reduce en dimensiones.

	Descriptor 1	Descriptor 2		Descriptor i	Descriptor j		Descriptor k
Doc 1	1	1	.	0	1	.	0
Doc 2	1	0	.	0	0	.	0
.	.	.	.	.	.	.	.
Doc i	0	1	.	1	0	.	0
Doc j	1	0	.	0	1	.	0
.	.	.	.	.	.	.	**.**
Doc N	1	0	.	0	0	.	1
Total	120	98	.	25	20	.	3
	C1	**C2**	**.**	**Ci**	**Cj**	**.**	**Ck**

Tabla 5: Matriz de ocurrencias “documentos x descriptores”.

Las celdas de la matriz de ocurrencia se llena de la siguiente forma: cuando un documento *i* contiene el descriptor *j* en la celda *(i, j)* colocamos 1 y en caso contrario se coloca 0. El número de veces que un descriptor *i* ocurre se denota por C_i.

La Heterogeneidad de los Descriptores. Otra diferencia entre datos estadísticos y descriptores, es la heterogeneidad de estos últimos. Los descriptores regularmente varían desde muy específicos a muy generales. Un documento esta indizado no por más de 15 descriptores, así la matriz de ocurrencia esta especialmente constituida por celdas de ceros.

Habitualmente, el análisis de datos contemplaria calcular, a partir de la matriz de ocurrencia, correlaciones o distancias. Aquí, se calcula la "interrelación" entre descriptores haciendo el producto escalar de los vectores boléanos (i.e., solamente compuestos de ceros y unos), correspondientes a los descriptores de la matriz de ocurrencia. Esto representa la "asociación" o "asociabilidad" de los descriptores en el contexto de documentos que constituyen nuestra base de datos. El número de veces que dos descriptores co-ocurren en los mismos documentos lo denotamos por C_{ij}. En la tabla 6, por ejemplo, se observa que la pareja *(i, j)* co-ocurre 20 veces, lo que significa que los descriptores *i* y *j* aparecen juntos en un total de 20 documentos. Nótese que la matriz de co-ocurrencia es una matriz de adyacencia cuadrada simétrica.

	Descriptor 1	Descriptor 2		Descriptor i	Descriptor j		Descriptor k
Desc 1	-	20	.	20	0	.	2
Desc 2	-	-	.	0	5	.	0
.	-	-	-	-	.	.	.
Desc i	-	-	-	-	20	.	0
Desc j	-	-	-	-	-	.	0
.	-	-	-	-	-	-	.
Desc k	-	-	-	-	-	-	-

Tabla 6: Matriz de co-ocurrencia.

Según, la medida de relación dos descriptores estarán más asociados entre sí cuanto mayor sea la co-ocurrencia entre ellas. Por tanto, la medida del enlace entre dos descriptores de una red será proporcional a la co-ocurrencia de esos dos descriptores en el conjunto de documentos que se tome como muestra. En teoría, a partir de esta matriz de adyacencia podríamos reconstruir completamente la red cienciométrica que genera el campo científico en estudio; pero en la práctica no es conveniente, ya que los valores de las co-ocurrencias, tal cual, dependen del tamaño de la muestra. Bajo estas circunstancias, los estudios de comparación de redes descritas por diferente número de documentos, serían incorrectos. Es por tanto conveniente recurrir a la normalización de los valores de las co-ocurrencias.

Medidas de Normalización. Un análisis de datos, en general, hará intervernir la medida de todas las asociaciones posibles entre los descriptores. Trabajará en un espacio: el espacio de todos los descriptores como si estos constituyeran un espacio homogéneo. A diferencia de esto, el análisis de palabras no pretende contar las asociaciones existentes entre cada pareja de descriptores tampoco pretende razonar en términos de un espacio más bien pretende razonar en términos de cadenas o encadenamientos entre descriptores.

Las medidas de normalización miden la fuerza del encadenamiento entre el descriptor *i* y el descriptor *j* según una ecuación, [Courtial et al., 1984]. Entre las medidas más populares se encuentran: el índice de Jaccard, índice Estadístico, índice de Inclusión, índice de Proximidad, índice de Salton y el índice de Equivalencia. Es costumbre aplicar la medida de normalización a las co-ocurrencias que tengan un valor por arriba de algun umbral establecido, así se consigue resaltar aun más los encadenamientos interesantes de la literatura analizada.

El índice de Jaccard J_{ij} se define como:

$$J_{ij} = \frac{C_{ij}}{C_i + C_j - C_{ij}}$$

Donde C_{ij} es el número de documentos en que aparecen conjuntamente los descriptores *i* y *j*. Mientras que C_i y C_j son el número de documentos en que aparece el descriptor *i* y *j* respectivamente. Este índice es muy utilizado en Biología para medir la semejanza entre dos especies. Considere dos conjuntos, un conjunto representa la especie *i* y el otro representa la especie *j*. El índice de Jaccard se basa en la relación de presencia-ausencia entre el número de especies comunes C_{ij} -de dos comunidades- y en el número total de especies. Los valores del índice oscilan entre 0 y 1, cuando la intersección es nula, $J_{ij} = 0$, y cuando los conjuntos son idénticos, $J_{ij} = 1$. Sin embargo, el índice falla cuando mide el grado de semejanza entre descriptores de baja ocurrencia y de alta ocurrencia, porque tendra valores bajos aun cuando descriptores de baja ocurrencia siempre aparezcan juntos con descriptores de alta ocurrencia, por lo tanto, se recomienda usarlo para explorar semejanzas entre descriptores con ocurrencias medias.

Antes de definir el índice estadístico H_{ij}, se discute brevemente la distribución hipergeométrica. La distribución hipergeométrica es una distribución de probabilidad discreta con parámetros *N*, *d*, y *n* cuya función de probabilidad es:

$$P(X = x) = \frac{\binom{d}{x}\binom{N-d}{n-x}}{\binom{N}{n}}$$

Donde (.) es el coeficiente binomial, *N* es el tamaño de la población, de los cuales *d* objetos son de una primera clase y *N* - *d* son de una segunda clase. Supongamos que de esta población tomamos una muestra aleatoria de tamaño *n*, la muestra es entonces sin reemplazo y el orden de los objetos seleccionados no importa. El espacio muestral de este experimento consiste entonces de todas las posibles muestras de tamaño *n* que se pueden obtener del conjunto mayor de tamaño. Si para cada muestra definimos la variable aleatoria *X* como el número de objetos de la primera clase contenidos en la muestra seleccionada, entonces *X* puede tomar los valores *0, 1, 2,..., n*, suponiendo *n* es menor o igual a *d*. La probabilidad de que *X* tome un valor *x* esta dada por la formula anterior. El valor esperado de una variable aleatoria *X* de distribución hipergeométrica es

$$E(X) = n\left(\frac{d}{N}\right)$$

y su varianza es

$$Var(X) = n\left(\frac{N-n}{N-1}\right)\left(\frac{d}{N}\right)\left(1 - \frac{d}{N}\right)$$

En tal caso, el índice de las co-ocurrencias se compara con el valor esperado. Si se asume que un conjunto de descriptores con ocurrencias dadas esta aleatoriamente distribuido en todos los artículos. La probabilidad de encontrar un cierto valor para la co-ocurrencia ente *i* y *j* esta dada por la distribución hipergeométrica con parámetros *N*, C_i y C_j. (*N* número de artículos).

El índice estadístico H_{ij} se define como:

$$H_{ij} = \frac{1}{\sigma}\left(C_{ij} - C_j\frac{C_i}{N}\right)$$

Donde σ es la desviación estándar y C_jC_i/N es el valor esperado de la distribución hypergeométrica. El índice estadístico H_{ij} es simplemente la desviación normalizada del valor esperado de la co-ocurrencia. Se usa para comparar la frecuencia observada C_{ij}/N de una pareja de descriptores con la frecuencia esperada de esa pareja, como si fuesen independientes los descriptores $(C_i/N)(C_j/N)$. El índice es simétrico y normalizado. Este índice no es muy usado porqué la fuerza del encadenamiento no es una variable importante en las gráficas. Además, su

cálculo es tardado, mientras que información extra no es esencial para la interpretación, [He, 1999]. Cálculos sobre el índice estadístico muestran que un umbral $H_{ij} > 2$ produce las mismas cadenas que el índice de Jaccard con un umbral de 0.19. Los mapas de Jaccard pueden retenerse como una imagen conservadora de los encadenamientos entre el conjunto de descriptores.

Cuando documentos y descriptores tienen otro aspecto, por ejemplo, corpus de documentos que son codificados con descriptores provenientes de recombinaciones del DNA, no hay un claro significado adjunto al universo de documentos y descriptores que ocurren en ellos. En tal caso, todas las conexiones son interesantes y es necesario utilizar otro índice.

El índice de inclusión I_{ij} se define como:

$$I_{ij} = \frac{C_{ij}}{C_j} \quad si \quad C_i < C_j$$

Donde C_{ij} es el número de documentos en que aparecen conjuntamente los descriptores *i* y *j*. Mientras que C_i y C_j son el número de documentos en que aparece el descriptor *i* y *j* respectivamente. Se usa expecialmente para resaltar jerarquias cuando se combinan descriptores con baja ocurrencia y descriptores con alta ocurrencia. El índice de inclusión es la probabilidad condicional de encontrar *j* dado *i*. Los valores de I_{ij} oscilan entre 0 y 1. Fijando un umbral, por ejemplo, de 0.5, los mapas presentan un "Descriptor Maestro", del cual se desprenden descriptores formando una especie de árbol. Las ocurrencias de los descriptores van disminuyendo conforme se alejan del descriptor maestro.

Este índice tiene la mala fortuna de producir encadenamientos "falsos" entre descriptores, por ejemplo, si tres descriptores *i*, *j* y *k* (con la ocurrencia de *i* menor que la ocurrencia de *j* y la ocurrencia de *j* menor que la ocurrencia de *k*) son encadenados en un triángulo, la co-ocurrencia entre *i* y *k* puede ser causada en gran parte por la unión de las co-ocurrencias de *i* y *k*, via *j*, así la inclusión entre *i* y *k* se considera "falsa". Por lo tanto, debe ser eliminada. Los programas de Análisis de Palabras Asociadas deben contener un algoritmo para calcular el valor esperado de C_{ik}, basados en las co-ocurrencias de C_{ij} y C_{jk}. Si el valor actual de C_{ik} excede el valor esperado por una cantidad que excede un valor umbral, la cadena entre *i* y *k* es significativa.

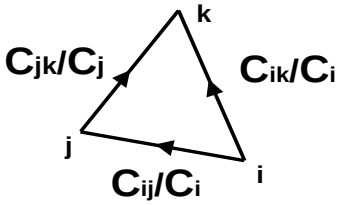


Ilustración 9: Encadenamiento de tres descriptores.

Cuando probabilidades son calculadas sobre la base de las frecuencias actuales de ocurrencia y co-ocurrencia, el argumento es el siguiente. El chance de encontrar *j*, dado *i*, es C_{ij}/C_i, y el chance de encontrar *k*, dado *j*, es C_{jk}/C_j. El producto de esas dos probabilidades condicionales es el chance de encontrar *k*, dado *i* y através de *j*. Si el valor actual C_{ik}/C_i excede el valor esperado, que es el producto de dos probabilidades condicionales, por una cantidad más grande que el umbral elegido, la cadena entre *k* y *i* se considera significativa. Ya que este criterio se aplica para decidir que cadenas bilaterales conservar, el mapa producido es homogéneo.

El índice de proximidad P_{ij} se define como:

$$P_{ij} = N\frac{C_{ij}}{C_i C_j}$$

Donde N es el número de artículos, C_{ij} es el número de documentos en que aparecen conjuntamente los descriptores i y j. Mientras que C_i y C_j son el número de documentos en que aparece el descriptor i y j respectivamente. El índice de proximidad P_{ij} es un índice compuesto, resulta de la división del índice de inclusión I_{ij} por la probabilidad de encontrar el descriptor j en el conjunto de documentos. Algunas veces, I_{ij} da valores bajos, pero que pueden ser significativamente más grandes que la probabilidad de encontrar i en los mismos articulos. Tal situación, implica la existencia de un descriptor mediador, que tiene una frecuencia baja de ocurrencia pero mantiene relaciones significativas con algunos descriptores perifericos. Los descriptores mediadores y perifericos resaltados por P_{ij} pueden representar nuevos desarrollos y/o áreas menores de investigación. Se usa porqué es fácil calcularlo; su interpretación estadística no es sencilla.

Para definir el índice de Salton S_{ij}, revisemos brevemente en que consiste el Modelo Básico del Espacio Vectorial y en que consiste un producto escalar o producto interno. Sea D el conjunto de documentos con elementos $D_1, D_2, \ldots, D_N$, y W el conjunto de descriptores con elementos $W_1, W_2, \ldots, W_k$. Las celdas de la matriz de ocurrencia se llena de la siguiente forma: cuando un documento i contiene el descriptor j en la celda (i, j) colocamos 1 y en caso contrario se coloca 0. Lo anterior se puede expresar como:

$$I = \begin{cases} 1 \;\; si \;\; W_j \in D_j \\ 0 \;\; si \;\; W_j \notin D_j \end{cases}$$

Cada documento D_i se representa por un vector de descriptores:

$$D_i = \{I_1, I_2, \ldots., I_N\}$$

A este modelo se le conoce como Modelo Básico del Espacio Vectorial. Cuando se entra al contenido textual de los documentos con un conjunto de palabras claves, el vector D_i ya no está necesariamente constituido por ceros y unos.

Sea V un espacio vectorial sobre un campo $K = \mathbb{R}$ o $\mathbb{C}$. Si $K = \mathbb{R}$, un producto escalar en V sobre $\mathbb{R}$ es una forma bilineal simétrica definida positivamente. Si $K = \mathbb{C}$, un producto escalar en V sobre $\mathbb{C}$ es una forma hermitiana definida positivamente. Es decir, Si V es un espacio vectorial real o complejo, la forma bilineal $< *, * >: V \times V \rightarrow K$ tal que:

- $\langle \alpha v_1 + \beta v_2, v_3 \rangle = \alpha \langle v_1, v_3 \rangle + \beta \langle v_2, v_3 \rangle$
- $\langle v_1, v_2 \rangle = \overline{\langle v_2, v_1 \rangle}$
- $\langle v_1, v_2 \rangle > 0 \;, \;\; si \;\; v \neq 0$

se llama *producto escalar* o *producto interno*.

El producto escalar en el caso particular de dos vectores en el plano, o en un espacio euclídeo n-dimensional, se define como el producto de sus módulos multiplicado por el coseno del ángulo θ que forman:

$$x \cdot y = \|x\| \|y\| cos\theta$$

Donde $x, y \in \mathbb{R}^n$, $\|x\| = \sqrt[2]{\langle x, x \rangle}$ es la norma del vector *x*, similarmente $\|y\|$ es la norma del vector *y*. Entre las medidas de normalización más populares se encuentra el índice de Salton, el cual mide el coseno del ángulo entre los documentos D_i y D_j.

El índice de Salton S_{ij} se define como:

$$S_{ij} = \frac{D_i \cdot D_j}{\|D_i\|\|D_j\|}$$

Donde D_i y D_j son los vectores de descriptores de los documentos *i* y *j* respectivamente, y $\|\cdot\|$ es la norma. Sus valores oscilan, como en el de Jaccard entre 0 y 1, aunque da valores de similitud más elevados que el de Jaccard. Si dos documentos presentan componentes muy parecidos, el indice $S_{ij} = 1$, en caso contrario, $S_{ij} = 0$, los documentos son ortogonales.

El índice de Equivalencia o de Asociación e_{ij} se define como:

$$e_{ij} = \frac{C_{ij}^2}{C_i C_j}$$

Donde C_{ij} es el número de documentos en que aparecen conjuntamente los descriptores *i* y *j*. Mientras que C_i y C_j son el número de documentos en que aparece el descriptor *i* y *j* respectivamente. El índice de asociación toma valores entre 0 y 1. Cuando dos descriptores no aparecen conjuntamente en ningun documento, $e_{ij} = 0$. En cambio, cuando dos descriptores siempre que aparecen lo hacen conjuntamente en los mismos documentos $e_{ij} = 1$. El índice se puede reescribir de la siguiente forma:

$$e_{ij} = \frac{C_{ij}}{C_i} \times \frac{C_{ij}}{C_j}$$

Donde el primer factor es la probabilidad condicional de encontrar *j* dado *i*, y el segundo factor es la probabilidad condicional de encontrar *i* dado *j*. El índice es una clase de medida de la relación "Y" entre los descriptores i y j, uno influye en el otro, el índice es débil si una de las probabilidades es débil. Por esta razón, también se le conoce como *Coeficiente de Inclusión Mutua*. Este índice es independiente del tamaño de la muestra, es homogéneo porque permanece constante cuando multiplicamos el conjunto de sus variables por un factor constante y es simétrico al contrario del índice de inclusión que es asimétrico. El índice de equivalencia da valores no decrecientes cuando la co-ocurrencia aumenta. Además, el índice de equivalencia entre dos descriptores no debe aumentar si un documento que contiene solamente uno de los dos descriptores es añadido al corpus de documentos. Es peligroso que tal adición modifique la influencia de un descriptor en el otro de tal manera, [Michelet, 1988].

Matriz de co-ocurrencia normalizada. El índice de equivalencia o de asociación tiene muy buenas propiedades que brinda mejores resultados que otros índices de normalización. En el capítulo 5 se utiliza para extraer conocimiento de un corpus de documentos MedLine indizados con el MeSH major topic "Nonlinear Dynamics".

	Descriptor 1	Descriptor 2		Descriptor i	Descriptor j		Descriptor k
Desc 1	-	0.034	.	0.133	0	.	0.011
Desc 2	-	-	.	0	0.013	.	0
	-	-	-	-	.	.	.
Desc i	-	-	-	-	0.800	.	0
Desc j	-	-	-	-	-	.	0
	-	-	-	-	-	-	.
Desc k	-	-	-	-	-	-	-

Tabla 7: Matriz de co-ocurrencia normalizada.

En la tabla 7, por ejemplo, si comparamos los valores de esta matriz con la de co-ocurrencias se observa que el par *(1, 2)* que poseía una elevada co-ocurrencia tiene sin embargo un índice de asociación menor que el par *(i, j)* que tiene la misma co-ocurrencia. Se comprueba, por tanto, que si dos descriptores aparecen juntos muchas veces pero proporcionalmente son aún mayores sus ocurrencias por separado, el índice de asociación será bajo y el análisis de las palabras asociadas considerará la unión poco fuerte. En cambio, dos descriptores poco frecuentes pero siempre que aparecen lo hacen en los mismos documentos, tendrán un índice de asociación muy elevado y por tanto su asociación será muy fuerte.

Supongamos, también, un descriptor que aparece en muchísimos documentos y que no tiene "predilección" por aparecer conjuntamente con algún otro en particular sino que se reparte homogéneamente con todos; en este caso, nunca llegará a formar asociaciones consistentes y el análisis lo considerará demasiado genérico y poco significativo. En definitiva, mediante el uso del índice de equivalencia, el análisis de palabras asociadas es capaz de discernir qué descriptores y qué asociaciones son realmente relevantes en la construcción de la red cienciométrica y eliminar aquellas que por su baja co-ocurrencia relativa o su elevada generalidad no lo son.

La red cienciométrica esta constituidad de conglomerados o sub-redes. Cada una de estas sub-redes representa un *centro de interés*, es decir, zonas de la red muy enlazadas y consistentes, asimilables a "polos de atracción" de gran intensidad informativa. Representan a los actores temáticos más relevantes, de más significado en el paradigma de la investigación en el período en estudio. Si algo es realmente importante, aparece como centro de interés; si su importancia es pequeña o está difuminada, no se manifiesta. Cada centro de interés viene definido por descriptores, aquellos que nos podrán recuperar de forma más óptima los documentos que se asocian a él. Esto es muy importante, ya que nos evita hacer una interrogación a priori equivocada. En definitiva, cada centro de interés tiene asociado el conjunto de documentos más representativo y puede ser identificado con los descriptores óptimos.

Construcción de Conglomerados. La matriz de asociaciones normalizada es la matriz de adyacencia del grafo que representa la red. Cada vértice de este grafo es un descriptor y cada índice de equivalencia entre cada dos descriptores es la ponderación de los arcos que une estas parejas de vértices. En principio sería reconstruible directamente la red imponiendo un umbral mínimo o bien realizar una representación gráfica en dos o tres dimensiones usando un análisis MDS o bien una estructura jerarquizada del tipo dendrograma. Se ha comprobado que posibilidades como éstas no son óptimas para nuestros propósitos, siendo necesario establecer un algoritmo o algoritmos que sean capaces de: [Ruiz-Baños et al., 1998]

- Extraer de la red cienciométrica (excesivamente extensa por el elevado número de vértices y enlaces) aquellas agrupaciones o subredes significativas. Estas subredes representarían los temas de investigación y definirían los actores que forman la red global.

- Ofrecer una estabilidad suficiente de los esqueletos (actores-red) frente a factores negativos como errores en la indización y tamaño de la muestra.
- Poder controlar perfectamente las dimensiones de las subredes que definen los actores (número de palabras y umbral de enlace de los temas).
- Capacidad de calcular parámetros que cuantifiquen los actores y los definan según suposición estratégica y poder seguir su evolución temporal o dinámica.
- En definitiva, que conceptos de la teoría actor-red como las traducciones puedan ser accesibles a un simple cálculo con un microordenador.

Para ello, Leximappe usa los siguientes algoritmos: El Algoritmo de Agrupación por Enlace Simple y El Algoritmo de Agrupación sobre Centros Simples.

Algoritmo de Agrupación por Enlace Simple. Los elementos de la matriz de co-ocurrencias normalizada son ordenados en una lista decreciente según su índice de asociación. Esta lista está formada tan solo por aquellos descriptores que tengan una ocurrencia mínima y pares de asociaciones también con una co-ocurrencia mínima preestablecidas. El programa recorre la lista desde el principio y va construyendo dobletes, tripletes, etc. de descriptores asociados de forma que suministra un grafo conexo que no exceda de un valor máximo de descriptores preestablecido (por ejemplo 10 ó 15). Cada vez que se obtiene un grafo, elimina los descriptores de éste de la lista y comienza el proceso de construcción de nuevos grafos hasta agotar el total de descriptores disponibles. Por ejemplo, suponga que se tiene los siguientes descriptores A, B, C, D, F, G, H e I. En la tabla 8 se muestran las parejas entre estos descriptores ordenados en una lista decreciente según su índice de asociación.

Parejas	e
AB	0.98
CD	0.96
AC	0.95
AE	0.92
FG	0.89
AF	0.86
FH	0.84
FI	0.82

Tabla 8: Parejas entre descriptores ordenados para usar el Algoritmo de Clasificación por Enlace Simple.

A continuación se muestran los pasos que sigue el algoritmo de clasificación por enlace simple para la construcción de conglomerados con un máximo de 5 descriptores.

- Paso 1: El algoritmo encadena a la pareja AB cuyo índice de asociación es 0.98

A → B

- Paso 2: El algoritmo encadena a la segunda pareja con mejor índice de asociación, en este caso, la pareja CD tiene 0.96

A → B
C → D

- Paso 3: La pareja AC tiene un índice de asociación de 0.95. El algoritmo los encadena.

A → B
↓
C → D

- Paso 4: La pareja AE tiene un índice de asociación de 0.92. El algoritmo los encadena y como ya se llego al umbral tope de 5 descriptores por conglomerado. El algoritmo inicia otro conglomerado.

E → A → B
↓
C → D

- Paso 5: Ahora, el algoritmo encadena a la pareja FG con índice de asociación de 0.89

F → G

- Paso 6: El algoritmo encadena a la pareja AF con índice de asociación de 0.86. Se aprecia la formación de dos conglomerados encadenados.

F → G
↑
E → A → B
↓
C → D

- Paso 7: Encadena a la pareja FH con índice de asociación de 0.84.

H ← F → G
↑
E → A → B
↓
C → D

- Paso 8: Finalmente, encadena a la pareja FI con índice de asociación de 0.82. Esta vez el segundo conglomerado contiene solamente 4 descriptores.

I
↑
H ← F → G
↑
E → A → B
↓
C → D

En la siguiente ilustración se aprecian los encadenamientos internos y externos de los dos conglomerados obtenidos por el algoritmo de clasificación por enlace simple.

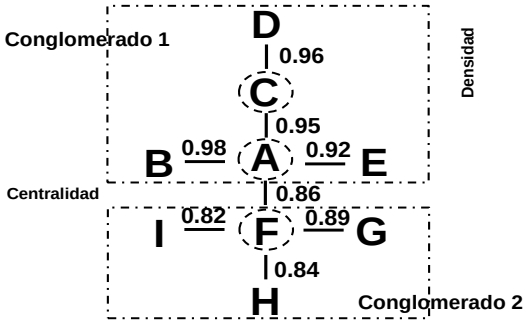


Ilustración 10: Reconstrucción de la red por enlace simple. Se obtuvieron 2 conglomerados de 5 descriptores como máximo.

Algoritmo de Agrupación sobre Centros Simples. Este algoritmo también ordena los pares de asociaciones por orden decreciente de índice de asociación y sólo pueden formar parte de esta lista los descriptores con una ocurrencia mínima y los pares con una co-ocurrencia mínima establecidas previamente. El algoritmo inicia un contador para cada descriptor y comienza a recorrer la lista desde el principio incrementando el contador de los descriptores que van apareciendo. Cuando el contador de un descriptor alcanza un valor igual al número de descriptores máximo estipulado para los temas menos uno, el algoritmo toma este descriptor como centro de un conglomerado. El conjunto resultante estará formado por las uniones de este descriptor central y todos aquellos que se han asociado con el. El resultado es una estructura en forma de estrella. Los descriptores que han aparecido se eliminan de la lista y se comienza de nuevo el proceso para generar más conglomerados. Si después de recorrer toda la lista ningún contador llega al valor máximo preestablecido, éste se disminuye en tantas unidades como sea necesario para formar un nuevo conglomerado. El proceso finaliza cuando el valor máximo del contador disminuya hasta un valor mínimo preestablecido o se terminen todos los descriptores de la lista ordenada de pares. Este algoritmo tiene la ventaja, frente al anterior, de que nos asegura que cualquier sub-red obtenida contiene al menos un descriptor unido a todos los demás. Este descriptor principal nos va a facilitar la identificación del tema de investigación.

Por ejemplo, suponga que se tienen los siguientes descriptores A, B, C, D, E, F, G, H e I. Ordenados en una lista decreciente según su índice de asociación, y además con sus respectivas co-ocurrencias. Vea tabla 9. En este caso el algoritmo considera al descriptor A como palabra principal pues tiene un valor de co-ocurrencia 4. Como se requieren conglomerados de 5 descriptores como máximo el algoritmo considera al descriptor A como centro del conglomerado. Posteriormente, encadena solamente aquellos descriptores que ocurren con A.

Pareja	e	A	B	C	D	E	F	G	H	I
AB	0.98	1	1	0	0	0	0	0	0	0
CD	0.96	1	1	1	1	0	0	0	0	0
AC	0.95	2	1	2	1	0	0	0	0	0
AE	0.92	3	1	2	1	1	0	0	0	0
FG	0.89	3	1	2	1	1	1	1	0	0
AF	0.86	4	1	2	1	1	2	1	0	0

Tabla 9: Parejas de descriptores ordenados para usar el Algoritmo de Agrupación sobre Centros Simples.

En la ilustración 11 se muestran la reconstrucción de la red obtenida con este algoritmo. Se observa que el resultado es diferente ya que resulta un solo conglomerado con una estructura en estrella que será identificado mediante el descriptor A.

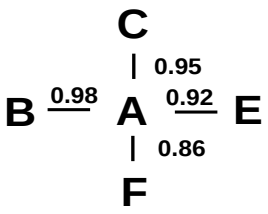


Ilustración 11: Reconstrucción de la red por centro simple.

Ambos métodos pueden dar resultados diferentes a la hora de definir los conglomerados, pero al reconstruir la red global, el resultado es parecido. La propia esencia de las redes cienciométricas es la presencia de fronteras difusas, por lo que no es de extrañar que no sea posible definirlas exactamente. Según el algoritmo utilizado, trazaremos más hacia un lado o hacia otro de la frontera difusa, la línea divisoria que nos servirá de referencia, pero debe entenderse que esta línea es sencillamente una guía para adentrarnos de forma simplificada en el estudio de las redes que de por sí son muy complejas.

La Red y sus Conglomerados. Ambos algoritmos construyen conglomerados o sub-redes a partir de variables umbrales, caracterizados por el valor de la primera cadena rechazada, que se denomina *umbral de saturación* del conglomerado. Estos umbrales son automáticamente determinados de tal manera que ningun conglomerado contiene más de 10, 15 o más descriptores. Puede suceder que el conglomerado tenga menos descriptores de los requeridos entonces se añaden otros descriptores y el conglomerado resultante tendra los descriptores requeridos. Cuando un conglomerado se satura un nuevo conglomerado se inicia, el valor de asociación de la primera cadena se vuelve el nuevo *umbral tope* del conglomerado. Cada conglomerado esta consecuentemente caracterizado por sus umbrales de tope y de saturación, los cuales varian de un conglomerado al siguiente.

La variación del umbral de saturación deja conglomerados arbitrarios. Las cadenas externas de estos conglomerados pueden ser de dos tipos. Cuando el valor de asociación de una cadena inter- conglomerado es más alto que el umbral tope del conglomerado examinado. Esto sugiere que el conglomerado examinado es simplemente la continuación de un conglomerado anterior y que fue ya saturado. Cuando el valor de asociación de la cadena inter-conglomerado es más bajo que el umbral de saturación, esto sugiere que el conglomerado examinado se esparcio en otro conglomerado. En ambos casos la partición de la red en conglomerados de 10, 15 o más descriptores es artificial y generalmente para identificarlos en la red global es necesario clasificarlos en tres tipos:

Conglomerado Aislado: Se caracterizan por la ausencia o baja intensidad de cadenas con otros conglomerados. *Conglomerado Secundario:* Sus cadenas externas con otros conglomerados están por encima del umbral tope son suficientemente fuertes que es legítimo considerar que son extensiones naturales de los siguientes conglomerados. *Conglomerado Principal:* uno o más conglomerados se asocian a él por cadenas cuyo valor es más bajo que el umbral de saturación.

Para dar a estas definiciones un valor operacional se necesitan algunas reglas generales para determinar la asociación de dos conglomerados (uno se vuelve principal y el otro secundario). Se a elegido considerar que dos conglomerados están asociados si están ligados por al menos tres cadenas. Asi, tenemos dos listas para el corpus. En una lista están todos los conglomerados aislados y en la otra lista hay grupos cada uno contiene conglomerados principales con sus respectivos conglomerados secundarios. Se puede dar el caso que el conglomerado aislado contenga conglomerados asociados a él, es decir, forma una grupo con su propia autonomia y coherencia.

Describir la dinámica de un sector consiste en identificar los conglomerados principales, conglomerados secundarios y conglomerados aislados para caracterizar su contenido y seguir su evolución. Para simplificar el analisis, una distinción adicional que ayuda a seleccionar conglomerados con una fuerte destreza para estructurar la red global son los denominados *"conglomerados crossroads".* Son conglomerados principales que tienen por lo menos dos conglomerados secundarios y juegan un papel esencial en la transformación de la red. Una última observación sobre esta clasificación. Se puede considerar que la distinción entre conglomerado principal y conglomerado secundario es artificial, ya que hay una unidad base entre grupos que ha sido arbitrariamente separada. De hecho, desde el punto de vista del análisis e interpretación, esta distinción es muy útil y debe retenerse. El conglomerado principal designa el corazón de una sub-red (en algún sentido el núcleo). Como resultado, conglomerados principales (y en particular conglomerados crossroads) identifican los problemas focales y estructurales del campo estudiado. Algún análisis deberá por lo tanto empezar con ellos.

Caracterización de los Conglomerados como una Función de su Participación en la Organización de la Red: Centralidad y Densidad. Un conglomerado puede definirse de dos formas. Primeramente, puede verse como un punto en la red general, el cual esta caracterizado por su posición, es decir, por el bulto de cadenas que lo unen a otros conglomerados/puntos en la red general. Secundariamente, puede verse como un conglomerado hecho de descriptores encadenados con otros (el mismo define mas o menos una red densa, la cual es mas o menos coherente y robusta). Es necesaria esta doble perspectiva analítica para apreciar la dinámica del total. En efecto, la red general se desarrolla de dos formas, por la reorganización de las relaciones entre conglomerados con una composición interna estable y por la reconstrucción, redefinición de los conglomerados hechos o por la apariencia de nuevos conglomerados (si esos emergen progresivamente o resultan de la fusión de conglomerados existentes) o por la desaparición de conglomerados (los cuales son progresivamente destruidos o fragmentados). Estos dos mecanismos raramente son mutuamente independientes. En general, uno observa una modificación del contenido de los conglomerados y de sus listas al mismo tiempo una redefinición de las cadenas que los unen. Esta complejidad es la escencia de investigación: hacer simplificaciones y buscar conglomerados en el que alguien pueda hablar sobre especialidades, campos o temas de investigación que son estables en el tiempo. El problema en el cual nos estamos posicionando es la identificación de que cambios y que transformaciones. Asi, evitamos respuestas a priori y nos proveemos con los medios y herramientas que permitan respuestas empiricas a la pregunta. Este es el fin de las siguientes dos nociones: *Centralidad* y *Densidad*.

Una vez identificados los conglomerados, definidos por sus descriptores y por los enlaces que los unen, es conveniente poder establecer parámetros numéricos que de alguna forma nos hagan referencia a sus estructuras internas y a su relación con la globalidad de la red. Se definen los índices siguientes:

a) Densidad. La densidad o índice de cohesión interna es la intensidad de las asociaciones internas de un conglomerado y representa el grado de desarrollo que posee. Se calcula como el cociente entre la suma de los índices de equivalencia internos y el número de descriptores que definen el conglomerado multiplicado por 100 para evitar números decimales.

$$d = 100\frac{1}{p}\sum_{i=1}^{L} e_i$$

Donde e_i es el índice de equivalencia del enlace interno *i*. *L* es el número de enlaces internos del conglomerado. *p* es el número de descriptores del conglomerado. Densidades elevadas corresponden a conglomerados altamente desarrollados, muy especializados y repetitivos en sus conceptos. Si ordenamos un conjunto de conglomerados por orden creciente de densidad, el rango de cada conglomerado es lo que se denomina rango densidad. Cuando se

normaliza, dividiendo entre el número total de conglomerados de la red, presenta valores entre 0 y 1. Se utiliza en la construcción del diagrama estratégico como sinónimo de densidad y es indispensable para hacer estudios comparativos con otras redes y en estudios dinámicos.

$$r_d = \frac{1}{N} rango$$

Donde *rango* es el rango del conglomerado según su densidad. *N* es el número de conglomerados de la red.

b) Centralidad. La centralidad o índice de cohesión externa es la suma de los índices de equivalencia de todos los enlaces externos que posee un conglomerado. Usualmente el valor de la centralidad se multiplica por 10.

$$c = 10 \sum_{j=1}^{T} e_j$$

Donde e_j es el índice de equivalencia del enlace externo *j*. *T* es el número total de enlaces externos. Un conglomerado con elevada centralidad está situado en el centro de la red y se relaciona muy bien con los demás conglomerados. Si de forma análoga a la densidad ordenamos un conjunto de conglomerados por orden creciente de centralidad, el rango de cada conglomerado (que puede ser también normalizado dividiendo entre el número total de conglomerados) es lo que se denomina rango centralidad. Se utiliza, junto con el rango densidad en la construcción del diagrama estratégico como sinónimo de centralidad y es imprescindible para hacer estudios comparativos entre redes y en estudios dinámicos.

$$r_c = \frac{1}{N} rango^t$$

Donde $rango^t$ es el rango del conglomerado según su centralidad. Si representamos en un diagrama cartesiano en el eje de abscisas la centralidad y en el eje de ordenadas la densidad, obtenemos lo que se denomina diagrama estratégico. Los cuatro cuadrantes de que consta nos definen las cualidades de los centros de interés (conglomerados) contenidos en ellos. La centralidad en términos de la sociología de traducción significa que el conglomerado en cuestión es un punto obligado de paso. Es esencial para cualquiera interesado en los conglomerados asociados a él, para investigar directamente o indirectamente, el sector.

Diagrama Estratégico: Identificar los conglomerados y describir las relaciones que los unen constituye una primera etapa de la descripción de la red. Queda después caracterizar la morfología de conjunto de esta red y la contribución de cada conglomerado a su estructuración. Con esta finalidad se introdujeron dos nociones, la de centralidad y la de densidad, que están destinadas a resaltar la contribución de los diferentes conglomerados a la estructuración de la red general. Las nociones de centralidad y de densidad permiten presentar de forma sintética y simplificada la morfología de la red y preparar un estudio dinámico. Como cada conglomerado se define por su centralidad y por su densidad es posible trazar un diagrama estratégico. Este diagrama se obtiene colocando los conglomerados horizontalmente (siguiendo el eje de las *X*) por orden de centralidad creciente y verticalmente (siguiendo el eje de las *Y*) por orden de densidad creciente. Esta operación permite clasificar, con la ayuda de valores medios, todos los conglomerados en cuatro categorías que corresponden a los cuatro cuadrantes del diagrama estratégico, ilustración 12.

Ilustración 12: Diagrama estratégico.

Los conglomerados del tipo 1 son a la vez los centrales en la red general (están solidamente conectados a otros conglomerados) y están recorridos por intensas relaciones que manifiestan su alto grado de desarrollo y de integración. Estos conglomerados constituyen de alguna forma el centro del campo. Su posición es estratégica y probablemente estén bajo tutela de un grupo de investigadores bien estructurado que se hace cargo de ellos de forma sistemática y durable. Los conglomerados del tipo 2 son centrales, es decir, están ampliamente conectados a otros conglomerados pero la densidad de sus relaciones internas es relativamente débil. Merecen que se les preste atención pues son susceptibles de convertirse en centrales y desarrollados (y consiguientemente desplazarse hacia el cuadrante 1): representan frecuentemente temas importantes para el desarrollo del campo. Los conglomerados del tipo 3 son poco centrales (podemos clasificarlos como periféricos) y la intensidad de sus relaciones (gran densidad) hacen pensar que se corresponden a problemáticas de investigación cuyo estudio esta bien desarrollado. Puede tratarse de conglomerados que en una fase anterior eran centrales pero que aun siguiendo siendo objeto de inversiones importantes que se han visto marginados progresivamente atrayendo cada vez menos interés a su alrededor: aparecen como especializaciones que interactúan débilmente con respecto a las otras subredes. Los conglomerados del tipo 4 son a la vez periféricos y poco desarrollados. Representan los márgenes de la red. Solo un análisis dinámico (la evolución de la red a lo largo de varios periodos) o comparativo (las relaciones de la red con otras redes) permitiría precisar su contribución al desarrollo del campo.

Esta clasificación de los conglomerados visualizada bajo la forma de un diagrama estratégico cuya lectura resulta fácil, proporciona una descripción más detallada del estado de una red determinada, de la posición y del grado de desarrollo de los temas que la constituyen. El hecho de que dos conglomerados estén cercanos uno del otro en el diagrama estratégico no significa que estén relacionados entre ellos: la única conclusión que puede deducirse de esta observación es que sus índices de centralidad y de densidad tienen valores próximos.

La Estructuración de un Campo: La red cienciométrica permite calificar la morfología de un campo de investigación o de un sector técnico. Es fácil imaginar que en algunos casos las temáticas de investigación sean poco numerosas y muy coherentes mientras que otras configuraciones se caracterizan por una gran diversidad de temas débilmente unidos unos con otros. El diagrama estratégico es un instrumento muy eficaz para proporcionar una representación de la estructura de un campo de investigación. Pueden considerarse tres tipos de organización, ilustración 13.

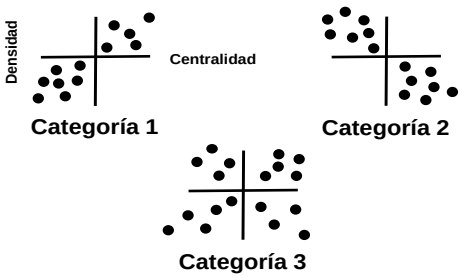


Ilustración 13: Categorías en que se estructura una red.

Categoría 1: Una distribución de los temas alrededor de la primera bisectriz (cuadrante 1 – cuadrante 4) indica que el campo se organiza en torno a un núcleo de temas bien estructurados y bien desarrollados con los cuales se relaciona una serie de temas periféricos y poco desarrollados. Categoría 2: Una distribución de los temas alrededor de la segunda bisectriz (cuadrante 2 – cuadrante 3) indica un campo en vías de estructuración o bien en vías de desintegración: pocos temas están en posición central y la red es muy policéntrica. Categoría 3: Una distribución igual de los temas por los diferentes cuadrantes caracteriza un campo cuya estructura es muy compleja y rica ya que encontramos todas las familias de temas: unos son centrales, otros periféricos y los diversos grados de desarrollo están presentes. Una configuración de este tipo que reúne todas las familias posibles de conglomerados (estabilizados, emergentes o en vías de desarrollo) sugiere una dinámica importante del campo.

Si las impresiones visuales son consideradas demasiado subjetivas siempre se pueden crear índices cuantificados que permitan representar la mayor o menor concentración de los conglomerados en ciertos cuadrantes del diagrama estratégico. En resumen, la red cienciométrica (estática) se describe como: (a) Una lista de conglomerados (principales, crossroads, secundarios y aislados). (b) Cada conglomerados es caracterizado por un índice de centralidad y un índice de densidad. (c) Se dibuja un diagrama estratégico, en el cual se hace una clasificacion de los conglomerados en cuatro familias, correspondientes a los valores relativos de sus dos indices –centralidad y densidad. Esta descripción proporciona la base para los análisis comparativos de las diferentes redes en estudio, [Courtial et al., 1991], [Callon et al., 1995].

El Análisis Dinámico: Queda ahora proponer los instrumentos para estudiar una red en el discurrir del tiempo. El análisis es capaz de poner de manifiesto la dinámica evolutiva de la estructura interna de la red. Esta evolución se debe al cambio que sufren en sus definiciones los actores, causada por el juego de intereses existente entre ellos. En definitiva, un estudio cienciometrico, a la luz de la teoría actor-red, debe contemplar este hecho y debe ser la guía y norte de la explicación del devenir científico y técnico. En este caso se construye una serie de corpus distintos. Cada corpus reúne documentos publicados en un intervalo de tiempo determinado y da lugar a un análisis independiente por el análisis de las palabras asociadas. Para trazar la dinámica del campo estudiado el análisis pasa por tres etapas.

Paso 1.- La comparación de los conglomerados. Se establece la lista de los conglomerados de las redes que se van a comparar. Se define un conjunto de indices para medir la similitud entre dos conglomerados cualesquiera que pertenezcan a diferentes redes. Hay que advertir que los índices que se van a definir a continuación se utilizan tanto para comparar temas en diferentes tiempos, análisis dinámicos, como temas en diferentes redes (por ejemplo, redes de investigación académica e investigación técnica) en el mismo periodo de tiempo. Sea cual fuere la intencionalidad, las definiciones son idénticas, aunque el alcance de las conclusiones que se extraigan serán diferentes.

El Índice de Intersección. Suponga dos conglomerados T_1 y T_2, y se quiere determinar su similitud. Se define índice de intersección como el número de descriptores comunes, W_{12}, que hay entre ambos conglomerados. Normalmente, se dira que dos conglomerados están relacionados por su similitud temática si su índice de intersección supera un umbral mínimo de, por ejemplo, 3. Este índice no es suficientemente ecuánime, ya que dependiendo del tamaño de los conglomerados que se comparan, el número de descriptores comunes puede representar fracciones de tema muy distintas y por tanto similitudes relativas variables: dos conglomerados de 4 descriptores en total con 3 comunes son, por supuesto, más similares que dos conglomerados de 15 descriptores en total y también con 3 comunes.

El Índice de Transformación. Sean de nuevo los conglomerados T_1 y T_2 y se quiere determinar cuanto se diferencian entre ellos, para ello se define el índice de transformación, *t*, como el cociente entre la suma de descriptores existentes en ambos conglomerados y el número de descriptores comunes:

$$t = \frac{W_1 + W_2}{W_{12}}$$

Donde, W_1 es el número de descriptores del conglomerado 1. W_2 es el número de descriptores del conglomerado 2. W_{12} es el número de descriptores comunes entre los conglomerados 1 y 2. Hay que hacer notar que si dos descriptores aparecen en los dos conglomerados a la vez, deben contarse dos veces. Cuando mas elevado es, menos se parecen los conglomerados (tienen relativamente pocos descriptores en común). Cuando mas bajo resulta mas comparables son los conglomerados (los descriptores que los definen son idénticos).

El Índice de Influencia e Índice de Procedencia. Miden el grado de continuidad entre dos generaciones de conglomerados. El índice de influencia es la proporción de descriptores de un conglomerado que reaparecen en otro conglomerado de la siguiente generación. Cuando la proporción citada se acerca a la unidad diremos que la influencia de un conglomerado de la primera generación sobre otro de la segunda generación es elevada. El índice de procedencia muestra la proporción de descriptores de un conglomerado de segunda generación que provienen de un conglomerado de primera generación. Ambos índices, el de influencia y el de procedencia presentan valores entre 0 y 1.

Una vez que todos los índices posibles han sido establecidos se dispone de un cuadro de correspondencias que proporciona adscripciones o proximidades entre los conglomerados de las diferentes redes estudiadas. Conjuntamente al uso de índices como los anteriores, es conveniente analizar los descriptores, una a una ya que van cambiando, para así ir valorando cuál es la carga conceptual que presenta el conglomerado cada año.

Series temáticas. Una serie temática es un conjunto de temas de generaciones encadenados por un valor de similitud umbral. Esta relación viene determinada por el Índice de Similitud Dinámica, ISD. Es una medida de la cantidad de significación que un tema conserva a lo largo de las traducciones sucesivas que sufren a lo largo del tiempo. Se calcula como el cociente entre el cuadrado del número de descriptores que se conservan y el producto del número de descriptores que contiene el tema antes y después.

$$t = \frac{W_{12}^2}{W_1 + W_2}$$

Donde, W_{12} son los descriptores comunes del tema en la generación 1 con el tema en la generación 2. W_1 son los descriptores de la generación 1. W_2 son los descriptores de la generación 2. El valor del ISD oscila entre 0 (no se parece en nada un tema de una generación con otro tema de la segunda siguiente) y 1 (cuando se mantiene idéntico, sin cambio alguno). Se sabe que si se toma como umbral un valor del ISD igual a 0.09, es posible construir cadenas de

temas similares o series temáticas de manera optima. Si estas cadenas las representamos gráficamente respecto del tiempo obtenemos lo que se denomina Diagrama Cronológico. Es muy útil para visualizar con comodidad el camino seguido por los temas a lo largo del tiempo, sus apariciones, desapariciones, fusiones y disgregaciones, [Ruiz-Baños et al., 1999].

Evolución de los descriptores de los conglomerados. Se parte de la serie temática en estudio y se construye una tabla que muestra la evolución conceptual de los temas (desaparecen unos para dar paso a otros). Esto es lo que se denomina sentido lingüístico de la traducción.

Paso 2.- Comparación de las posiciones de los conglomerados sobre los diagramas estratégicos. Una vez establecidas las (eventuales) similitudes entre los diferentes conglomerados de las redes consideradas se examinan sus posiciones sobre los diagramas estratégicos correspondientes, es decir, establecer pautas de comportamiento evolutivo.

Generalización: los temas nacen en el cuadrante 4 (poco desarrollados y periféricos), pasan al cuadrante 2 (centrales y de elevado interés pero todavía poco desarrollados), pasan al cuadrante 2 (centrales y de elevado interés pero todavía poco desarrollados), para luego continuar en el cuadrante 1 (alta centralidad y desarrollo) y seguir en el cuadrante 3 (desarrollados pero cada vez menos interesantes y mas alejados del centro de la red). Finalmente, vuelven al cuadrante 4 en que mueren por indefinición interna y alejamiento de la red. Este movimiento circular en el sentido contrario de las agujas del reloj no suele observarse completo debido a las continuas traducciones y redefiniciones de los actores temáticos. En otros casos, el movimiento es a la inversa, cuando temas objeto procedentes del exterior de la red aparecen en el cuadrante 3 (arriba a la izquierda), pasan al cuadrante 1 produciéndose una "proliferación" y convergencia de líneas correspondientes a la construcción de nuevos conocimientos.

Paso 3.- El ciclo de vida de los conglomerados. Cada periodo estudiado se caracteriza por una lista de conglomerados. La evolución de cada conglomerado en el curso del tiempo es analizada mediante tres índices: el índice de densidad, el índice de centralidad y el índice de transformación. Supongamos que seguimos la evolución de un conglomerado durante un periodo largo de años a través del estudio de sus propiedades. Tendremos sobre la serie temática informaciones abundantes y muy significativas. Esta evolución dinámica de las propiedades de un conglomerado es lo que denominamos *Ciclo de Vida*. En una representación gráfica se pueden incluir variables como las siguientes:

- Índice de transformación que nos expresará los cambios conceptuales del conglomerado.
- Centralidad y densidad. Nos ofrecerá una visión cuantificada y progresiva de la cercanía o alejamiento al centro de la red y del desarrollo interno.
- Número de artículos: nos proporcionara información sobre el tamaño que adquiere en cada momento el conglomerado.

Se han descripto tres patrones de ciclo de vida. El primero presenta un máximo de centralidad y densidad. Normalmente hay un desfase entre centralidad y densidad, sobre todo en la ciencia académica, precediendo la primera a la segunda. El porcentaje de artículos suele atribuir este valor máximo. El segundo tipo es aquel en que la centralidad y/o la densidad presentan dos máximos o picos. En este caso suele haber una profunda transformación en el contenido del conglomerado, a veces tan grande que en el periodo de tiempo entre picos puede incluso desaparecer momentáneamente. Por ultimo, el tercer tipo es aquel en que todos los parámetros incrementan o decrementan constantemente.

Análisis Predictivo. El conocimiento de los patrones de comportamiento de los conglomerados con el tiempo o ciclos de vida nos puede permitir realizar predicciones a corto y medio plazo. Usualmente, en los ciclos de vida se han seguido las variables centralidad,

densidad, índice de transformación y número de artículos referentes al tema. Para llegar a profundizar aun más en estos patrones y sobre todo para poder encontrar causas que induzcan a comportamientos concretos se hace necesario incluir, más variables en los ciclos de vida. Aparte de los patrones antedichos se esta intentando encontrar el origen causal en el concepto de *"neguentropía"* y en la interpretación y modernización de las traducciones que pueden regir estas evoluciones. Esta cuestión del análisis predictivo es el reto investigador, difícil y complejo por su naturaleza, que se vislumbra para los próximos años.

En el capitulo 5 se utiliza la medida de asociación o de equivalencia e_{ij} para resaltar los encadenamientos internos y externos de las redes red biomédicas dadas por corpus de documentos MedLine.

2.5.-La Actividad Científica

La ciencia y la tecnología adquirieron 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, [Sancho, 1990]. 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. Cabe destacarse que a nivel internacional, la *Investigación Científica* es un subsistema de Actividad Científica. La Actividad Científica puede ser considerada como un análogo a los modelos teóricos económicos de Entradas-Salidas (Input–Output). Esta forma de ver a la Actividad Científica puede resultar un poco simplista para algunos científicos pero tiene validez desde el punto de vista económico, [Venegas, 2003]. 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.

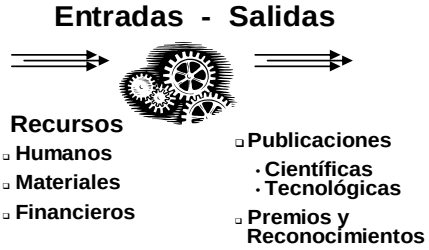


Ilustración 14: Entradas y salidas de la Actividad Científica.

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.

Algunos resultados de la Actividad Científica son: las 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 [Spinak, 2001] 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 como Rosa Sancho 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.

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. En México, El Consejo Nacional para la Ciencia y la Tecnología, CONACYT, publica anualmente los indicadores de actividades científicas y tecnológicas referentes a las acciones que se desarrollan en el país de manera sistemática para la producción, diseminación y aplicación de los conocimientos en estas áreas. Es importante mencionar que los indicadores cienciométricos tienen cabida en la medición del Output, 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.

3 PubMed

En la actualidad existen bases de datos especializadas en todas las áreas científicas y técnicas, lo que permite analizar cualquier de ellas a través de estas fuentes. Sin embargo, la validez del análisis cienciométrico 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 del análisis.

3.1.-El Artículo Científico

De todas las publicaciones científico-tecnológicas, solamente el *artículo científico* es considerado pieza clave para estudiar a la ciencia por medio de análisis cuantitativos, es decir, a través de los análisis cienciométricos. El artículo científico tiene la característica de ser un texto estructurado 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 partes y cada una de ellas tiene una misión informativa diferente. La estructura más general es la siguiente: [Oscar, 2005]

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

Según los expertos un artículo científico debe 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. 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 último *la actualidad/accesibilidad*, [Sánchez-Paus, 2002].

3.2.-Las Bases de Datos Científicas - Tecnológicas

Estas bases de datos son la principal fuente de información que se utiliza en los análisis cienciométricos, [Guzman, 2001]. 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 implementar programas de gestión documental. Estos programas se encargan de estructurar y controlar la información,

para facilitar, en cualquier momento, su rápida y precisa localización y recuperación. En la tabla 10 se muestran algunas bases de datos cientificas-tecnologicas reconocidas 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 Estados Unidos	PubMed	Biomédicina	17 millones de citas	1966
The Thomson Corporation	Web of Science	Varias disciplinas cientificas	Más de 5.4 millones de enlaces a documentos en texto completo	1960

Tabla 10: Algunas bases de datos científicas-tecnológicas.

PASCAL proporciona información sobre más de 6000 titulos de revistas, articulos de revista, procedimientos, disertaciones, libros, patentes, reportes. *Latindex* brinda información básica sobre más de 12,000 títulos a través del directorio. Ofrece información adicional sobre un conjunto seleccionado de revistas. Brinda también acceso a recursos electrónicos a través del índice. *ICYT* proporciona información de 747 publicaciones periódicas editadas en España, fundamentalmente revistas además de monografías, actas de congresos, informes y tesinas. *PubMed* entre sus ventajas como fuente de información está su amplia cobertura de revistas, pues contiene aproximadamente 17 millones de citas bibliográficas que provienen de MedLine y de otras áreas de la Biomedicina. *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).

3.3.-PubMed

En Estados Unidos de America, el Department of Health and Human Services, tiene el objetivo cuidar la salud del pueblo norteamericano, para lo cual, ha desarrollado una infraestructura tanto material como intelectual. Los National Institutes of Health, NIH, son el principal motor para cumplir cabalmente dicho objetivo, pues a través de estos institutos, se desarrolla toda la investigación médica que se realiza en ese país. Los Institutos se localizan en Bethesda, Maryland, Estados Unidos de America. A cada instituto se le asignan fondos y desarrollan sus propias investigaciones, [Pubmed, 2008]. Algunos de estos institutos son: National Cancer Institute (NCI), National Eye Institute (NEI), National Heart, Lung, and Blood Institute (NHLBI), National Institute of Child Health and Human Development (NICHD), National Institute of Dental and Craniofacial Research (NIDCR), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute on Drug Abuse (NIDA), National Library of Medicine (NLM), Center for Information Technology (CIT formerly DCRT, OIRM, TCB), Center for Scientific Review (CSR), John E. Fogarty International Center (FIC), etc.

La National Library of Medicine, NLM, tiene un lugar privilegiado entre dicho conjunto, pues desarrolla el papel de administrador de la información relacionada con las ciencias biomédicas, haciendo uso de la mas moderna tecnología de bases de datos. Toda esta información esta disponible para toda persona interesada a través de una gran variedad de medios. Las bases de datos pertenecientes al National Library of Medicine se dividen en dos grupos: *ELHILL* y *TOXNET*. Las bases que se encuentran en *ELHILL* tratan sobre temas biomedicos y sus áreas relacionadas. Algunas bases pertenecientes a este grupo son: *AIDSDRUGS; CATLINE; HISTLINE; POPLINE; TOXLINE; AIDSLINE; ChemID; HSRPROJ; PREMEDLINE; AIDSTRIALS; DIRLINE; MEDLINE; SDILINE; AVLINE; DOCUSER; MeSH Vocabulary; SERLINE; BIOETHICSLINE; HealthSTAR; OLDMEDLINE; SPACELINE.* Por ejemplo, *AIDSLINE* (AIDS Information onLine): contiene citas bibliográficas sobre sida y temas relacionados. *HealthSTAR*: recoge los aspectos clínicos y no clínicos de la gestión sanitaria. *HISTLINE* (HISTory of medicine onLINE): proporciona referencias bibliográficas sobre historia de la medicina. *HSPROJ* (Health Services Research Projects in Progress): trata sobre los proyectos de investigación en curso que están financiados por instituciones privadas o públicas de Estados Unidos. En cambio, las bases de datos que están en *TOXNET* proporcionan una colección informatizada de archivos en toxicología, elementos químicos peligrosos y sus áreas relacionadas. Algunas bases pertenecientes a este grupo son: *CCRIS; EMIC; GENE-TOX; IRIS; DART; ETICBACK; HSDB; TRI.* Por ejemplo: *CCRIS* (Chemical Carcinogenesis Research Information System), contiene citas sobre carcinógenos químicos, mutagens, promotores del tumor, e inhibidores del tumor; *GENE-TOX* (Toxicología Genética), trata sobre productos químicos que se probaron para la mutagenicidad. *IRIS* (Integrated Risk Information System), trata sobre químicos potencialmente tóxicos; *TRI* (Toxic chemical Release Inventory) series, trata sobre químicos tóxicos al ambiente, reciclados, etc.

Para acceder a las bases de datos de cualquier grupo, es necesario emplear un sistema de búsqueda y recuperación de información. El sistema más moderno pero también el más complejo se denomina ENTREZ. Otros sistemas aun en uso son: MEDLARS, el cual fue implementado en la década de los sesentas; y MEDLINE, que se empezó a usar en la década de los setentas.

El Sistema MEDLARS *(MEDical Literature Analysis and Retrieval System)* fue desarrollado por la National Library of Medicine, en la década de los sesentas. El cual, permite acceder a más de 18 millones de citas provenientes de los grupos ELHILL y TOXNET. Mientras que el sistema MEDLINE (*MEDlars online*) permite acceder a casi 12 millones de citas provenientes exclusivamente de las bases de datos bibliográficas denominada MedLine, OldMedLine y PreMedLine. Tanto el sistema MEDLARS como el sistema MEDLINE son usados por universidades, institutos, escuelas médicas, hospitales, agencias gubernamentales, organizaciones comerciales y no lucrativas de Estados Unidos y de otras partes del mundo. El sistema ENTREZ permite un acceso global a todas las bases de datos pertenecientes al National Library of Medicine, a toda persona, a toda institución que cuente con una conexión a Internet. El acceso global significa que se puede acceder a las bases de datos bibliográficas pertenecientes a PubMed (MedLine, OldMedLine, PreMedLine, PubMed Central) y a las bases de datos no bibliográficas pertenecientes al National Center for Biotechnology Information, NCBI.

En la tabla 11 se muestran algunas bases de datos no bibliográficas pertenecientes al National Center for Biotechnology Information, estas almacenan información secuencial de animales, humanos, proteínas, moléculas, etc.

En la ilustración 15 se muestra la integración de las bases de datos bibliográficas y no bibliográficas bajo el sistema ENTREZ. Cada base de datos es representada por un color, el color indica el número aproximado de registros en cada base de datos. Las líneas muestran las relaciones de la base de datos PudMed con el resto de ellas.

Nucleotide Databases	**Protein Databases**
dbEST dbGSS dbSNP dbSTS Nucleotide GenBank HomoloGene	Domains Proteins PROW RefSeq
Structure Databases	**Taxonomy Databases**
Domains 3D Domains Structure (MMDB)	Taxonomy
Genome Databases	**Expression Databases**
Cancer Chromosomes Gene Genomes LocusLink COGs	GEO GEO Datasets SAGE
Chemical Databases	
PubChem BioAssay PubChem Compound PubChem Substance	

Tabla 11: Las bases de datos no bibliográficas que pertenecen a NCBI.

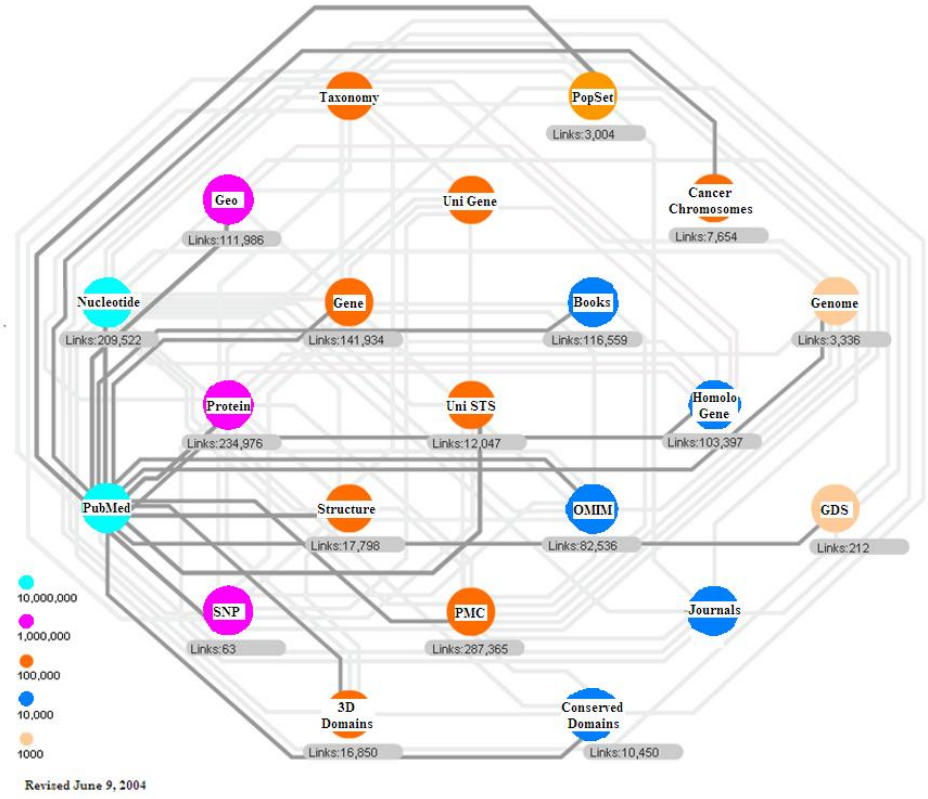


Ilustración 15: Red del sistema ENTREZ.

ENTREZ es un sistema complejo de búsqueda y recuperación de información que utiliza algoritmos muy potentes para realizar dichas tareas, por ello, National Library of Medicine, National Center for Biotechnology Information y algunos editores de revistas han desarrollado un subsistema de búsqueda y recuperación de información denominado *Entrez – PubMed* que permita acceder a las bases de datos bibliográficas que conforman a PubMed.

La base de datos PubMed cuenta con un conjunto de más de 17 millones de citas provenientes de MedLine, OldMedLine, PreMedLine y Pubmed Central. Entre las ventajas de PubMed se destacan:

- Los registros de MedLine se incorporan a PubMed semanalmente.
- Acceso gratuito a MedLine sin necesidad de registro, ni inclusión de contraseñas, por medio de *Entrez - PubMed*.
- Posibilidad de elegir entre varias pantallas o interfaces, con diferentes grados de dificultad y potencia de búsqueda.
- Enlaces con los textos completos de algunos artículos a través de las sedes web de los editores.
- Modalidades de búsqueda adicional como la posibilidad de buscar artículos relacionados a partir de un artículo encontrado en una búsqueda previa.
- Búsquedas clínicas a partir de filtros metodológicos preconfigurados.
- Utilizando sintaxis HTML, se pueden configurar enlaces directos con búsquedas bibliográficas de MedLine desde páginas web externas al sistema PubMed.
- Algunas citas de artículos aparecen en la base de datos de PubMed antes o al mismo tiempo que se publica el artículo. Esto se debe a que las editoriales de revistas científicas que participan en el proyecto aportan su información de forma directa. Además, si el editor tiene un lugar WWW que ofrece el texto completo de su revista, PubMed ofrece enlace a ese sitio de Internet.

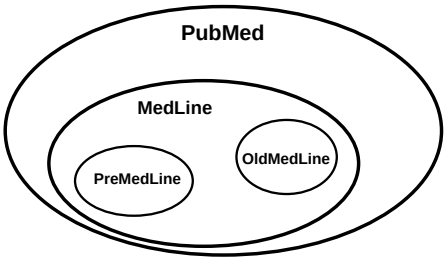


Ilustración 16: Las bases de datos que conforman Pubmed.

MedLine es la base de datos bibliográfica producida por National Library of Medicine. La cual, contiene aproximadamente 12 millones de citas bibliográficas a artículos de revistas sobre temas de biomedicina y sus temas afines. Las citas bibliograficas provienen de más de 4,600 revistas que cubren los temas de biomedicina principalmente, medicina, enfermería, odontología, oncología, medicina veterinaria, salud pública, ciencias preclínicas y de otras áreas de las ciencias de la vida. Este número reducido de revistas seleccionadas, también llamadas revistas MedLine, se debe a la característica principal de MedLine, que es la indización de la literatura biomédica. La indización obliga a National Library of Medicine a establecer ciertos parámetros que las revistas tienen que cumplir, dando como consecuencia, una selección de revistas muy especifica. Las revistas MedLine son editadas en 30 idiomas, para aquellas revistas no editadas en ingles, National Library of Medicine cuenta con equipos de profesionales que traducen los elementos bibliográficos de las revistas al idioma ingles. Como dato curioso, el 52% de las revistas en MedLine son editadas en los Estados Unidos. A la base

de datos MedLine, se le han agregado todas las citas publicadas en los antiguos Index Medicus, International Nursing Index y el Index to Dental Literatura. En general, estos tres índices, contienen citas de artículos de medicina, enfermería y odontología.

MedLine se actualiza de martes a sábado con alrededor de 2,000 citas por día, excepto en los meses de noviembre y diciembre debido a que La National Library of Medecine actualiza el MeSH Vocabulary, teniendo como consecuencia, que el proceso de indización de las revistas se detenga. El acceso a la base de datos MedLine puede hacerse por varias vías:

- Se puede acceder utilizando los servicios que ofrece National library of Medicine como son: Entrez - PubMed, ENTREZ y NLM Gateway.
- Por medio de desarrollos específicos para MedLine como son: Community of Science MedLine, Infotrieve.
- Por medio de portales médicos: Doctor's Guide, Medical World Search, MEDifusión, Medynet, Saludalia Médica, etc.
- Por medio de empresas distribuidoras de bases de datos como Dialog, DataStar, DIMDI, Knowledge Finger, Silver Platter, etc.; que la ofrecen en cd-rom y a través de Internet; además, proporcionan al usuario una gama de herramientas para la búsqueda y recuperación de la información.

Cabe destacar, que MedLine no almacena el texto completo de las revistas, sino que almacena los "registros" elaborados a partir de los elementos bibliográficos que contienen las revistas. Y estos registros posteriormente forman las "citas" que visualizamos en el subsistema Entrez - PubMed. Por eso es muy común decir que MedLine almacena o contiene citas. Para ver, los textos completos de los artículos, MedLine ofrece el servicio MedLine Plus o PubMed Central. Estos registros han sido almacenados y segmentados en los llamados MedLine Backfiles: MEDLINE, MED90, MED85, MED80, MED75 y MED66. Entendiéndose que los registros actuales se almacenan en MEDLINE y los registros más viejos están almacenados en el MED66. Además de estos grupos de registros, MedLine esta complementada con otras dos bases de datos, llamadas OldMedLine y PreMedLine. Estas dos bases de datos, juegan un papel importantes durante el proceso de indización realizado por National Library of Medicine.

OldMedLine contiene las citas publicadas entre los años de 1953 y 1965. Cubriendo así, los campos de la medicina, ciencias preclínicas, y ciencias ligadas a la salud de aquellos años. Hay aproximadamente 1.5 millones de citas OldMedLine que no incluyen resumen, tampoco han sido indexados con los términos MeSH; pero sus datos bibliográficos sin han sido revisados y aprobados por MedLine. A mediados de la década de los noventa, se creó la base de datos PreMedLine, con el objetivo de almacenar las citas "in process", es decir, las citas que se encuentran en la sección de Indizado del National Library of Medicine. Una vez que estas citas han sido indizadas exitosamente, entonces, reciben un número de identificación PMID (PubMed Identify), se suprimen de PreMedLine y se incorporan en formato completo a MedLine. Algunas de las citas provienen directamente de los editores de las revistas, mientras que otras, son incorporados por los equipos del National Library of Medicine. En la siguiente sección se describe brevemente el sistema Entrez-PubMed, el cual nos permitirá tener acceso a PubMed.

3.4.-El Sistema Entrez–Pubmed

National Library of Medicine pone a disposición el sistema Entrez–Pubmed para la búsqueda y recuperación de las citas bibliográficas localizadas en PubMed. El servicio es gratuito y está disponible en la página electrónica *http:/ www.ncbi.nlm.nih.gov /.* El núcleo del sistema Entrez–Pubmed es el Algoritmo Mapeo Automático de Términos[18]. Básicamente, el

[18] Automatic Term Mapping

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:

- *Tabla de traducción MeSH:* Contiene una lista alfabética de los términos MeSH; los sinónimos, las referencias cruzadas y los términos de entrada para los términos MeSH; los tipos de publicaciones; términos derivados del Sistema de Lenguaje Médico Unificado; Los conceptos de nombres suplementarios correspondientes a los nombres de sustancias y sus sinónimos.
- *Tabla de traducción de revistas:* Contiene un listado alfabético de todos los títulos de las revistas; las abreviaturas de revistas en formato MedLine; el número de identificación unívoco de una revista: International Standard Serial Numbers, ISSN.
- *Lista de frases:* Contiene una lista de frases derivadas del Sistema de Lenguaje Médico Unificado; nombres de sustancias.
- *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*. 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:

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 instrucción *All Fields* busca en todos los campos (Tags) del formato MedLine. Los campos de un registro bibliográfico, se identifica mediante una etiqueta de dos o más letras (calificadores de campo).

Algunos 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 12: Algunos campos del formato MedLine.

Las etiquetas se emplean para realizar búsquedas especificas, por ejemplo, si se busca *"mycobacterium"* como término MeSH basta escribir lo siguiente *"mycobacterium bovis [mh]"*. Realizar búsquedas por medio de las etiquetas es muy complicado, por ello, solamente personas calificadas realizan este tipo de búsquedas. Por ejemplo,

(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 revistas. Este aspecto del formato MedLine es muy importante ya que puede ser procesado con cualquier programa de gestión de documentos (por ejemplo, el programa ProCite) para recuperar los campos que se vayan a utilizar en los análisis cienciométricos.

3.6.-MeSH Vocabulary

MedLine es una base de datos bibliográficos[19] producida por la National Library of Medicine de 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 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 diferentes términos del MeSH Vocabulary[20] que identifican el contenido o los contenidos de cada documento en PubMed. 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.

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:

- *MeSH Headings.* Representan conceptos o temas generales que se encuentran en la literatura biomédica. Algunos ejemplos de MeSH Headings son los siguientes: body weight, dental cavity preparation, radioactive waste, kidney, self medication, brain edema, etc.
- *MeSH Subheadings.* Son palabras o frases, con las cuales, se califica un MeSH Headings, es decir, estas palabras o frases se usan para caracterizar a los temas generales en sus aspectos más específicos. Algunos ejemplos de MeSH Subheadings son los siguientes: diagnosis, surgery, metabolism, pathology, etc.
- *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

[19] 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.
[20] Acrónimo de **Me**dical **S**ubject **H**eadings Vocabulary.

Un aspecto importante del MeSH es su *estructura jerárquica*. En esta estructura en forma de árbol (MeSH Tree Structure), los términos MeSH se ramifican en series de términos cada vez más concretos o específicos. La tabla 13 muestra las 16 categorías del MeSH Vocabulary 2008.

	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	Natural Sciences
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology, Industry, Agriculture
K	Humanities
L	Information Science
M	Named Groups
N	Health Care
V	Publication Characteristics
Z	Geographicals

Tabla 13: Categorías del MeSH Vocabulary.

A continuación 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».

Face
- Cheek
- Chin
- Eye
  - Eyebrows
  - Eyelids
  - Eyelashes
- Forehead
- Mouth
  - Lip
- Nose

Ilustración 17: Encabezados y subencabezados.

Conozcamos la relación entre Headings y Subheadings en el MeSH Tree Structure. En la ilustración 17 se muestra una parte de la categoría A, la cual incluye a "Face" (Rostro). Por ejemplo, "Eye" *(Ojo)* se considera MeSH Headings mientras que "Eyebrows" *(Ceja)* y

"Eyelids" *(Párpado)* son sus correspondientes MeSH Subheadings. Nótese que los MeSH Subheadings están debajo de los MeSH Headings.

Ahora se presenta en forma sencilla como indizan los documentos con el MeSH Vocabulary. 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 asigna los siguientes Headings MeSH y Subheadings MeSH a los 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 Headings MeSH. Mientras que los términos después de la diagonal invertida representan a los Subheadings MeSH. Un Headings MeSH puede tener varios Subheadings MeSH. El símbolo asterisco (*) se utiliza para distinguir la importancia de los términos dentro de los documentos, es decir, el término 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 término MeSH ser *"Major"*, la toma el grupo de personas encargadas de la indización después de analizar el artículo.

National Library of Medicine indiza varios tipos de documentos, por ejemplo, Ensayos Clínicos, Editoriales, Cartas, Meta Análisis, Normas de Practica, Revisiones, Ensayos Controlados Aleatorizados, Noticias, Diccionarios, Estudios Gemelos, Festschrift, etc., A continuación se hace una breve descripción de algunos documentos.

Ensayo Clínico: 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: 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: 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 va acompañada de comentarios sobre el artículo.

Meta Análisis: 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: 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 desarrollados 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: 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 numeros 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: 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.

En el capítulo 5, solamente se consideran los siguientes documentos: los ensayos clínicos, los meta análisis y los ensayos controlados aleatorizados, para realizar el Análisis de Palabras Asociadas.

3.5.-Ventajas de PubMed para el Análisis Cienciométrico

Como se ha mencionado la validez del análisis cienciométrico 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 PubMed 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 los campos bibliográficos.
- 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.

Se debe tener cuidado en campos como: nombres de autores, nombre de institución, citas, referencias, etc., ya que cada autor escribe algunos elementos distintos aún en sus propios artículos.

4 Mapas Auto - Organizantes

El algoritmo SOM[21] de Teuvo Kohonen es muy utilizado debido a su eficiencia para llevar a cabo las siguientes tareas:

- *Conglomerado:* 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 bidimensional. El mapa 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. 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 el algoritmo SOM.

4.1.-Elementos de las Redes Neuronales

Después de la segunda mitad del siglo XX, los investigadores en las áreas de Inteligencia Artificial, Aprendizaje Maquina, etc., han tratado de imitar, por medio de simulaciones por computadora el funcionamiento del cerebro humano. Hasta la fecha, la mejor simulación del cerebro humano son las *Redes Neuronales Artificiales*. 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, [Villaseñor, 2004]. En la ilustración 18 se muestran algunas partes que integran a una neurona biológica.

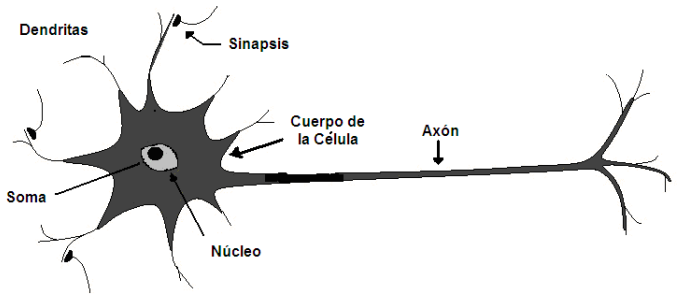


Ilustración 18: Esquema de una neurona biológica.

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. 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.

[21] *Self-Organizing Maps.*

La ilustración 19 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*.

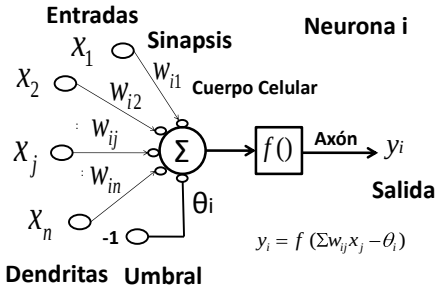


Ilustración 19: Modelado una red neuronal artificial.

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: Feedforward Networks (Hacia - Adelante) y Feedback Networks (Retroalimentación). Como se aprecia en la ilustración 20, 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 sí. 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.

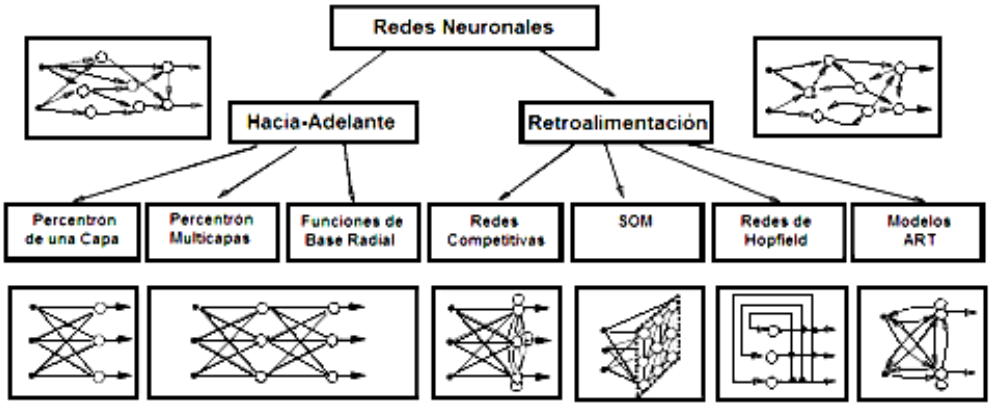


Ilustración 20: Las arquitecturas más representativas de cada categoría.

El Proceso de Aprendizaje. A groso modo, las redes neuronales artificiales simulan al proceso de aprendizaje del cerebro humano, por medio de pesos sinápticos *w*. Veamos brevemente en qué consiste la simulación. Supongamos que tenemos una neurona *i* como la de la ilustración 21, 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 ilustración 21, siendo θ_i el umbral.

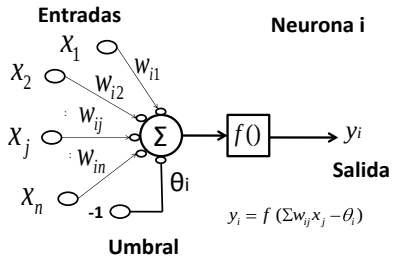


Ilustración 21: La neurona artificial.

Esta simulación se sustenta en las ideas de Hebb, [Hebb, 1949]. 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.

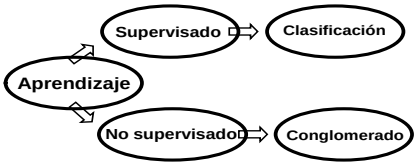


Ilustración 22: Paradigmas de Aprendizaje.

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.

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, se mencionan 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. El conglomerado 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 sí 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 los conceptos es: 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 conglomerados. 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.

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.

4.2.-Mapas Auto Organizantes

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. 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.

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 ilustración 23 se aprecia que las capas están conectadas entre sí por medio de *pesos sinápticos*. La idea básica del modelo es, a partir 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 auto-organizante* donde se representan los rasgos más sobresalientes del espacio de entrada.

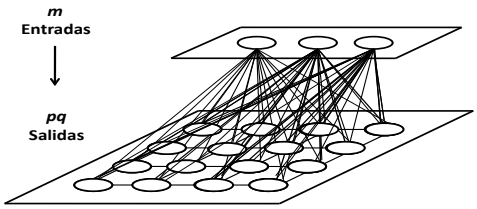


Ilustración 23: La estructura del SOM.

A continuación se presenta el funcionamiento del algoritmo SOM. El SOM constituye un importante ejemplo dentro del contexto del paradigma de las redes neuronales. Para determinar una red neuronal es necesario definir: las neuronas, la arquitectura y el algoritmo de entrenamiento. Por esta razón, la presentación del algoritmo SOM se establece mediante la definición de estos aspectos, [Villaseñor, 2004], [Kohonen, 2001], [Kohonen, 1998] y [Kaski, 1997].

Arquitectura. Sea $X = \{x_1, x_2, \dots, x_m\}$ el conjunto de datos de entrada, cada x_i se denomina *componente* y cada componente se representa como $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, es decir $x_i \in \mathbb{R}^n$. Para simplificar la notación se escribirá $(x_i =)\ x \in \mathbb{R}^n$ y se dirá que *x* es el *vector de entrada* o simplemente *el dato de entrada*. 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 $x \in \mathbb{R}^n$ e interactúan entre ellas por medio de relaciones laterales que se activan durante la actualización de los pesos. Estas relaciones responden a la relación (ilustración 24) de distancia física entre una neurona y sus 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:

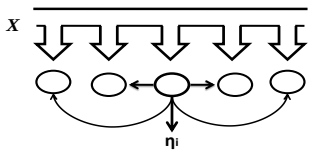


Ilustración 24: Representación de una neurona y sus conexiones con la entrada *x* y las neuronas vecinas.

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

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 las siguientes características:

- 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 se denomina *vector de referencia*.

En la ilustración 25 se muestran las configuraciones o tipos 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. 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.

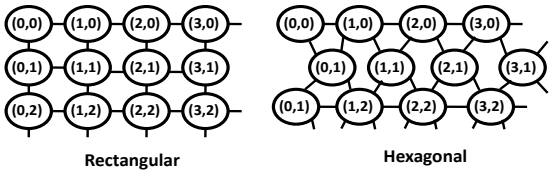


Ilustración 25: Configuraciones más comunes en la retícula del SOM.

Entrenamiento. Ahora se detalla en qué 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 cuál

de ellas es capaz de representar de mejor manera al dato de entrada *x(t)*. En los métodos de Cuantización esto se determina de la siguiente manera: dado $x(t) \in \mathbb{R}^n$, el vector $\eta_{c(x)}$ es tal que

$$d\left(x, \eta_{c(x)}\right) = min\{d(x - \eta_i): i = 1, \dots, k\}$$

Donde *d* es una función de distancia. Nótese que el subíndice *c* es función de *x*; para cada *x* existe un $\eta_{c(x)} \in \Im$. En caso de que este índice no esté bien definido, es decir cuando para un dato *x* existan dos $\eta_e, \eta_d \in \Im$, tal que:

$$d(x, \eta_e) = d(x, \eta_d)$$

La selección de un único *c(x)* debe hacerse de manera aleatoria. Asi, cada elemento del conjunto *X* es asociado a un elemento de $\Im$, es decir, cada elemento de $x \in X$ queda representado por $\eta_{c(x)} \in \Im$. Por simplicidad esta asociación se denotará de la siguiente manera:

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

Generalmente, se utiliza la distancia euclidiana para determinar el nodo que mejor representa a un dato. En general, las *normas* – L_r se definen como:

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

Donde *r* es un número real positivo. Para el caso $r = 1$ la distancia es denomina métrica Manhattan. Sin embargo, se pueden usar otras funciones de distancia si el problema lo requiere. Para variables que son medidas en unidades que no son comparables en términos de escala, la siguiente función de distancia es especialmente apropiada:

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

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 distancia Euclidiana ya que esta última se obtiene cuando $B = I$ donde *I* es la matriz identidad en $\mathbb{R}^n$.

En el SOM dado cualquier $x \in X$, la competencia consiste en encontrar la neurona η_c 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*. 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 (vecindad de Voronoi) de $\eta_{c(t)}$ en el arreglo bidimensional (ver ilustración 26) 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:

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

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, (ver ilustración 26):

1. $Si\ t_i \leq t_j \Longrightarrow \rho(t_i) \geq \rho(t_j)$
2. $Si\ \rho(t) \to 0\ cuando\ t \to \infty$

Debe tenerse cuidado al escoger el tamaño inicial de *ρ(0),* si desde el 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 *ρ(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 $w_i(0)$.

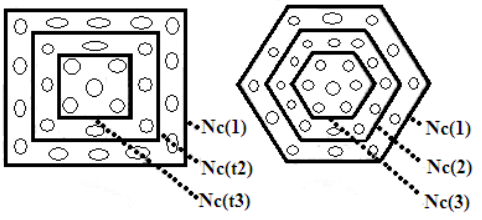


Ilustración 26: Variación en el tiempo del radio de la vecindad.

En las versiones más simples del SOM los valores sucesivos para los vectores de referencia se determinan recursivamente por el siguiente mapeo de iteraciones:

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

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 *ρ(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 $h_{ci}(t) \to 0$ mientras $\|r_{c(t)} - r_i\|$ se incrementa. Una de las definiciones más simples que se encuentran de la función vecindad es la siguiente:

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

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

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

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

Ordenamiento Global: Según lo reportado por Teuvo Kohonen, [Kohonen, 2001] durante aproximadamente las primeras 1,000 competencias se lleva a cabo 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 iníciales de los pesos han sido seleccionados de manera aleatoria, durante estos primeros 1,000 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 final. 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 100,000, sin embargo en ciertas aplicaciones, como el reconocimiento de voz, es de alrededor de 10,000. 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.

4.3.-El SOM y la 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.

El SOM proporciona un mapa de baja dimensión del espacio de datos. El objetivo de la visualización es entender el área mapeada y validar la investigación de nuevas muestras de datos con respecto al área mapeada. Para entender lo que realmente muestra el SOM, es importante entender que realiza dos tareas: vector de cuantización y vector de proyección. El vector de cuantización crea de los datos originales un conjunto de datos más pequeño pero representativo, a ser trabajados. El conjunto de vectores prototipos refleja las propiedades del espacio de datos. La proyección hecha por el SOM es no lineal y restringida a una malla regular (la cuadrícula del mapa). El SOM trata de preservar la topología del espacio de datos más que las distancias relativas.

En contraste, hay muchas formas de proyectar datos multidimensionales a dimensiones menores. Un método bien conocido es el Análisis de Componentes Principales, ACP. La idea central del Análisis de Componentes Principales es reducir la dimensión de un conjunto de datos que consiste de un gran número de variables interrelacionadas; la redución debe retener la mayor variación posible presente en el conjunto de datos. Las variables se transforman a un nuevo conjunto de variables denominados *los componentes principales*, que son no correlacionados y son ordenados de tal forma que el primero retiene mucha de la variación

presente de todas las variables originales, [Jolliffe, 1986]. Esta operación lineal es rápida, pero da resultados engañosos si se ignoran direcciones con información significativa. Otro método de proyección consiste en mantener las distancias relativas entre muestras tan cercanas a las originales según una función de costo. Diferentes funciones de costo conducen a distintos algoritmos no lineales de proyección, i.e. proyección de Sammon o el Análisis de Componentes Curvilinear (CCA). Grandes conjuntos de datos causan a menudo problemas a los métodos de proyección, en especial a los métodos de proyección iterativos puesto que los procedimientos se vuelven computacionalmente pesados. Una posibilidad es reducir la tarea computacional, primeramente cuantizando los datos usando algún método propio, i.e., K-means y entonces aplicamos el método de proyección. Por su puesto, esto es similar a lo que el SOM hace, excepto que el SOM lo hace simultáneamente más que secuencialmente y solamente la topología (no distancias) es preservada.

Métodos Básicos para la Visualización SOM. La cuadrícula SOM proporciona una base para varios tipos de visualizaciones. Los valores de las variables u otras características pueden mostrarse con respecto a la cuadrícula, [Himberg et al., 2001].

Unified Distance Matrix. La Unified Distance Matrix (U-Matrix) es una herramienta simple y efectiva para mostrar la posible estructura del conglomerado sobre la cuadrícula SOM. Muestra las distancias entre unidades vecinas usando una representación de escala gris sobre la cuadrícula del mapa. Esto da una impresión de "montañas" (distancias prolongadas) que dividen el mapa en "campos" (partes densas, i.e., conglomerados). Vea ilustración 27 (c).

Planos Componentes. El SOM a menudo se "divide" en planos componentes para ver como los valores de un cierto componente (variable) varían sobre diferentes posiciones del mapa. Cada plano representa el valor de una variable (componente) del vector prototipo en cada nodo del SOM usando una representación de escala gris. Uno puede ahora ver el comportamiento general de los valores componente en diferentes partes del SOM. Vea ilustración 27 (c). Los planos componentes juegan un papel importante en la detección de correlación: comparando estos planos aun variables correlacionadas parcialmente pueden ser detectadas por inspección visual. Esta clase de comparación podría hacerse usando graficas de dispersión tan bien, pero esto requerirá una cantidad cuadrática de imágenes con respecto al número de variables: cada variable en contra de las otras variables. Cuando usamos los planos componentes el número de imágenes crece linealmente. Además, el vector de cuantización ejecutado por el SOM remueve ruido. Los planos componentes pueden también ser fácilmente comparados con la representación de conglomerados de la U-Matrix.

Hits. Cuando investigamos nuevos datos con el SOM, la pregunta es: ¿Qué parte del mapa representa mejor a los datos? Tradicionalmente, esto ha sido respondido encontrando el vector prototipo más cercano (the best matching unit, BMU) para cada dato investigado y entonces indicándolo desde el SOM. Vea ilustración 27 (c). Para múltiples vectores, un histograma se obtiene. Comparando diferentes histogramas se puede evaluar la semejanza de diferentes conjuntos de datos en términos del mapa.

Trayectorias. Si los datos han sido adquiridos de un proceso, uno puede estar interesado en visualizar la evolución del estado del proceso en el tiempo. El BMU del actual vector de características puede ser considerado como el punto operacional sobre el mapa el cual a su vez puede ser considerado como una proyección del espacio de estado multidimensional. La trayectoria (ilustración 27 (d)) es una línea que conecta una sucesión de puntos operacionales que muestra los cambios del proceso en el tiempo.

Combinando Diferentes Proyecciones. Para obtener una idea de la forma del mapa en el espacio de datos, los vectores prototipos del SOM pueden ser proyectados a baja dimensión usando algún método de proyección, que trate de preservar distancias entre puntos proyectados.

Una practica común es usar la proyección de Sammon y mostrar las relaciones topológicas del mapa según conectando puntos que corresponden a unidades vecinas.

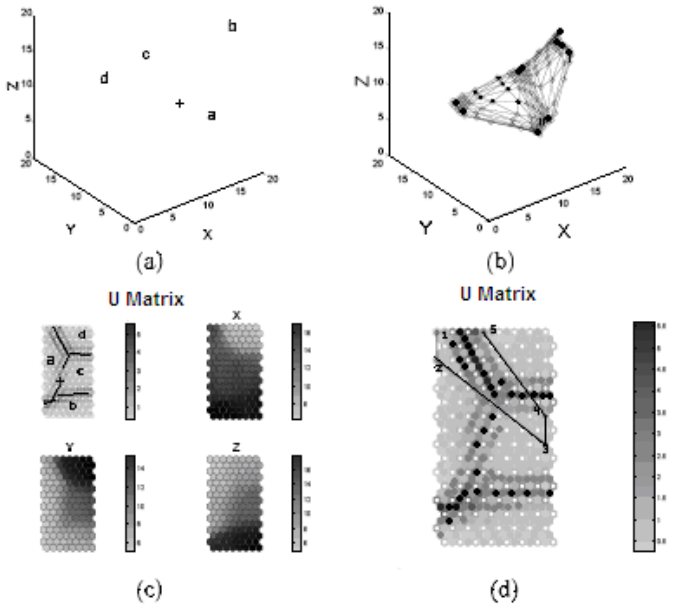


Ilustración 27: Visualización SOM. La figura (a) muestra un conjunto de datos artificiales en dimensión tres con cuatro conglomerados (a, b, c, d). La figura (b) muestra los vectores prototipos de un entrenamiento SOM con los datos de (a). Las conexiones topológicas se muestran también. La figura (c) muestra la U-Matrix y planos componentes. Grises oscuros representan extensas distancias inter-unidad y grises claros distancias cortas. Los conglomerados –que pueden ser vistos como claros "campos" entre las oscuras "montañas" –han sido etiquetados por conveniencia. El signo + muestra el BMU para la muestra marcada por signo + en (a) (localizada entre los conglomerados a y c). Los planos componentes muestran como las variables X, Y y Z varían a lo largo del mapa. La figura (d) muestra una trayectoria de cinco muestras en la U-Matrix.

El SOM puede ser considerado inestable si las estructuras topológicas están torcidas o plegadas. La ilustración 27 (b) es un ejemplo de una proyección lineal de vectores prototipos y sus conexiones topológicas. Para clarificar las conexiones entre visualizaciones, pueden vincularse usando color, que es una ayuda visual dominante para agrupar objectos. Aplicamos esta idea para vincular diferentes presentaciones de los mismos datos, i.e. la cuadrícula SOM y el diagrama de dispersión o la proyección de Sammon. Vea ilustración 29.

Cazando Correlación. Correlaciones entre parejas de componentes son reveladas como patrones similares en posiciones idénticas de los planos componentes. La detección de correlación puede ser hecha fácilmente si los planos componentes son reorganizados tal que la posiblemente correlaciones son presentadas cerca una de otras. Ilustración 28.

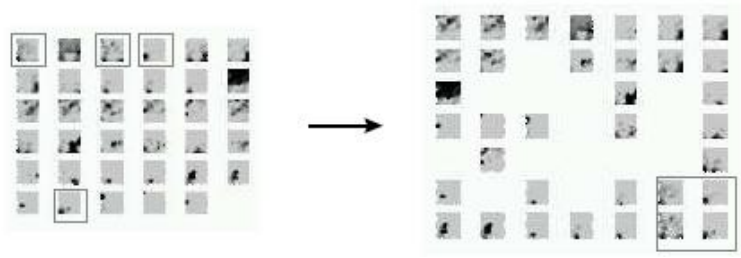


Ilustración 28: Planos Componentes. Las correlaciones entre componentes pueden cazarse de la visualización de los planos componentes de la izquierda. La tarea es muy fácil si los planos son reorganizados tal que los planos componentes que parezcan tener alta correlación son puestos en lugares cercanos a otros, como se muestra en el lado derecho. Por ejemplo, esta reorganización muestra cuatro componentes correlacionados.

Usando los planos componentes para cazar correlación de esta manera es fácil pero también bastante vago y algunas veces aun engañoso. Sin embargo, es fácil seleccionar combinaciones de interesantes componentes para investigación adicional. Un estudio más detallado de interesantes combinaciones puede hacerse usando diagramas de dispersión, que puede ser vinculados a la unidades del mapa por color como se hace en muchos casos de estudio.

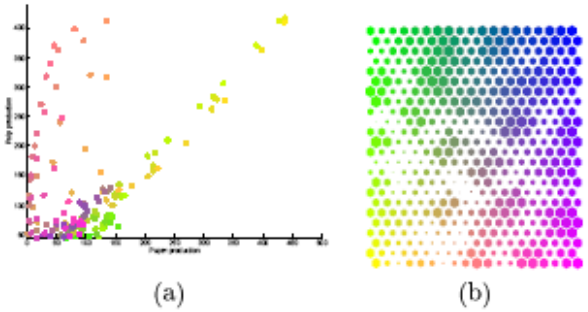


Ilustración 29: Correlaciones entre componentes. Las correlaciones entre los componentes del vector pueden ser eficientemente visualizados usando diagramas de dispersión. En figura (a) cada punto corresponde a una unidad del mapa. Las coordenadas *x* y *y* de los puntos han sido tomadas de dos componentes de los vectores prototipos. Para enlazar el diagrama de dispersión a otras visualizaciones, cada punto está dado por un color según el código de color de las unidades del mapa mostrado en Figura (b). Además, el código de color, Figura (b) también usa el tamaño para indicar el conglomerado sobre el mapa: unidades pequeñas corresponden a conglomerados fronterizos. Puede ser visto que para muchas unidades, especialmente esas con código de color amarillo, los dos componentes son linealmente correlacionados pero hay distintas excepciones.

Descubriendo Novedad. Cuando investigamos nuevos datos con el SOM, el BMU de cada muestra de datos es encontrado e indicado en el mapa. El problema con este simple tema es que no se obtiene información de la exactitud de la compatibilidad. Típicamente, hay muchas unidades con tan buenas compatibilidades como el BMU. Alternativamente, la muestra de datos puede estar muy lejos del mapa –una novedad en términos del mapa. En lugar de simplemente señalar el BMU, la respuesta de todas las unidades del mapa a los datos puede mostrarse. La resultante respuesta superficie muestra la relativa bondad de cada unida del mapa en representar los datos. La respuesta puede ser una función del error de cuantización como:

$$g(x, m_i) = 1/(1 + (q_i/a)^2) \qquad [[6]]$$

Donde $q_i = \| x - m_i \|$ es el error de cuantización, es decir, distancia, entre la muestra *x* y la unidad del mapa *i*. El factor de escala *a* es la distancia promedio entre cada muestra de datos

entrenada y su BMU. Vea ilustración 30 (a). Quizás una más interpretativa función de respuesta resulte si el SOM es usado como una base para reducir la densidad de kernel evaluada de los datos. Entonces uno puede estimar la probabilidad $P(i/x)$ de cada unidad del mapa represente la muestra de datos.

En ambos casos, la respuesta superficie es añadida hacia los mapas después, mientras que el algoritmo SOM original tiene un *"crisp"*, el ganador toma todas las funciones de activación. Hay algoritmos que tienen un subordinador probabilístico intrínseco como los S-Map. Sin embargo, parece que un modelo de evaluación para la densidad de kernel añadido al SOM da resultados comparables con esos métodos. Otra manera de mostrar la exactitud de la compatibilidad es usar i.e., el tamaño de la muestra marcada. En la ilustración 30 (b), la función de respuesta borrosa (ecuación [[6]]) ha sido usada para controlar el tamaño de las muestras marcadas (círculos). Ahora muestras individuales pueden ser vistas a lo largo con sus BMU´s (posición) y exactitud (tamaño).

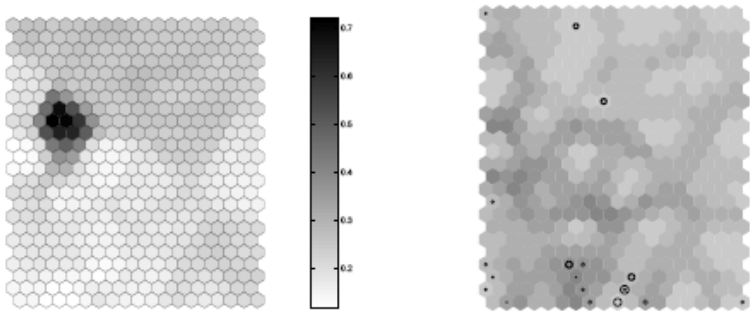


Ilustración 30: Exactitud de compatibilidad. La figura (a) muestra la respuesta superficie (ecuación [[6]]) para un dato de la muestra. La figura (b) muestra los BMU´s y la correspondiente exactitud de 20 muestras. La textura del fondo es el promedio de la U-Matrix del SOM. Cada círculo representa una muestra. La posición de los círculos indica el BMU, y su tamaño la exactitud de la compatibilidad.

Auto – Organización: 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.

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". 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.

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 graficas 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"*. 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 cienciométricos.

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 conglomerado sobre los vectores de referencia. Algunos métodos de conglomerados utilizados son el *SOM-Ward Clusters, SOM-Single-Linkage Clusters, etc.*

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 $\boldsymbol{x}$ 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 se desea, los datos extremos pueden ser descartados y el análisis puede continuar con el resto del conjunto de datos.

5 Análisis de Palabras Asociadas

El objetivo de este capítulo es demostrar el interés existente sobre el Análisis de Palabras Asociadas. Dos limitaciones deberán considerarse cuando se lea este capítulo. No se realiza un análisis exhaustivo del campo biomédico, más bien, se presenta la técnica, se muestra como puede ser usada y se dan ejemplos para ilustrar como diferentes tipos de resultados pueden ser interpretados. Solamente, nos concentramos en una parte muy limitada de la red para simplificar el capítulo.

El fin del Análisis de Palabras Asociadas es producir las relaciones entre descriptores que podrían en un momento dado ser considerados como las más significativas, (Courtial et al., 1991A). Se usan los descriptores MeSH asignados a documentos MedLine y se mapea la literatura analizada según la fuerza de co-ocurrencia de los descriptores MeSH. La literatura analizada consiste de un concepto principal pero en ella hay otros conceptos principales no mencionados. Swanson (Stegmann, 2003) se refiere a este tipo de literatura como literatura CBD (Complementary But Disjoint).

El Análisis de Palabras Asociadas fue desarrollado para mapear la estructura y dinámica del campo tecno-científico y es una herramienta potente para descubrir conocimiento. El concepto "Dinámica" aparece en un sin fin de problemas técnicos de la física, química, biología, sociología, economía, etc. En particular, el área de la biomedicina ha mostrado una marcada tendencia a considerar los aspectos dinámicos de ciertos procesos fisiológicos, ya sea humano o animal. Desde este punto de vista, variaciones súbitas en los factores físicos o químicos del proceso fisiológico, pueden provocar cambios cualitativos en la dinámica del proceso fisiológico correspondiente. Además, algunos procesos fisiológicos pueden presentar caos, representaciones apropiadas de procesos caóticos usualmente revelan auto-similaridad en el tiempo. Por ejemplo, la dinámica no lineal y el caos están presentes en la variabilidad del ritmo cardíaco, y formas fractales están presentes en las vías de ventilación de los pulmones. Se quieren localizar conglomerados o sub-redes que estén firmemente encadenados sus descriptores entre si y que correspondan a centros de interés o a problemas de investigación biomédica asociados al MeSH Major Topic "Nonlinear Dynamics". Estos centros de interés contemplan los objetos de inversión significativa por parte de los investigadores.

5.1.-Selección de Documentos

Los documentos MedLine son aptos para los análisis cienciométricos, están normalizados por campos, existe gran diversidad y se accede a ellos gratuitamente.

Recuperación Online. La búsqueda online se realizo en PubMed, la versión web de MedLine, usando el sistema de recuperación de la US National Center for Biotechnology Information denominado Entrez. La búsqueda se realizo por descriptor MeSH. El conjunto de documentos recuperados fue bajado en formato MedLine de PubMed usando su opción *save*.

Descriptor MeSH. MedLine contempla el concepto "Dinámica no Lineal" como descriptor MeSH valido en la indización de publicaciones biomédicas. En este trabajo, se consideran las publicaciones biomédicas indizadas con el concepto "Dinámica no Lineal" en su carácter de *MeSH Major Topic* en los últimos 4 años. Las siguientes sintaxis se emplearon para recuperar los documentos:

- "Nonlinear Dynamics"[MAJR] AND "2004"[MHDA] NOT "Animals"[MH] NOT "Editorial"[PT] NOT "Letter"[PT] NOT "Practice Guideline"[PT] NOT "Review"[PT]
- "Nonlinear Dynamics"[MAJR] AND "2005"[MHDA] NOT "Animals"[MH] NOT "Editorial"[PT] NOT "Letter"[PT] NOT "Practice Guideline"[PT] NOT "Review"[PT]
- "Nonlinear Dynamics"[MAJR] AND "2006"[MHDA] NOT "Animals"[MH] NOT "Editorial"[PT] NOT "Letter"[PT] NOT "Practice Guideline"[PT] NOT "Review"[PT]
- "Nonlinear Dynamics"[MAJR] AND "2007"[MHDA] NOT "Animals"[MH] NOT "Editorial"[PT] NOT "Letter"[PT] NOT "Practice Guideline"[PT] NOT "Review"[PT]

Para limitar la cantidad de documentos a analizar, solamente se consideraron estudios en humanos sin importar la edad (existen estudios en animales). Los documentos denominados ensayos clínicos, meta análisis y ensayos controlados aleatorizados se consideran fuentes importantes para reportar esta clase de estudios al contrario de los documentos denominados editorial, carta, normas de práctica y revisión. En la siguiente tabla se muestran la cantidad de documentos recuperados entre los años 2004 y 2007.

Corpus	Año de indización	Total de documentos
	2002	40
	2003	78
C1	2004	144
C2	2005	95
C3	2006	108
C4	2007	115

Tabla 14: Total de documentos recuperados entre los años 2002 y 2007.

5.2.-Preprocesamiento de los Corpus

El preprocesamiento consistió en reacomodar los descriptores MeSH de los corpus C1, C2, C3 y C4. En el formato MedLine, los descriptores MeSH aparecen ordenados de la siguiente forma:

MH - Animals
MH - Central Nervous System Stimulants/*pharmacology
MH - Computer Simulation
MH - Diptera/*physiology
MH - Female
MH - Models, Neurological
MH - Motion Perception/drug effects/*physiology
MH - Neurons/drug effects/*physiology
MH - *Nonlinear Dynamics
MH - Picrotoxin/*pharmacology
MH - Vision/drug effects/*physiology

Donde MH denota descriptor MeSH. El término después de **"-"** denota Headings MeSH. El término después de **"/"** denota Subheadings MeSH. El asterisco indica que el Heading MeSH o el Subheading MeSH es considerado MeSH Major Topic. Reacomodar los descriptores MeSH consiste en ponerlos en una lista separada por **"/"**, así:

Animals/Central Nervous System Stimulants**:**pharmacology/**……../** Vision**:**drug effects**:**physiology

Se cambia "/" por "**:**" para distinguir el descriptor principal de sus respectivas especificidades y se elimina "*****", esto último, para no tener simbología en los descriptores y hacerlos más legibles a la vista. De esta forma se conserva el descriptor principal junto con sus respectivas especificidades. Este preprocesamiento se llevo a cabo con Procite for Windows, version 5.0 y Microsoft Excel 2003.

5.3.-Minería de Datos y Visualización de Conglomerados

En estas dos etapas se usa el programa Redes 2005: Análisis de Redes Tecnocientíficas, desarrollado para el proyecto CognoSfera por el Dr. Rafael Bailón Moreno y Rosario Ruiz Baños. Redes 2005 utiliza el algoritmo de agrupación sobre centros simples. Posteriormente, se usara el programa Viscovery SonMine para obtener una representación visual alterna de los conglomerados de la red.

5.4.-Análisis Estático de la Red

Considere que la red evoluciono durante los años 2004, 2005, 2006 y 2007. En este breve periodo de tiempo, los conglomerados que integran la red pudieron conservarse, desvanecerse, partirse o fusionarse. Esta dinámica de conglomerados afecta la estructura de la red durante el transcurso del tiempo. El análisis estático de la red consiste en identificar las relaciones entre los conglomerados y así poder cuantificar su importancia en la estructura de la red.

Para la construcción de los conglomerados, se fijo el umbral de saturación en 10, es decir, los conglomerados contendrán un máximo de 10 temáticas encadenadas entre sí. Para descartar cadenas con índices de equivalencia por debajo de 0.006, se fijaron los valores de ocurrencia mínima y de co-ocurrencia mínima en 2 y 2 respectivamente. Nota: En el Apéndice C se muestran los índices de equivalencia entre los encadenamientos internos y externos de los conglomerados que integraron la red biomédica durante los años 2004, 2005, 2006 y 2007.

Red en el año 2004 asociada a C1. La red asociada al corpus C1 consiste de 8 conglomerados. "Motion" y "Nonlinear Dynamics" son conglomerados principales. Solamente, "Nucleic Acid Conformation" y "Models, Neurological" comparten cadenas externas con ambos conglomerados principales. Aunque estos no comparten cadenas externas entre sí. El conglomerado "Models, Genetic" no comparte cadenas externas con ninguno de los otros conglomerados, es conglomerado aislado.

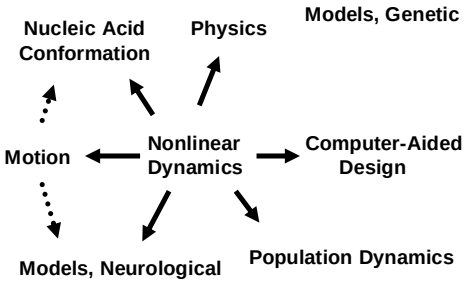


Ilustración 31: Red asociada al corpus C1. *"Nonlinear Dynamics"* representa un conglomerado principal con conglomerados secundarios *"Motion", "Physics", "Models, Neurological", "Population Dynamics", "Computer-Aided Design"* y *"Nucleic Acid Conformation"*. El conglomerado *"Motion"* es principal con conglomerados secundarios *"Nucleic Acid Conformation"* y *"Models, Neurological"*. Mientras que *"Nucleic Acid Conformation", "Computer-Aided Design", "Models, Neurological", "Physics", "Population Dynamics"* y *"Models, Genetic"* son conglomerados aislados.

Para la construcción de los conglomerados se fijo el umbral de saturación en 10 y los valores de ocurrencia mínima y de co-ocurrencia mínima en 2 y 2 respectivamente. Con estos valores fijos se identificaron 320 temáticas, de las cuales 103 temáticas superan los valores de ocurrencia mínima y de co-ocurrencia mínima y 217 temáticas se descartaron.

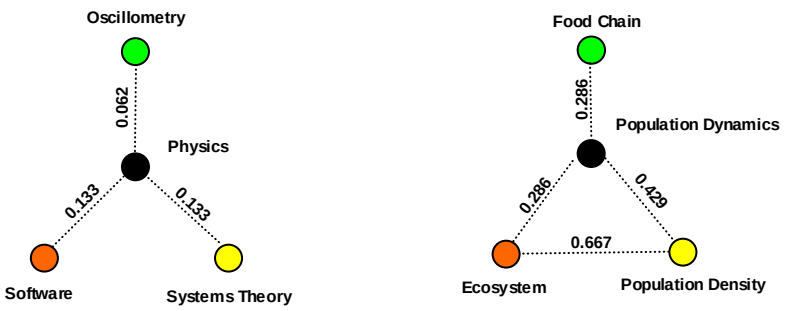


Ilustración 32: Conglomerados de la red en 2004-A. El mapa del conglomerado *"Physics"* contiene 4 temáticas. El mapa del conglomerado *"Population Dynamics"* contiene 4 temáticas.

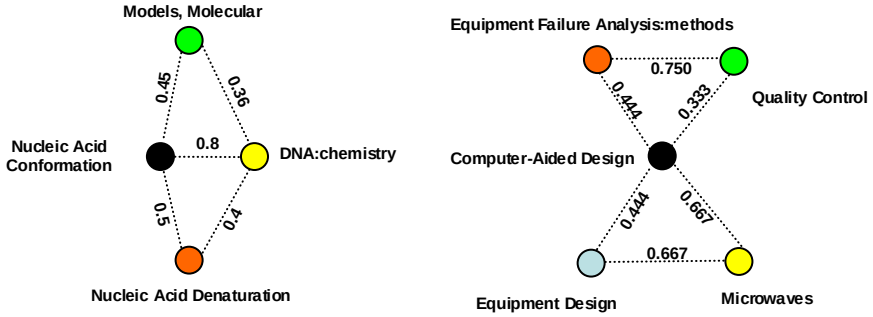


Ilustración 33: Conglomerados de la red en 2004-B. El mapa del conglomerado *"Nucleic Acid Conformation"* contiene 4 temáticas. El mapa del conglomerado *"Computer-Aided Design"* contiene 5 temáticas.

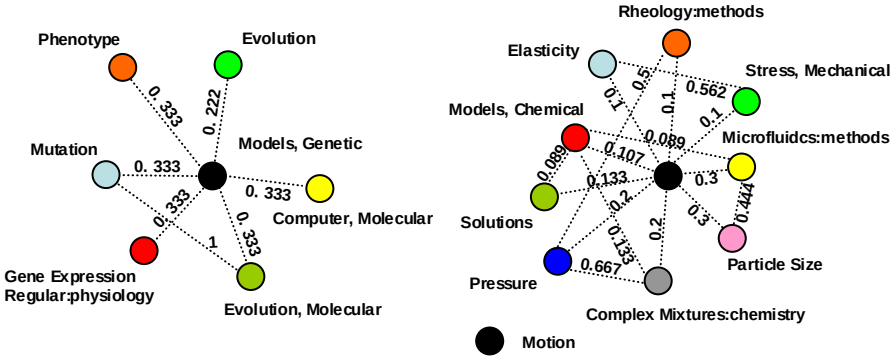


Ilustración 34: Conglomerados de la red en 2004-C. El mapa del conglomerado *"Models, Genetic"* contiene 7 temáticas. El mapa del conglomerado *"Motion"* contiene 10 temáticas.

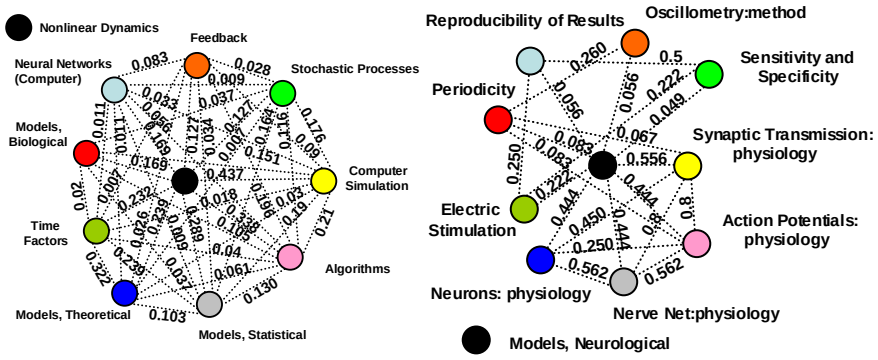


Ilustración 35: Conglomerados de la red en 2004-D. El mapa del conglomerado *"Nonlinear Dynamics"* contiene 10 temáticas. El mapa del conglomerado *"Models, Neurological"* contiene 10 temáticas.

Una vez identificados los conglomerados, definidos por sus temáticas y por las cadenas que los unen, veamos sus respectivos índices de densidad y centralidad, los cuales, hacen referencia a sus estructuras internas y a su relación con la globalidad de la red.

Conglomerados	Centralidad	Densidad
Models, Neurological	14.020	71.600
Computer-Aided Design	16.380	66.100
Nucleic Acid Conformation	4.000	62.750
Nonlinear Dynamics	124.090	48.760
Population Dynamics	0.000	41.700
Models, Genetic	1.900	41.243
Motion	18.640	40.240
Physics	1.370	8.200

Tabla 15: Centralidad y densidad de los conglomerados de C1.

Se tiene gran diversidad de conglomerados asociados al corpus C1 pero débilmente unidos unos con otros. El conglomerado principal "Nonlinear Dynamics" se coloca en una posición central dentro la estructura de la red. Está conectado sólidamente a otros conglomerados pero la densidad de sus relaciones internas es débil. Para obtener una descripción más detallada de la posición y del grado de desarrollo de los conglomerados que constituyen la red veamos el diagrama estratégico.

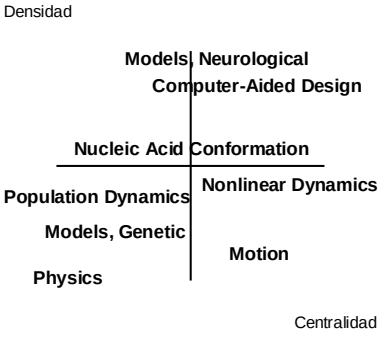


Ilustración 36: Diagrama estratégico del corpus C1. *"Computer-Aided Design"* es central y desarrollado (tema motor). *"Models, Neurological"* y *"Nucleic Acid Conformation"* son periféricos y desarrollados (temas especializados). *"Nonlinear Dynamics"* y *"Motion"* son centrales y no desarrollados (temas puentes). Mientras que *"Physics", "Models, Genetic"* y *"Population Dynamics"* son periféricos y no desarrollados (temas marginales).

En el diagrama estratégico están representadas todas las familias de temas: motores, especializados, puentes y marginales. Esta diversidad de temas con distintos grados de desarrollo describe contenidos complejos y ricos en la literatura analizada.

Red en el año 2005 asociada a C2. La red asociada al corpus C2 consiste de 7 conglomerados. "Nonlinear Dynamics" y "Image Interpretation, Computer-Assisted:methods" son conglomerados principales. El conglomerado "Numerical Analysis Computer-Assisted" mantiene encadenamiento externo con ambos conglomerados principales. "Hydrostatic Pressure" es conglomerado aislado.

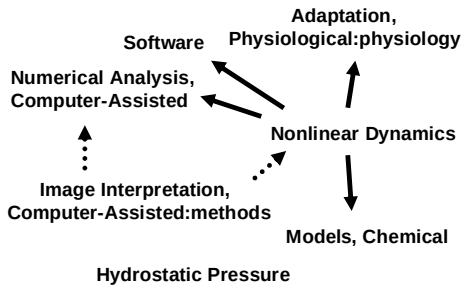


Ilustración 37: Red asociada al corpus C2. El conglomerado *"Nonlinear Dynamics"* es principal con conglomerados secundarios *"Software", "Models, Chemical"* y *"Numerical Analysis, Computer-Assisted".* El conglomerado *"Image Interpretation, Computer-Assisted:methods"* es principal con conglomerados secundarios *"Nonlinear Dynamics"* y *"Numerical Analysis, Computer-Assisted".* Mientras que los conglomerados *"Software", "Adaptation, Physiological:physiology" , "Numerical Analysis, Computer-Assisted", "Models, Chemical"* y *"Hydrostatic Pressure"* son aislados.

Para la construcción de los conglomerados se fijo el umbral de saturación en 10 y los valores de ocurrencia mínima y de co-ocurrencia mínima en 2 y 2 respectivamente. Con estos valores fijos se identificaron 230 temáticas, de los cuales 79 temáticas superan los valores de ocurrencia mínima y de co-ocurrencia mínima y 151 temáticas se descartaron.

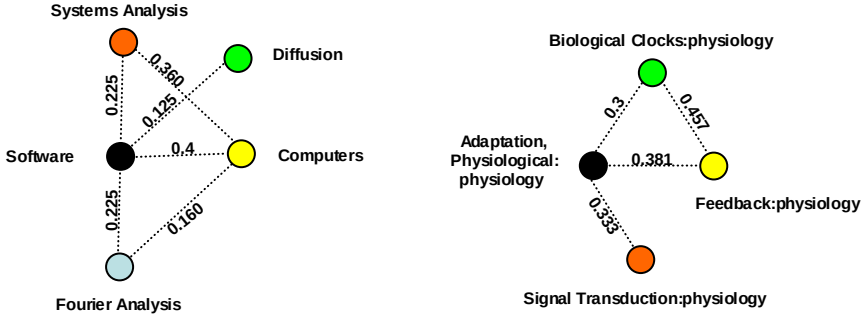


Ilustración 38: Conglomerados de la red en 2005-A. El mapa del conglomerado *"Software"* contiene 5 temáticas. El mapa del conglomerado *"Adaptation, Physiological:physiology"* contiene 4 temáticas.

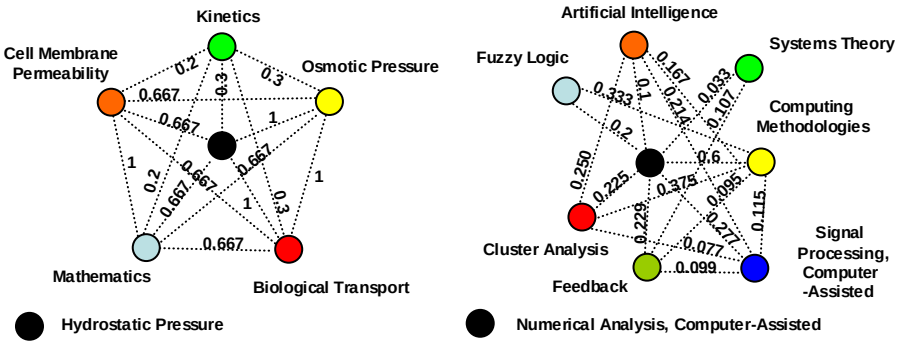


Ilustración 39: Conglomerados de la red en 2005-B. El mapa del conglomerado *"Hydrostatic Pressure"* contiene 6 temáticas. El mapa del conglomerado *"Numerical Analysis, Computer-Assisted"* contiene 8 temáticas.

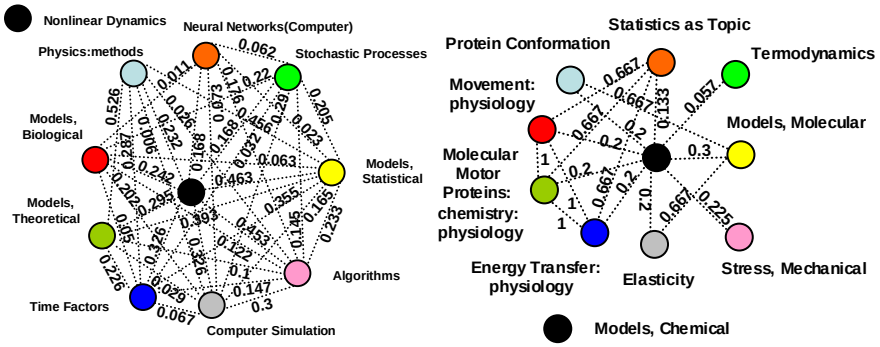


Ilustración 40: Conglomerados de la red en 2005-C. El mapa del conglomerado *"Nonlinear Dynamics"* contiene 10 temáticas. El mapa del conglomerado *"Models, Chemical"* contiene 10 temáticas.

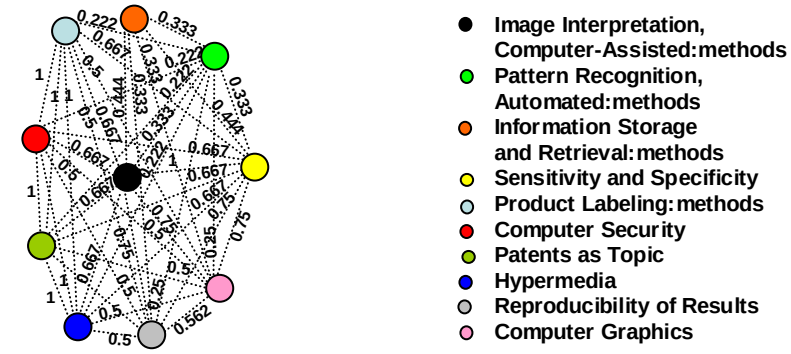


Ilustración 41: Conglomerados de la red en 2005-D. El mapa del conglomerado *"Image Interpretation, Computer-Assisted:methods"* contiene 10 temáticas.

Para obtener una referencia de la estructura interna de cada conglomerado y de su relación con la globalidad de la red veamos sus respectivos índices de centralidad y densidad. La alta centralidad del conglomerado principal "Nonlinear Dynamics" lo coloca en una posición central en la estructura de la red. El conglomerado "Image Interpretation, Computer-

Assisted:methods" posee una alta densidad que indica un encadenamiento interno muy fuerte entre sus temáticas. Exceptuando a los conglomerados "Nonlinear Dynamics" y "Image Interpretation, Computer-Assisted: methods", los otros conglomerados están débilmente unidos unos con otros.

Conglomerados	Centralidad	Densidad
Image Interpretation, Computer-Assisted:methods	64.370	238.390
Hydrostatic Pressure	0.000	155.033
Models, Chemical	5.550	80.500
Nonlinear Dynamics	147.580	76.630
Numerical Analysis, Computer-Assisted	12.040	43.700
Adaptation, Physiological:physiology	0.000	36.775
Software	4.670	29.900

Tabla 16: Centralidad y densidad de los conglomerados de C2.

En términos generales, la literatura analizada en 2005 ostenta contenidos complejos y ricos con diferentes grados de desarrollo.

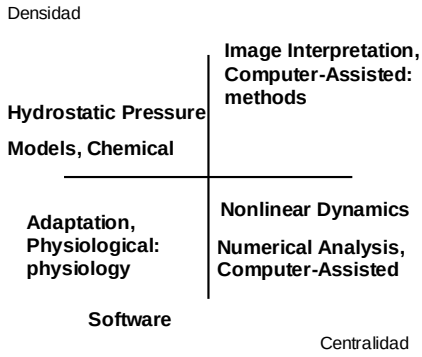


Ilustración 42: Diagrama estratégico del corpus C2. *"Image Interpretation, Computer-Assisted:methods"* es tema motor. Como temas puente están *"Nonlinear Dynamics"* y *"Numerical Analysis, Computer-Assisted"*. Como temas especializados están *"Models, Chemical"* y *"Hydrostatic Pressure"* y como temas marginales están "*Software"* y *"Adaptation, Physiological: physiology"*.

Red en el año 2006 asociada a C3. La red asociada al corpus C3 exhibe cuatro conglomerados y un único conglomerado principal. El conglomerado principal "Algorithms" esta encadenado externamente con los otros conglomerados. La red no exhibe ningún conglomerado aislado.

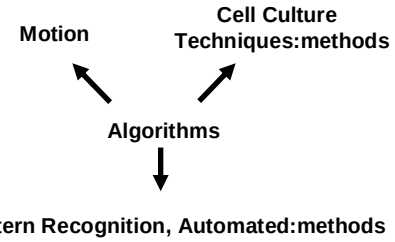


Ilustración 43: Red asociada al corpus C3. El conglomerado *"Nonlinear Dynamics"* es principal con conglomerados secundarios *"Motion", "Pattern Recognition, Automated:methods"* y *"Cell Culture Techniques:methods".* Mientras que los conglomerados *"Motion", "Pattern Recognition, Automated:methods"* y *"Cell Culture Techniques:methods"* son aislados.

Para la construcción de los conglomerados se fijo el umbral de saturación en 10 y los valores de ocurrencia mínima y de co-ocurrencia mínima en 2 y 2 respectivamente. Con estos valores fijos se identificaron 211 temáticas, de los cuales 62 temáticas superan los valores de ocurrencia mínima y de co-ocurrencia mínima y 149 temáticas se descartaron.

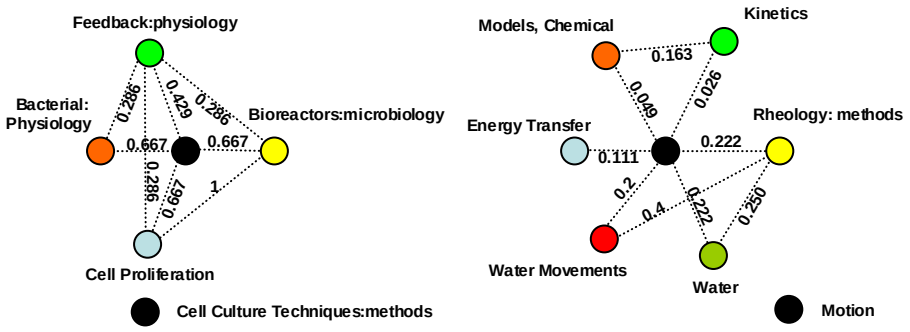


Ilustración 44: Conglomerados de la red en 2006-A. El mapa del conglomerado *"Cell Culture Techniques: methods"* contiene 5 temáticas. El mapa del conglomerado *"Motion"* contiene 7 temáticas.

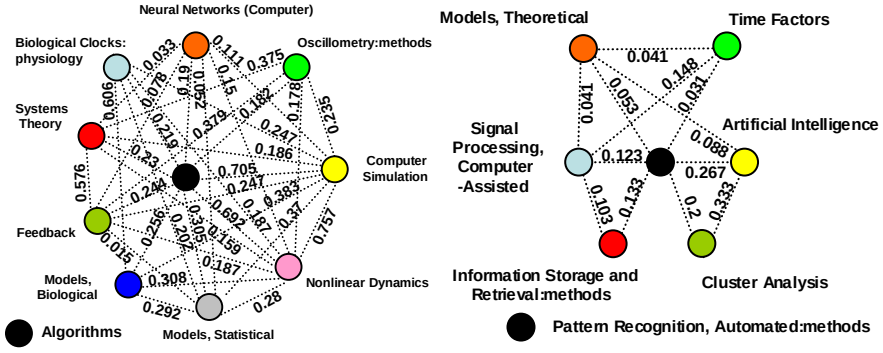


Ilustración 45: Conglomerados de la red en 2006-B. El mapa del conglomerado *"Algorithms"* contiene 10 temáticas. El mapa del conglomerado *"Pattern Recognition, Automated:methods"* contiene 7 temáticas.

En este periodo, el conglomerado principal "Algorithms" está conectado sólidamente a otros conglomerados y sus enlaces internos son muy fuertes. Es central en la estructura y desarrollo de la red.

Conglomerados	Centralidad	Densidad
Algorithms	128.210	96.160
Cell Culture Techniques:methods	0.000	85.760
Motion	13.960	23.471
Pattern Recognition, Automated:methods	1.800	22.300

Tabla 17: Centralidad y densidad de los conglomerados de C3

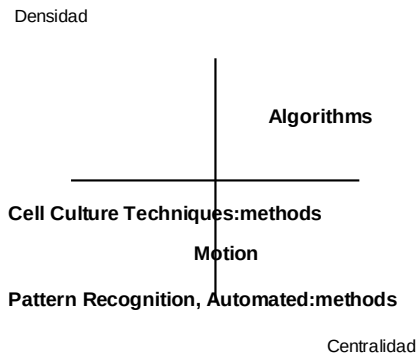


Ilustración 46: Diagrama estratégico del corpus C3. Como tema motor se tiene a *"Algorithms"* y como temas marginales están *"Motion", "Cell Culture Techniques:methods"* y *"Pattern Recognition, Automated: methods"*

El diagrama estratégico muestra la distribución de los 4 conglomerados en el cuadrante 1 y en el cuadrante 4. Esto significa que el contenido de la literatura analizada se organiza en torno a un tema bien estructurado y bien desarrollado. Los contenidos "Cell Culture Techniques: methods", "Motion" y "Pattern Recognition, Automated:methods" (perifericos y no desarrollados) se organizan en torno a "Algorithms" (central y desarrollado).

Red en el año 2007 asociada a C4. La red asociada al corpus C4 exhibe 6 conglomerados. Los conglomerado principales son "Time Factors" y "Models, Theoretical". Se tiene de nuevo un conglomerado aislado: "Scattering, Radiation".

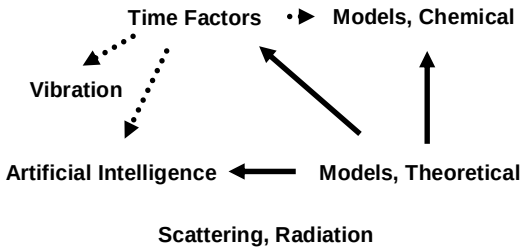


Ilustración 47: Red asociada al corpus C4. El conglomerado *"Models, Theoretical"* es principal con conglomerados secundarios *"Artificial Intelligence", "Models, Chemical"* y *"Time Factor"*. El conglomerado *"Time Factor"* es principal con conglomerados secundarios *"Artificial Intelligence"* y *"Models, Chemical"*. Mientras que los conglomerados *"Vibration", "Scattering, Radiation", "Artificial Intelligence"* y *"Models, Chemical"* son aislados.

Para la construcción de los conglomerados se fijo el umbral de saturación en 10 y los valores de ocurrencia mínima y de co-ocurrencia mínima en 2 y 2 respectivamente. Con estos valores fijos se identificaron 261 temáticas, de los cuales 77 temáticas superan los valores de ocurrencia mínima y de co-ocurrencia mínima y 184 temáticas se descartaron.

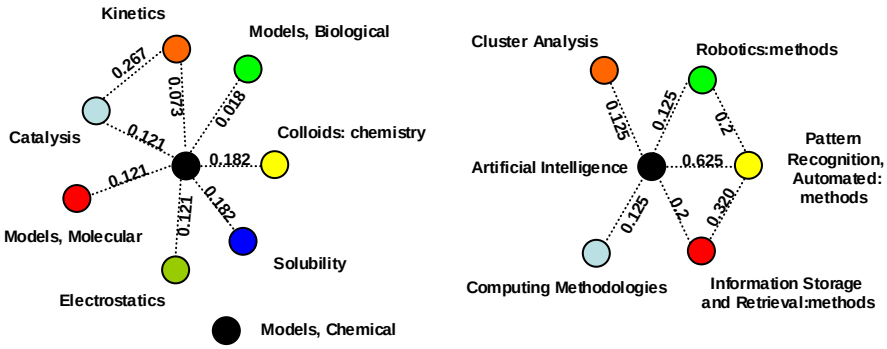


Ilustración 48: Conglomerados de la red en 2007-A. El mapa del conglomerado *"Models, Chemical"* contiene 8 temáticas. El mapa del conglomerado *"Artificial Intelligence"* contiene 6 temáticas.

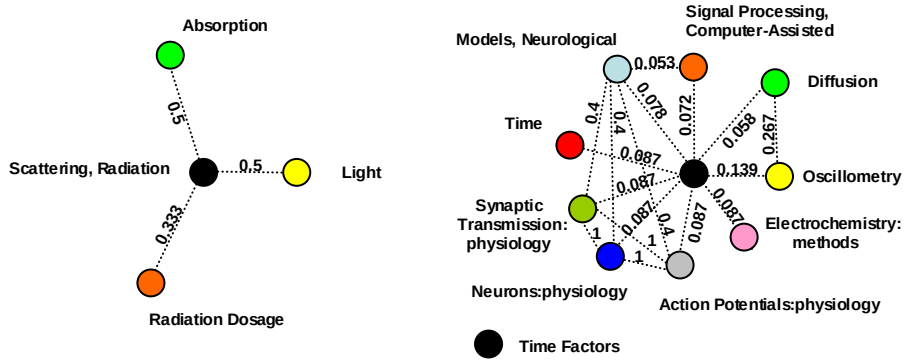


Ilustración 49: Conglomerados de la red en 2007-B. El mapa del conglomerado *"Scattering, Radiation"* contiene 4 temáticas. El mapa del conglomerado *"Time Factors"* contiene 10 temáticas.

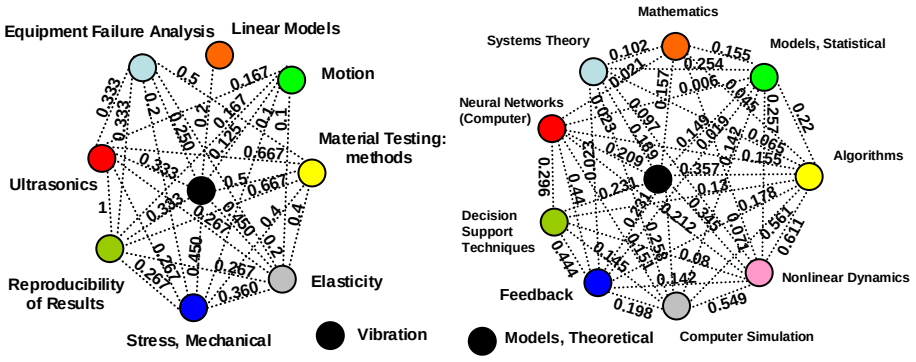


Ilustración 50: Conglomerados de la red en 2007-C. El mapa del conglomerado *"Vibration"* contiene 9 temáticas. El mapa del conglomerado *"Models, Theorical"* contiene 10 temáticas.

El conglomerado principal "Models, Theoretical" está conectado sólidamente a otros conglomerados aunque no posee una fuerte estructura interna. Exceptuando al conglomerado "Time Factors" los otros conglomerados están débilmente unidos unos con otros.

Conglomerados	Centralidad	Densidad
Vibration	5.000	103.367
Models, Theorical	114.940	79.180
Time Factors	42.160	53.020
Scattering, Radiation	0.000	33.325
Artificial Intelligence	0.000	28.667
Models, Chemical	9.190	13.563

Tabla 18: Centralidad y densidad de los conglomerados de C4

La literatura analizada de forma semejante que en los periodos 2004 y 2005 ostenta contenidos complejos y ricos con diferentes grados de desarrollo.

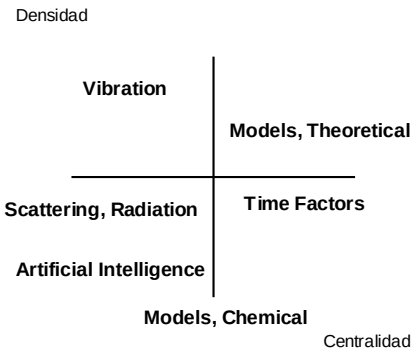


Ilustración 51: Diagrama estratégico del corpus C4. Como tema motor esta *"Models, Theoretical"*. Como tema puente esta *"Time Factors"*. Como tema especializado esta *"Vibration"* y como temas marginales están *"Scattering, Radiation", "Artificial Intelligence"* y *"Models, Chemical"*.

La estructura de la red se caracteriza por contener dos conglomerados principales durante los años 2004, 2005 y 2007. Solamente, contiene un conglomerado principal en el año 2006. Un conjunto de conglomerados secundarios acompañan a los conglomerados principales durante el transcurso del tiempo. Los conglomerados principales están encadenados fuertemente a los conglomerados secundarios pero los conglomerados secundarios están débilmente unidos unos con otros. Uno de los conglomerados principales forma una serie más o menos estable durante el transcurrir del tiempo. Esta serie principal inicia con el conglomerado "Nonlinear Dynamics" en 2004 y 2005, el conglomerado se redefine y se reorganiza en "Algorithms" en 2006 y posteriormente en "Models, Theoretical" en 2007. Durante la reorganización y redefinición de esta serie principal se detecto que las siguientes temáticas se mantienen estables: Algorithms", "Computer Simulation", "Models, Statistical", "Neural Networks (Computer)" y "Nonlinear Dynamics". A los conjuntos de temáticas que se mantienen estables en el tiempo se les denomina serie temática. La literatura analizada revela contenidos temáticos que son de interés constante por parte de los investigadores. Mientras que el otro conglomerado principal no forma ninguna serie estable más bien se obtiene un conjunto de conglomerados principales. Cada conglomerado principal es tan distinto al anterior que ninguna temática sobrevive a la reorganización y redefinición. El conjunto de conglomerados principales esta dado por "Motion" en 2004, "Image Interpretation, Computer - Assisted:methods" en 2005 y "Time Factors" en 2007.

En cuanto, al conjunto de conglomerados secundarios algunos desaparecen y otros progresivamente reorganizan sus relaciones en el transcurso del tiempo. Algunos revelan contenidos computacionales, neurológicos, químicos, genéticos, moleculares, etc., Otros

revelan contenidos poco habituales como: Dinámica de Poblaciones, Presión Hidrostática, Técnicas in Vitro de Crecimiento de Células y Desviación de Energías.

Otro conjunto de temáticas que no llegan a formar una serie temática debido a que emergen en diferentes conglomerados (principales y secundarios) pero manifiestan un sentido de modelación en los contenidos de la literatura analizada son "Models, Theoretical", "Models, Chemical", "Models, Biological", "Time Factors", "Feedback" y "Systems Theory".

En resumen, la serie principal constituye el centro de interés asociado al MeSH Major Topic "Nonlinear Dynamics". Simular aspectos dinámicos de ciertos procesos fisiológicos, ya sean químicos, biológicos, genéticos, neurológicos, moleculares, genéticos, etc., requiere uso de ecuaciones matemáticas, computadoras o algún otro equipo electrónico.

5.5.-Análisis Comparativo de la Red

Una vez que se han identificado, relaciones e importancia de los conglomerados en la estructura de la red. El paso siguiente es analizar la evolución de la red en el transcurso del tiempo. El análisis comprende: la comparación de conglomerados, su distribución sobre los diagramas estratégicos y el ciclo de vida de los conglomerados.

(a) Comparación de Conglomerados

Los conglomerados durante el transcurso del tiempo pueden conservarse, desvanecerse, partirse o fusionarse, etc. Para detectar conglomerados similares a otros conglomerados durante el transcurso del tiempo se recurre al índice de transformación.

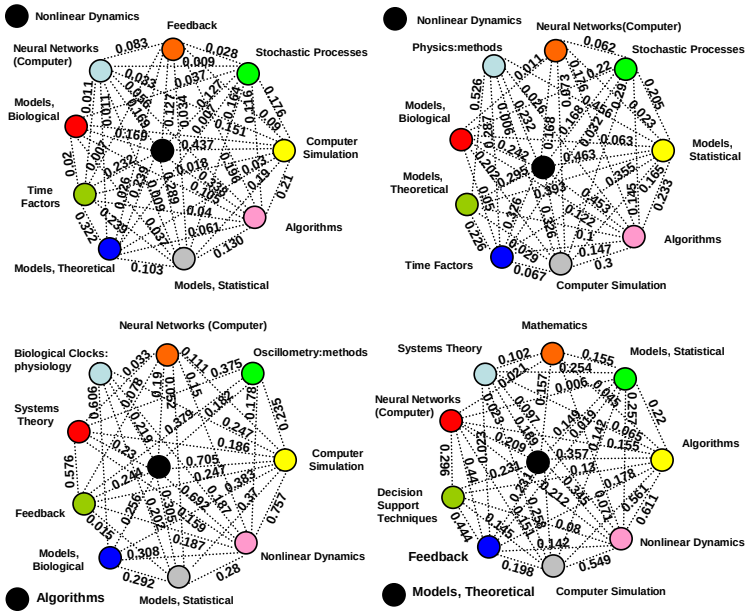


Ilustración 52: Conglomerados que integran la serie principal. "Nonlinear Dynamics" de C1, "Nonlinear Dynamics" de C2, "Algorithms" de C3 y "Models, Theoretical" de C4.

La gran mayoría de los conglomerados que integran las redes no comparte las suficientes temáticas para ser considerados similares a otros. Solamente, los conglomerados que integran la serie principal (Ilustración 52) comparten durante el transcurso del tiempo varias temáticas entre ellos, como para ser considerados similares. Los conglomerados "Nonlinear Dynamics" de C1 y "Nonlinear Dynamics" de C2 poseen el mejor índice de transformación de 20/9.

Mientras que los conglomerados “Nonlinear Dynamics” de C2 y “Algorithms” de C3 poseen el peor índice de transformación de 20/5. En el Apéndice se muestran las correspondientes tablas de índices de transformación para la serie principal.

En la tabla 19, se observa la evolución de las temáticas que integran los conglomerados de la serie principal. La “x” indica que la temática está presente en dicho año y en caso contrario la temática está ausente, tal vez se integro a otro conglomerado o desapareció y reapareció en el año siguiente.

Temáticas	Años			
	2004	2005	2006	2007
Algorithms	x	x	x	x
Biological Clocks:physiology			x	
Computer Simulation	x	x	x	x
Decision Support Techniques				x
Feedback	x		x	x
Mathematics				x
Models, Biological	x	x	x	
Models, Statistical	x	x	x	x
Models, Theoretical	x	x		x
Neural Networks(Computer)	x	x	x	x
Nonlinear Dynamcs	x	x	x	x
Oscillometry:methods			x	
Physics:methods		x		
Stochastic Processes	x	x		
Systems Theory			x	x
Time Factors	x	x		

Tabla 19: Evolución de las temáticas de la serie principal.

La serie principal conserva una colección de temáticas que constantemente están encadenadas con “Algorithms”, “Computer Simulation”, “Models, Statistical”, “Neural Networks (Computer)” y “Nonlinear Dynamics”, constituyendo, una especie de temáticas complementarias. Sin embargo, hay temáticas como “Mathematics”, “Biological Clocks:physiology”, “Decision Support Techniques”, “Oscillometry:methods” y “Physics:methods” que se ausentan rápidamente de la serie temática.

(b) Comparación de las posiciones de los conglomerados sobre los diagramas estratégicos. Una vez establecidas las (eventuales) similitudes entre los diferentes conglomerados de la red considerada se examinan sus posiciones sobre los diagramas estratégicos correspondientes. Esta clasificación de los conglomerados visualizada bajo la forma de un diagrama estratégico cuya lectura resulta fácil, proporciona una descripción más detallada del estado de la red, de la posición y del grado de desarrollo de los temas que la constituyen.

En términos generales, durante el transcurso del tiempo la organización de los conglomerados en los diagramas estratégicos revela una gran diversidad de contenidos en la literatura analizada. Hay contenidos centrales, contenidos categorizantes, contenidos especializados y contenidos marginales. El grado de desarrollo de estos contenidos reflejan los intereses cambiantes de los investigadores.

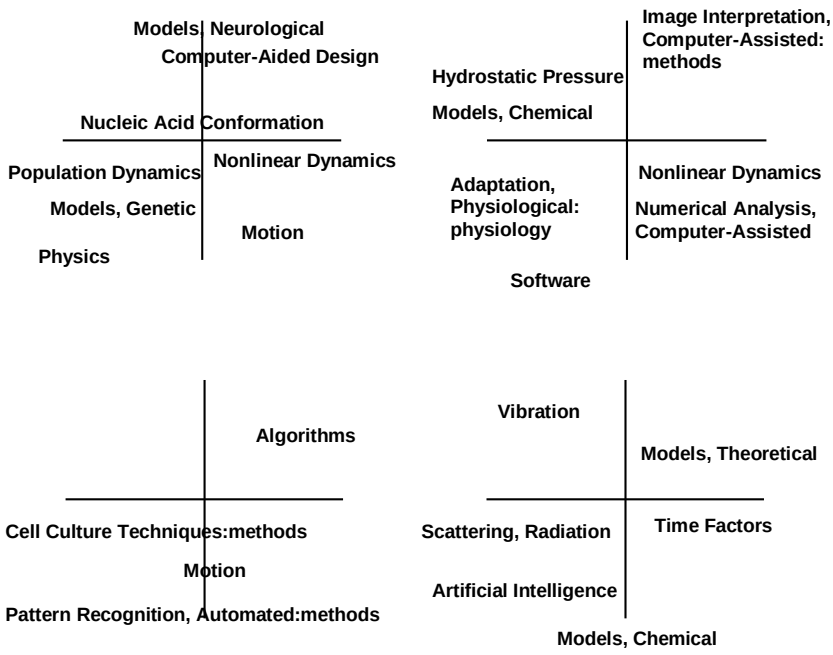


Ilustración 53: Diagramas estratégicos de los corpus C1, C2, C3 y C4.

Se aprecia que la serie principal integrada por los conglomerados "Nonlinear Dynamics" de C1, "Nonlinear Dynamics" de C2, "Algorithms" de C3 y "Models, Theoretical" de C4 se mantiene durante el transcurso del tiempo. Pasan de ser temática puente a temática motor. Lo cual indica que la serie puede ser considerada como *centro de interes* vigente por parte de los investigadores. Otra serie de conglomerados que se mantiene durante el transcurrir del tiempo está dada por "Motion" y "Models, Chemical". Esta serie nunca ocupan un lugar importante en los diagramas estratégicos.

Percibamos el grado de integración y desarrollo de la red. Para ello, se calculo la centralidad media y densidad media de todos los conglomerados que integran la red durante el transcurso del tiempo. Vea tabla 20. La red estuvo aumentando su integración en los años 2004, 2005 y 2006 pero la redujo en 2007. El grado de coherencia de las temáticas que integran cada uno de los conglomerados alcanzo un máximo en 2005.

Año	Centralidad Media	Densidad Media
2004	22.55	47.57
2005	33.45	94.41
2006	35.99	56.92
2007	28.54	51.85

Tabla 20: Variación de los índices de centralidad media y densidad media durante el transcurso del tiempo.

Se tiene una red que estuvo al mismo tiempo aumentando la fuerza del encadenamiento externo entre los conglomerados (principales y secundarios) y diversificando sus temáticas. Hay que ser prudentes pues solamente se está considerando una pequeña parte de la red en un breve periodo de tiempo.

(c) El ciclo de vida de la serie principal. Se analiza la evolución de la serie principal integrada por "Nonlinear Dynamics" de C1, "Nonlinear Dynamics" de C2, "Algorithms" de C3 y "Models, Theoretical" de C4 mediante los índices de densidad, de centralidad y de transformación.

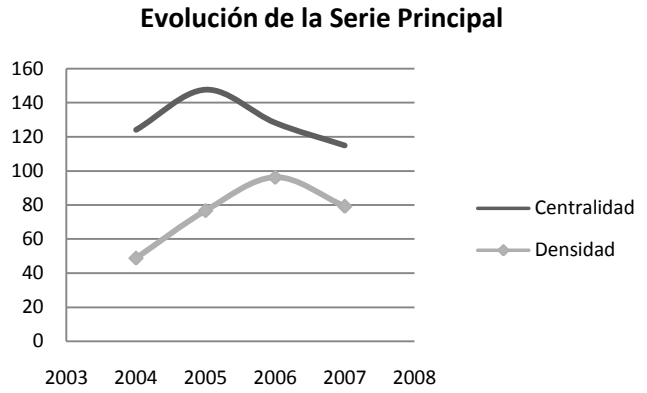


Ilustración 54: Visualización de la evolución de la serie principal.

La fuerza del encadenamiento externo de la serie principal es alta (centralidad) pero la fuerza del encadenamiento interno es baja (densidad). Esto indica que la serie principal se conserva en el centro de la red con una estructura interna débil pero aun así es importante en el desarrollo de la red.

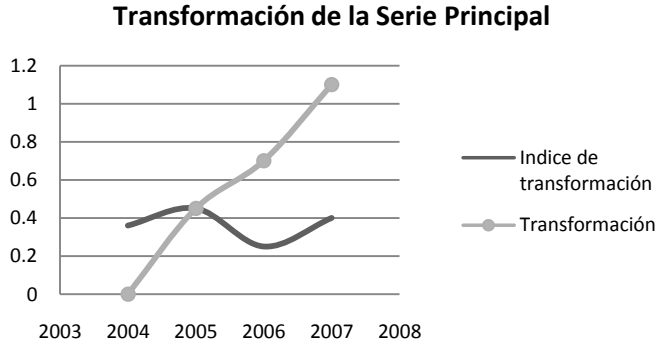


Ilustración 55: Transformación de la serie principal.

Los índices de transformación de los conglomerados que integran la serie principal no presentan cambios significativos en el tiempo. La serie principal se reorganiza y se redefine suavemente durante el transcurso del tiempo.

La cantidad de documentos indizados en el año 2004 con el MeSH Major Topic "Nonlinear Dynamics" es atípica al resto de los años. En términos generales, la indización tiende a crecer en cada año, (Vea tabla 14). En cuanto, a la serie principal, el volumen de documentos indizados con al menos tres de las temáticas siguientes "Algorithms", "Computer Simulation", "Models, Statistical", "Neural Networks (Computer)" y "Nonlinear Dynamics" se

mantiene sin variaciones significativas en cada año. Además, este volumen no se ve afectado por el tamaño atípico de la indización en 2004. (Ilustración 56).

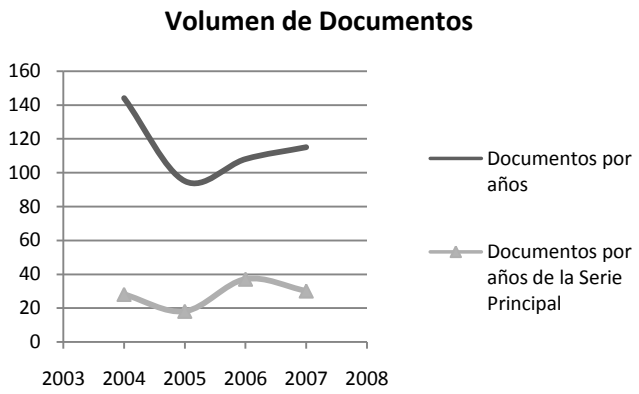


Ilustración 56: Volumen de documentos de la serie principal.

En resumen, el centro de interés "Nonlinear Dynamics" es central en la estructura de la red durante los años 2004, 2005, 2006 y 2007. Su reorganización y redefinición es suave durante el transcurso del tiempo. Además, mantiene un núcleo de publicaciones asociado a unas cuantas temáticas que conforman la serie temática.

Veamos que países, instituciones e investigadores contribuyen a la formación de este núcleo de publicaciones.

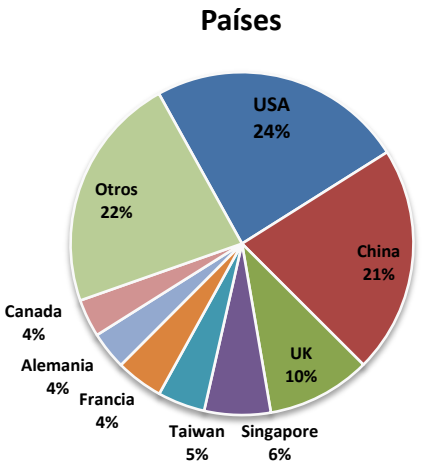


Ilustración 57: Producción de documentos por país. (Total de documentos 113).

Estados Unidos de America, China y Reino Unido son los países con mayor número de documentos en el núcleo. En Estados Unidos de América, la producción de documentos la hacen las universidades, laboratorios y centros como Universidad de Arizona, Universidad de San Diego, Universidad de California, Universidad de Michigan, Sandia National Laboratories, US Naval Research Laboratory, National Centers for Environmental Prediction, entre otros. En

China, la producción de documentos principalmente la hacen las universidades como Universidad de Zhejiang, Universidad de Xinjiang, Universidad de Zhong Shan, Universidad de Beihang, entre otras. En Reino Unido, al igual que China, la producción de documentos la hacen las universidades como Universidad de Loughborough, Universidad del Colegio Londres, Universidad de Bristol, entre otras. En Singapore, la producción de documentos la hace principalmente la Universidad Nacional de Singapore.

Solamente, cuatro investigadores han publicado tres documentos en los últimos cuatro años. Los investigadores Ge, S. S. y Zhang, J ambos de la Universidad Nacional de Singapore. Lai, Y. C. de la Universidad de Arizona y Lin, C. M. de la Universidad de Tung en Taiwan. Los investigadores Ge, S. S. y Zhang, J han colaborado en una publicación y Zhang, J tiene una publicación afiliada a la Universidad de Wisconsin en Estados Unidos. Hay 17 investigadores que han publicado dos documentos y el resto solamente han publicado un documento.

Las revistas favoritas para publicar resultados son Chaos, IEEE transactions on neural networks y Physical review E, Statistical, nonlinear, and soft matter physics. La revista Chaos es editada por American Institute of Physics; está clasificada en ciencias básicas y experimentales: matemáticas; tiene un factor de impacto de 1.76. La revista IEEE transactions on neural networks (abreviado como IEEE Trans Neural Netw) es editada por IEEE Neural Networks Council; abarca la teoría, diseño, aplicaciones y desarrollos biológicos y lingüísticos que motivan paradigmas computacionales que incluyen redes neuronales de todo tipo; tiene un factor de impacto de 2.205. La revista Physical review E, Statistical, nonlinear, and soft matter physics (abreviado como Phys Rev E Stat Nonlin Soft Matter Phys) es editada por American Institute of Physics; abarca desarrollos quánticos y caos clásico, física de materia, física estadística, fluidos, polímetros, física de plasma, física computacional, física biológica y materiales granulados; tiene un factor de impacto de 2.418, (Vea Apéndice C).

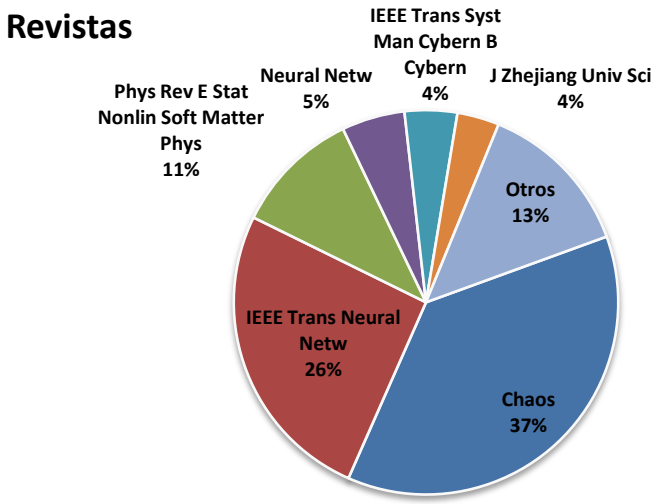


Ilustración 58: Principales revistas del núcleo.

En México, el investigador Sánchez, E. N. del CINVESTAV, Unidad Guadalajara ha publicado dos documentos. En colaboración con Alanis, A. Y., y Loukianov, A. G., publico "Discrete-time adaptive backstepping nonlinear control via high-order neural networks" y con Felix, R. A., y Chen, G. publico "Reproducing chaos by variable structure recurrent neural networks". Los investigadores Jorge, M. C., Cruz-Pacheco, G. Mier-y-Teran-Romero, L. y Smyth, N. F., del Departamento de Matemáticas y Mecánica, I.I.M.A.S., UNAM, han publicado

un solo documento, titulado "Evolution of two-dimensional lump nanosolitons for the Zakharov-Kuznetsov and electromigration equations".

En resumen, el centro de interés "Nonlinear Dynamics" no esta bajo tutela de ningún grupo de investigadores que se haga cargo del tema, de forma sistemática y durable.

5.6.-Técnicas Estadísticas y Mapa Auto-Organizante

Con el fin de comparar los mapas de conglomerados obtenidos por medio del algoritmo de agrupación sobre centros simples, AACS, se aplico a las matrices de co-ocurrencias normalizadas técnicas estadísticas como el método jerárquico de conglomerado "Vecino más Cercano" y el algoritmo K-means para agrupar temáticas similares. Además, se aplico el Análisis Discriminante para clasificar las temáticas de cada uno de los Corpus. Estas técnicas estadísticas están implementadas el software SPSS 15 for Windows.

Para simplificar la presentación de los resultados, los dendrogramas obtenidos con el método jerárquico de conglomerado "Vecino más Cercano" están en el Apéndice D y para su construcción se considero la distancia euclidiana. Para el Algoritmo K-means se uso un "Quick Cluster Algorithms" debido a que se conocen el número de conglomerados deseados. En los resultados no se muestra la "distancia desde el centro del conglomerado" solamente "conglomerado de pertenencia". El Análisis Discriminante uso la función discriminante lineal de Fisher. La clasificación se realizo con "Cross Validation", es decir SPSS cicla las temáticas del conjunto de datos. En cada ciclo deja una temática afuera y es tratada como un dato de prueba. El resto de las temáticas es tratado como un nuevo conjunto de datos. Para simplificar los resultados se descartan las matrices W (inter grupos), B (entre grupos) y T con T = B + W; Scores Discriminantes. No se realizo ninguna prueba como la prueba de significancia multivariada para examinar si hay una diferencia significativa entre los centroides (medias) o la prueba para igualdad de matrices de varianzas-covarianzas. Ya que los mapas de conglomerados están integrados por un número distinto de temáticas (a lo máximo 10 temáticas) se considero que las probabilidades a priori utilizadas por la función discriminante lineal de Fisher para la clasificación se calcularan a partir del tamaño de los conglomerados obtenidos por el algoritmo de agrupación sobre centros simples. Vea tabla 21.

	Probabilidades a priori			
Conglomerados	**C1**	**C2**	**C3**	**C4**
0	0.074	0.094	0.172	0.170
1	0.074	0.075	0.241	0.128
2	0.074	0.113	0.241	0.085
3	0.093	0.151	0.345	0.213
4	0.130	0.189		0.191
5	0.185	0.189		0.213
6	0.185	0.189		
7	0.185			
Total	**1**	**1**	**1**	**1**

Tabla 21: Probabilidades a priori

En las tablas 22, 23, 24 y 25 se muestran los resultados. Los conglomerados están indicados por 0, 1, 2, 3, 4, 5, 6 y 7. Cada corpus tiene un número distinto de conglomerados, por ejemplo, el corpus C1 tiene 8 conglomerados. En la primera fila de cada tabla se muestran los siguientes encabezados, *Temáticas* hace referencia a descriptores MeSH Major Topic. *CP-AACS* se refiere al conglomerado de pertenencia indicado por el algoritmo de agrupación sobre centros simples. *CP-CrossV* se refiere al conglomerado de pertenencia indicado por el algortimo de clasificación "Cross Validation" usado por SPSS. *CP-KM* se refiere al conglomerado de pertenencia indicado por el algoritmo K-means. CP-*VC* se refiere al conglomerado de pertenencia indicado por el método jerárquico de conglomerado "Vecino más Cercano".

Para el Corpus C1, “Vecino más Cercano” y el algoritmo K-means ambos forman un conglomerado que agrupa muchas temáticas. Estos conglomerados son grandes con respecto a los otros que forman e incluso a los conglomerados hechos con AACS. Estos conglomerados están indicados por 0 y 6 respectivamente, Vea tabla 22. El análisis discriminante clasifico las 54 temáticas del corpus C1 y dio un 38.9% de confianza a las temáticas agrupadas con el AASC.

Temáticas	*CP-AACS*	*CP-CrossV*	*CP-KM*	*CP-VC*
Action Potentials:physiology	6	3	7	4
Algorithms	7	7	6	0
Complex Mixtures:chemistry	5	7	5	7
Computer Simulation	7	0	6	0
Computer-Aided Design	3	6	1	6
Computers, Molecular	4	4	6	5
DNA:chemistry	2	1	3	2
Ecosystem	1	7	2	0
Elasticity	5	2	3	2
Electric Stimulation	6	6	6	3
Equipment Design	3	2	1	6
Equipment Failure Analysis:methods	3	3	4	6
Evolution	4	4	6	0
Evolution, Molecular	4	4	0	5
Feedback	7	7	6	0
Food Chain	1	1	6	0
Gene Expression Regulation:physiology	4	4	6	5
Microfluidics:methods	5	7	6	0
Microwaves	3	0	1	6
Models, Biological	7	7	2	0
Models, Chemical	5	6	5	2
Models, Genetic	4	1	0	5
Models, Molecular	2	0	3	2
Models, Neurological	6	3	7	4
Models, Statistical	7	6	6	0
Models, Theoretical	7	7	6	0
Motion	5	6	6	0
Mutation	4	4	0	5
Nerve Net:physiology	6	7	7	4
Neural Networks(Computer)	7	1	6	1
Neurons:physiology	6	6	7	4
Nonlinear Dynamics	7	7	6	0
Nucleic Acid Conformation	2	2	3	2
Nucleic Acid Denaturation	2	4	3	2
Oscillometry	0	0	6	0
Oscillometry:methods	6	5	6	0
Particle Size	5	2	6	0
Periodicity	6	6	6	0
Phenotype	4	1	6	5
Physics	0	6	6	0
Population Density	1	1	2	0
Population Dynamics	1	0	2	0
Pressure	5	7	6	0
Quality Control	3	7	4	6
Reproducibility of Results	6	6	6	3
Rheology:methods	5	7	6	0
Sensitivity and Specificity	6	6	6	3
Software	0	0	6	0
Solutions	5	7	5	7
Stochastic Processes	7	7	6	0
Stress, Mechanical	5	2	6	2
Synaptic Transmission:physiology	6	3	7	4
Systems Theory	0	0	6	0
Time Factors	7	7	6	0

Tabla 22: Comparación entre conglomerados del corpus C1

Para el Corpus C2, "Vecino más Cercano" formo un conglomerado grande de aproximadamente 27 temáticas. Este conglomerado está indicado por 0. El algoritmo K-means hizo conglomerados de entre 3 y 14 temáticas. Vea tabla 23. El análisis de discriminante clasifico las 53 temáticas del corpus C2 y dio un 49.1% de confianza a las temáticas agrupadas con el AACS.

Temáticas	*CP-AACS*	*CP-CrossV*	*CP-KM*	*CP-VC*
Adaptation, Physiological:physiology	1	1	6	0
Algorithms	5	6	6	0
Artificial Intelligence	3	3	4	0
Biological Clocks:physiology	1	5	6	0
Biological Transport	2	2	1	5
Cell Membrane Permeability	2	2	1	5
Cluster Analysis	3	3	6	0
Computer Graphics	6	3	2	3
Computer Security	6	1	2	3
Computer Simulation	5	4	6	0
Computers	0	5	5	0
Computing Methodologies	3	3	6	0
Diffusion	0	0	5	4
Elasticity	4	0	3	2
Energy Transfer:physiology	4	4	0	6
Feedback	3	5	6	0
Feedback:physiology	1	5	6	0
Fourier Analysis	0	3	5	0
Fuzzy Logic	3	4	6	0
Hydrostatic Pressure	2	2	1	5
Hypermedia	6	6	2	3
Image Interpretation, Computer-Assisted:methods	6	6	2	3
Information Storage and Retrieval:methods	6	6	4	0
Kinetics	2	3	1	1
Mathematics	2	2	1	5
Models, Biological	5	6	6	0
Models, Chemical	4	5	3	2
Models, Molecular	4	4	3	2
Models, Statistical	5	5	5	0
Models, Theoretical	5	6	5	0
Molecular Motor Proteins:chemistry:physiology	4	4	0	6
Movement:physiology	4	4	0	6
Neural Networks(Computer)	5	6	6	0
Nonlinear Dynamcs	5	5	5	0
Numerical Analysis, Computer-Assisted	3	5	6	0
Osmotic Pressure	2	2	1	5
Patents asTopic	6	1	2	3
Pattern Recognition, Automated:methods	6	3	4	0
Physics:methods	5	6	5	0
Product Labeling:methods	6	6	2	3
Protein Conformation	4	3	3	2
Reproducibility of Results	6	5	2	3
Sensitivity and Specificity	6	6	2	3
Signal Processing, Computer-Assisted	3	3	4	0
Signal Transduction:physiology	1	1	6	0
Software	0	6	5	0
Statistics as Topic	4	5	0	6
Stochastic Processes	5	5	6	0
Stress, Mechanical	4	2	6	2
Systems Analysis	0	0	5	0
Systems Theory	3	3	5	0
Themodynamics	4	0	6	1
Time Factors	5	0	5	0

Tabla 23: Comparación entre conglomerados del corpus C2

Para el Corpus C3, “Vecino más Cercano” y el algoritmo K-means formaron ambos un conglomerado grande de aproximadamente 15 temáticas cada uno. Estos conglomerados están indicados por 0 y 1 respectivamente. Vea tabla 24. El análisis discriminante clasifico las 29 temáticas del corpus C3 y dio un 34.5% de confianza a las temáticas agrupadas con el AACS.

Temáticas	*CP-AACS*	*CP-CrossV*	*CP-KM*	*CP-VC*
Algorithms	3	2	3	0
Artificial Intelligence	2	2	1	0
Bacterial Physiology	0	3	0	3
Biological Clocks:physiology	3	3	3	0
Bioreactors:microbiology	0	1	1	3
Cell Culture Techniques:methods	0	2	0	3
Cell Proliferation	0	1	1	3
Cluster Analysis	2	1	1	0
Computer Simulation	3	1	3	0
Energy Transfer	1	1	1	2
Feedback	3	0	2	1
Feedback:physiology	0	3	0	3
Information Storage and Retrieval:methods	2	0	1	0
Kinetics	1	3	2	1
Models, Chemical	1	1	1	1
Models, Statistical	3	0	3	0
Models, Theoretical	2	1	1	0
Models,Biological	3	1	3	0
Motion	1	1	1	2
Neural Networks(Computer)	3	3	1	0
Nonlinear Dynamics	3	3	3	0
Oscillometry:methods	3	3	2	1
Pattern Recognition, Automated:methods	2	1	1	0
Rheology:methods	1	1	1	2
Signal Processing, Computer-Assisted	2	3	1	0
Systems Theory	3	1	2	1
Time Factors	2	3	1	0
Water	1	1	1	2
Water Movements	1	0	1	2

Tabla 24: Comparación entre conglomerados del corpus C3

Para el Corpus C4, “Vecino más Cercano” y el algoritmo K-means formaron ambos un conglomerado grande de aproximadamente 35 temáticas y de 28 temáticas cada uno. Estos conglomerados están indicados por 0 y 1 respectivamente. Vea tabla 25. El análisis discriminante clasifico las 58 temáticas del corpus C4 y dio un 48.9% de confianza a las temáticas agrupadas con el AACS.

Temáticas	*CP-AACS*	*CP-CrossV*	*CP-KM*	*CP-VC*
Absorption	2	2	0	3
Action Potentials:physiology	3	3	2	2
Algorithms	5	5	1	0
Artificial Intelligence	1	3	1	0
Catalysis	0	0	1	0
Cluster Analysis	1	4	1	4
Colloids:chemistry	0	5	1	0
Computer Simulation	5	4	1	0
Computing Methodologies	1	4	1	4
Decision Support Techniques	5	5	1	0
Diffusion	3	3	1	0
Elasticity	4	0	5	0
Electrochemistry:methods	3	1	1	0
Electrostatics	0	0	1	0
Equipment Failure Analysis	4	4	5	0
Feedback	5	1	1	0
Information Storage and Retrieval:methods	1	5	1	0
Kinetics	0	0	1	0
Light	2	2	0	3
Linear Models	4	4	1	1
Materials Testing:methods	4	4	5	0
Mathematics	5	0	4	0
Models, Biological	0	0	1	0
Models, Chemical	0	2	1	0
Models, Molecular	0	0	1	0
Models, Neurological	3	5	2	2
Models, Statistical	5	1	1	0
Models, Theoretical	5	5	1	0
Motion	4	1	3	0
Neural Networks(Computer)	5	1	1	0
Neurons:physiology	3	3	2	2
Nonlinear Dynamics	5	5	1	0
Oscillometry	3	5	1	0
Pattern Recognition, Automated:methods	1	4	1	0
Radiation Dosage	2	2	0	3
Reproducibility of Results	4	4	3	0
Robotics:methods	1	2	1	0
Scattering, Radiation	2	0	0	3
Signal Processing, Computer-Assisted	3	5	1	0
Solubility	0	0	1	0
Stress, Mechanical	4	0	5	0
Synaptic Transmission:physiology	3	4	2	2
Systems Theory	5	5	1	0
Time	3	4	4	5
Time Factors	3	0	1	0
Ultrasonics	4	4	3	0
Vibration	4	4	5	0

Tabla 25: Comparación entre conglomerados del corpus C4

Se han explorado posibles métodos para reducir el grado de arbitrariedad en la construcción de conglomerados y en construcción de la red global (Tijssen et al., 1989). Imponiendo métricas que construyen mapas de conglomerados con distancias definidas geométricamente. Sin embargo, esta aproximación se considera no apropiada, principalmente por razones teóricas; Callon, señala:

"…una representación métrica tiene dos limitaciones: es difícil interpretarla porqué asocia con cada pareja de elementos de un conjunto un número que constituyen su distancia pero cuyo significado teórico no es siempre claro; hace necesaria la existencia de un espacio bidimensional que es también teóricamente difícil de justificar…"

La aplicación de MDS no métrico a la matriz de co-ocurrencia normalizada, arroja resultados insatisfactorios. Rip y Courtial, señalan:

"…es posible transformar el encadenamiento de diferentes intensidades ha distancias en un plano bidimensional con la ayuda, e. g., técnicas de escalamiento de Kruskal. Esto… también se probó en el primer análisis de palabras asociadas. El problema, sin embargo, es que la proyeccion hacia el plano introduce fuerte "stress" y descriptores… que son cercanos a otros en el plano pueden apartarse en términos de la intensidad del encadenamiento. Ninguna metrización global deberá ser probada para eso, aunque metrización local puede ser útil para clarificar conjuntos complejos de encadenamientos..."

El uso de redes neuronales artificiales, en especial, la red neuronal de Teuvo Kohonen denominada *Self-Organizing Map* ha sido exitosamente aplicada a una gran diversidad de problemas económicos, financieros, etc. En el campo cienciométrico se usa en la exploración de conglomerados. En la siguiente ilustración se muestra solamente el mapa auto-organizante del corpus C4. El mapa auto-organizante muestra 59 conglomerados. Cada temática formo su propio conglomerado. Cuando tratamos de representar los 6 conglomerados del corpus C4 en el mapa auto-organizante obtenemos 11 conglomerados. Los conglomerados "Vibration", "Models, Theoretical" y "Scattering, Radiation" están muy bien definidos, es decir, mantienen cercanas las temáticas que los integran. Los conglomerados "Models, Chemical", "Time Factors" y "Artificial Intelligence" están parcialmente fragmentados en conglomerados pequeños. Tal vez esto indique que los encadenamientos entre las temáticas son tan débiles que en el espacio multidimensional han de estar muy separados.

Ilustración 59: Mapa Auto-Organizante de C4.

Conclusiones

En el capítulo 5 se intento establecer la utilidad de los indicadores cienciométricos, en especial, el indicador relacional de segunda generación denominado análisis de las palabras asociadas, para estudiar las relaciones en la investigación biomédica. El análisis de palabras asociadas ayudo a identificar áreas temáticas que caracterizan la investigación biomédica en diferentes periodos. No solamente se describió el contenido de la investigación en marcha, se desarrollaron y usaron dos índices para poner cada área temática en su correspondiente red de investigación (centralidad) y determinar su grado de desarrollo interno (densidad). Las cadenas entre las áreas temáticas fueron visualizadas para identificar los ejes de investigación que permanecen constantes sobre el tiempo, esos que desaparecen y esos que emergen. Un coeficiente fue calculado y aplicado para medir cambios en la estabilidad del contenido de los ejes temáticos en el tiempo. Este índice de transformación es útil para describir las trayectorias de investigación.

La literatura analizada por medio del análisis de palabras asociadas revelo ser muy rica y compleja en cuanto a contenidos. Hay contenidos computacionales, neurológicos, químicos, genéticos, moleculares, etc., Así, como contenidos poco habituales dinámica de poblaciones, presión hidrostática, técnicas de crecimiento de células in Vitro y desviación de energías. Sin embargo, hay contenidos que se mantienen vigentes durante el transcurso del tiempo al que denominamos serie principal. Esta constituye el centro de interés asociado al MeSH Major Topic “Nonlinear Dynamics”.

La serie principal es central en la estructura de la red. Su reorganización y redefinición es suave durante el transcurso del tiempo. Y mantiene un núcleo de publicaciones asociado a unas cuantas temáticas. La serie principal muestra que el centro de interés “Nonlinear Dynamics” no está bajo tutela de ningún grupo de investigadores que se haga cargo del tema, de forma sistemática y durable.

En resumen, simular aspectos dinámicos de ciertos procesos fisiológicos, ya sean químicos, biológicos, genéticos, neurológicos, moleculares, genéticos, etc., requiere uso de ecuaciones matemáticas, computadoras o algún otro equipo electrónico.

Apéndice

Apéndice A

El Análisis de 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.

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 (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)

El rendimiento del equipo	
Nombre	**flops (potencia)**
Megaflop	10(6)
Gigaflop	10(9)
Teraflop	10(12)
Petaflop	10(15)
Exaflop	10(18)
Zettaflop	10(21)
Yottaflop	10(24)
Xeraflop	10(27)

Apéndice B

Las distribuciones bibliométricas de Bradford, Lotka y Zipf exponen un comportamiento estadístico de la ciencia. Aunque para Gorbea [1] es más apropiado referirse a estas distribuciones bibliométricas como modelos matemáticos que como leyes.

EL modelo matemático 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 [2] y [3]. El modelo se deduce de la siguiente observación: 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 [4], 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:

$$N_r = k^r N_n$$

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 [5] 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 matemático 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.

El modelo de Zipf. En 1932, George Zipf [6], describió el comportamiento estadístico de la distribución de las palabras en un texto a través de un modelo. Él propuso que en un texto, existía una relación matemática entre, la frecuencia de repetición de cada palabra y el lugar que ocupa en el listado de las palabras usadas en el texto, ordenadas por su frecuencia decreciente. Identificaba una palabra en particular y le asignaba un índice ***s*** igual al lugar de la palabra en el listado y un *f(s)*, que es la frecuencia de la repetición de esa palabra, es decir, el número de veces que aparece la palabra en el texto. La ley de Zipf, sostiene la siguiente relación matemática:

$$f(s) = \frac{A}{s^{\alpha}}$$

Donde A es una constante y α es un valor cercano a 1.

[1] Gorbea P. S; ″Modelación Matemática de la Actividad Bibliotecaria: Una Revisión″. Investigación Bibliotecológica v. 12. No. 24 enero/junio de 1998.

[2] Bookstein A; "Robustness Properties of the Bibliometric Distributions". Journal of the American Society for Information Science. 1984.

[3] Ungern-Sternberg S; "Bradford´s Law in the Context of Information Provision", Scientometrics, V.49. No. 1 (2000) 161-186.

[4] 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

[5] Urbizagástegui A; "La Ley de Lotka y la Literatura de Bibliometría". Investigación Bibliotecológica, V.13, NO. 27, 1999.

[6] Felipe C. S; Sergio C. H; "Distribución Estadística en Textos Literarios". Universidad Católica del Norte, Departamento de Física, Av. Angamos 0610, Antofagasta, Chile.

Apéndice C

Factor de Impacto: El factor de impacto de una revista científica se define como el cociente del número de veces que son citados artículos de la revista en cuestión, publicados durante los dos años anteriores al año del reporte del índice, entre el número total de artículos publicados por la revista en ese lapso. En forma breve se puede decir que representa el promedio de citas por artículo y constituye una medida cuantitativa del impacto que los trabajos allí publicados tienen sobre la comunidad científica relacionada específicamente con el tema de la revista.

Índices de Transformación

Se presentan las tablas de los índices de transformación de la serie principal dada por "Nonlinear Dynamics" de C1, "Nonlinear Dynamics" de C2, "Algorithms" de C3 y "Models, Theoretical" de C4. Vea Ilustración 56. (Algunos nombres de columnas y filas están abreviados por cuestiones de tamaño)

		C2						
		Nonlinear	**Numerical**	**Adaptation**	**Software**	**Image**	**Hydrostatic**	**Models,C**
C1	**Nonlinear**	20/9	18/1	-	-	-	-	-
	Nucleic	-	-	-	-	-	-	14/1
	Computer	-	-	-	-	-	-	-
	Models, N	-	-	-	-	20/2	-	-
	Population	-	-	-	-	-	-	-
	Models, G	-	-	-	-	-	-	-
	Motion	-	-	-	-	-	-	20/3
	Physics	-	12/1	-	9/1	-	-	-

		C3			
		Algorithms	Cell	Motion	Pattern
C1	Nonlinear	20/7	-	-	17/2
	Nucleic	-	-	-	-
	Computer	-	-	-	-
	Models, N	-	-	-	-
	Population	-	-	-	-
	Models, G	-	-	-	-
	Motion	-	-	17/3	-
	Physics	-	-	-	-

		C4					
		Models, T	Vibration	Time	Scattering	Artificial	Models, C
C1	Nonlinear	20/7	-	-	-	-	18/1
	Nucleic	-	-	-	-	-	-
	Computer	-	-	-	-	-	-
	Models, N	-	20/1	20/4	-	-	-
	Population	-	-	-	-	-	-
	Models, G	-	-	-	-	-	-
	Motion	-	20/3	-	-	-	18/1
	Physics	14/1	-	14/1	-	-	-

		C3			
		Algorithms	Cell	Motion	Pattern
C2	Nonlinear	20/5	-	-	17/2
	Numerical	18/2	-	-	15/3
	Adaptation	14/1	9/1	-	-
	Software	-	-	-	-
	Image	-	-	17/2	-
	Hydrostatic	-	-	-	-
	Models, C	-	-	17/1	-

		C4					
		Models, T	Vibration	Time	Scattering	Artificial	Models, C
C2	Nonlinear	20/6	-	20/1	-	-	18/1
	Numerical	18/2	-	18/1	-	-	-
	Adaptation	-	-	-	-	-	-
	Software	-	-	15/1	-	-	-
	Image	-	20/1	-	-	16/2	-
	Hydrostatic	-	-	-	-	-	14/1
	Models, C	-	20/2	-	-	-	18/2

		C4					
		Models, T	Vibration	Time	Scattering	Artificial	Models, C
C3	Algorithms	20/8	-	-	-	-	18/1
	Cell	-	-	-	-	-	-
	Motion	-	17/1	-	-	-	15/2
	Pattern	17/1	-	17/2	-	13/3	-

Índices de equivalencia

Cada conglomerado mantiene cadenas internas entre los descriptores que lo integran y cadenas externas con descriptores de otros conglomerados. La fuerza del encadenamiento entre los descriptores ***D1*** y ***D2*** es medida por el indice de equivalencia ***e***. Por ejemplo, en la tabla **Physics** se observan las cadenas internas y las cadenas externas de dicho conglomerado. La fuerza del encadenamiento interno entre los descriptores 28 (Oscillometry) y 36 (Physics) es 0.133. La fuerza del encadenamiento externo entre los descriptores 12 (Systems Theory) y 17 (Artificial Intelligence) es 0.137.

Physics		
Cadenas internas		
D1	***D2***	***e***
28	36	0.133
29	36	0.133
12	36	0.062
Cadenas externas		
12	17	0.137

Conglomerados pertenecientes a la red en 2004

No.	Conglomerados
1	**Nonlinear Dynamics**
2	Computer Simulation
3	Algorithms
4	Models, Statistical
5	Models, Theoretical
6	Time Factors
7	Models, Biological
8	Neural Networks(Computer)
9	Feedback
10	Stochastic Processes
16	**Models, Neurological**
37	Synaptic Transmission:physiology
39	Action Potentials:physiology
40	Nerve Net:physiology
41	Neurons:physiology
66	Electric Stimulation
13	Periodicity
19	Reproducibility of Results
21	Oscillometry:methods
15	Sensitivity and Specificity
48	**Computer-Aided Design**
80	Microwaves
52	Equipment Design
55	Equipment Failure Analysis:methods
44	Quality Control
43	**Nucleic Acid Conformation**
85	Nucleic Acid Denaturation
30	DNA:chemistry
35	Models, Molecular
36	**Physics**
28	Oscillometry
29	Software
12	Systems Theory
22	**Models, Genetic**
64	Computers, Molecular
71	Evolution, Molecular
73	Gene Expression Regulation:physiology
84	Mutation
89	Phenotype
56	Evolution
14	**Motion**
58	Microfluidics:methods
59	Particle Size
70	Complex Mixtures:chemistry
91	Pressure
62	Solutions
11	Models, Chemical
38	Elasticity
45	Rheology:methods
46	Stress, Mechanical

No.	Conglomerados
23	**Population Dynamics**
61	Population Density
65	Ecosystem
72	Food Chain
No.	**Descriptores**
17	Artificial Intelligence
18	Numerical Analysis, Computer-Assisted
20	Signal Processing, Computer-Assisted
24	Biological Clocks:physiology
25	Feedback:physiology
26	Fractals
27	Homeostasis:physiology
31	Diffusion
32	Fuzzy Logic
33	Kinetics
34	Linear Models
42	Normal Distribution
47	Temperature
49	Electromagnetic Fields
50	Energy Transfer
51	Environment
53	Equipment Design:methods
54	Equipment Failure Analysis
57	Forecasting
60	Pattern Recognition, Automated
63	Thermodynamics
67	Electrostatics
68	Entropy
69	Environmental Monitoring
74	History, 20th Century
75	Information Storage and Retrieval:methods
76	Lasers
77	Least-Squares Analysis
78	Logistic Models
79	Metabolism:physiology
81	Models, Economic
82	Adaptation, Physiological
83	Multivariate Analysis
86	Oceans and Seas
87	Acoustics
88	Artifacts
90	Physics:methods
92	Probability
93	Protein Conformation
94	Regression Analysis
95	Bayes Theorem
96	Biophysics
97	Catalysis
98	Anisotropy
99	Computers
100	Torque
101	Water Movements
102	Water Purification:methods
103	Water:chemistry

Índices de equivalencia de los conglomerados pertenecientes a la red en 2004

Nonlinear Dynamics

Cadenas internas

D1	D2	e
1	2	0.437
1	3	0.338
1	4	0.289
1	5	0.239
1	6	0.232
1	7	0.169
1	8	0.169
1	9	0.127
1	10	0.127
2	3	0.21
2	4	0.19
2	10	0.176
2	7	0.151
2	8	0.033
2	9	0.09
2	5	0.03
2	6	0.018
3	8	0.056
3	6	0.04
3	4	0.13
3	10	0.116
3	7	0.105
3	5	0.061
4	6	0.239
4	10	0.164
4	7	0.037
4	9	0.034
4	5	0.103
4	8	0.009
5	6	0.322
5	9	0.026
5	8	0.011
5	10	0.007
7	10	0.037
8	9	0.083
9	10	0.028
6	7	0.02
6	8	0.011
6	9	0.007

Cadenas externas

D1	D2	e
7	23	0.214
7	24	0.174
7	27	0.133
1	11	0.106
1	12	0.092
1	13	0.085
1	14	0.07
1	15	0.063
2	13	0.163
2	14	0.161
2	11	0.13
2	21	0.099
2	24	0.097
2	16	0.088
2	20	0.088
2	35	0.081
2	37	0.081
2	38	0.065
2	39	0.065
2	40	0.065

Cadenas externas

D1	D2	e
2	43	0.065
2	46	0.065
3	17	0.113
3	18	0.113
3	32	0.067
4	18	0.098
4	31	0.098
4	37	0.078
4	20	0.068
4	28	0.065
5	42	0.118
5	48	0.088
5	44	0.066
6	28	0.182
6	36	0.097
6	31	0.068
7	13	0.125
7	61	0.125
7	18	0.116
7	21	0.083
7	65	0.083
7	72	0.083
7	82	0.083
7	25	0.075
8	17	0.227
9	12	0.346
9	17	0.302
9	75	0.111
10	13	0.116
10	50	0.074

Models, Neurological

Cadenas internas

D1	D2	e
13	21	0.26
13	16	0.083
13	39	0.083
13	37	0.067
15	19	0.5
15	16	0.049
15	66	0.222
16	37	0.556
16	39	0.444
16	40	0.444
16	41	0.444
16	19	0.056
16	21	0.056
16	66	0.222
37	39	0.8
37	40	0.8
37	41	0.45
39	40	0.562
39	41	0.25
40	41	0.562

Cadenas externas

D1	D2	e
13	23	0.048
13	72	0.167
13	24	0.125
13	25	0.067
15	49	0.148

Physics

Cadenas internas

D1	D2	e
28	36	0.133
29	36	0.133
12	36	0.062

Cadenas externas

D1	D2	e
12	17	0.137

Population Dynamics

Cadenas internas

Models, Genetic

Cadenas internas

D1	D2	e
71	84	1
22	64	0.333
22	71	0.333
22	73	0.333
22	84	0.333
22	89	0.333
22	56	0.222

Cadenas externas

Models, Neurological

Cadenas externas

D1	D2	e
15	27	0.089
19	27	0.1
21	24	0.333
21	25	0.225
21	27	0.1

Motion

Cadenas internas

D1	D2	e
11	70	0.133
11	14	0.107
11	58	0.089
11	62	0.089
14	58	0.3
14	59	0.3
14	70	0.2
14	91	0.2
14	62	0.133
14	38	0.1
14	45	0.1
14	46	0.1
38	46	0.562
45	91	0.5
58	59	0.444
62	70	0.667

Cadenas externas

D1	D2	e
11	24	0.044
11	35	0.213
11	43	0.15
11	67	0.133
11	85	0.133
11	93	0.133
11	30	0.12
11	15	0.03
13	45	0.083
14	21	0.05
21	45	0.125
30	38	0.2
35	38	0.2
38	43	0.25

Computer-Aided Design

Cadenas internas

D1	D2	e
44	55	0.75
44	48	0.333
48	80	0.667
48	52	0.444
48	55	0.444
52	80	0.667

Cadenas externas

D1	D2	e
44	53	0.75
52	54	0.444
53	55	0.444

Nucleic Acid Conformation

Cadenas internas

D1	D2	e
30	43	0.8
30	85	0.4
30	35	0.36
35	43	0.45
43	85	0.5

Cadenas externas

D1	D2	e
35	67	0.4

D1	D2	e
23	61	0.429
23	65	0.286
23	72	0.286
61	65	0.667

Cadenas externas

N/A

23	56	0.19

Conglomerados pertenecientes a la red en 2005

No.	Conglomerados
1	**Nonlinear Dynamcs**
2	Models, Statistical
3	Algorithms
4	Computer Simulation
5	Time Factors
6	Models, Theoretical
7	Models, Biological
8	Physics:methods
9	Neural Networks(Computer)
10	Stochastic Processes
37	**Image Interpretation, Computer-Assisted:methods**
47	Sensitivity and Specificity
31	Computer Graphics
33	Reproducibility of Results
58	Hypermedia
72	Patents asTopic
73	Computer Security
75	Product Labeling:methods
38	Information Storage and Retrieval:methods
16	Pattern Recognition, Automated:methods
14	**Models, Chemical**
40	Models, Molecular
34	Stress, Mechanical
52	Elasticity
53	Energy Transfer:physiology
65	Molecular Motor Proteins:chemistry:physiology
67	Movement:physiology
76	Protein Conformation
48	Statistics as Topic
21	Themodynamics
15	**Numerical Analysis, Computer-Assisted**
23	Computing Methodologies
11	Signal Processing, Computer-Assisted
19	Feedback
30	Cluster Analysis
57	Fuzzy Logic
17	Artificial Intelligence
12	Systems Analysis
22	**Adaptation, Physiological:physiology**
20	Feedback:physiology
78	Signal Transduction:physiology
27	Biological Clocks:physiology

No.	Conglomerados
18	**Software**
24	Computers
25	Fourier Analysis
29	Systems Theory
32	Diffusion
36	**Hydrostatic Pressure**
44	Osmotic Pressure
45	Biological Transport
60	Mathematics
71	Cell Membrane Permeability
13	Kinetics
No.	**Descriptores**
26	Normal Distribution
28	Oscillometry
35	Heat
39	Linear Models
41	Adaptation, Physiological
42	Neurons:physiological
43	Oceans and Seas
46	Periodicity
49	Artifacts
50	Temperature
51	Data Interpretation, Statistical
54	Equipment Design
55	Equipment Failure Analysis
56	Blood Pressure:physiology
59	Bayes Theorem
60	Mathematics
62	Models, Genetic
63	Climate
64	Models, Neurological
66	Motion
68	Nanotubes, Carbon:chemistry
69	Nerve Net:physiology
70	Nucleic Acid Denaturation
74	Principal Component Analysis
77	Rheology:instrumentation:methods
79	Aging:physiology

Índices de equivalencia de los conglomerados pertenecientes a la red en 2005

Nonlinear Dynamics

Cadenas internas

D1	D2	e
1	2	0.463
1	3	0.453
1	4	0.326
1	5	0.326
1	6	0.295
1	7	0.242
1	8	0.232
1	9	0.168
1	10	0.168
2	8	0.456
2	6	0.393
2	5	0.355
2	7	0.063
2	3	0.233
2	10	0.205
2	4	0.165
2	9	0.023
3	4	0.3
3	9	0.176
3	5	0.147
3	10	0.145
3	7	0.122
3	8	0.026
4	9	0.073
4	10	0.29
4	5	0.067
4	7	0.202
4	6	0.029
4	8	0.006
5	8	0.287
5	6	0.226
5	7	0.05
5	10	0.032
6	8	0.526
7	10	0.22
7	9	0.011
9	10	0.062
3	6	0.1

Cadenas externas

1	13	0.105
1	14	0.105
1	11	0.137
1	12	0.126
2	28	0.114
2	18	0.182
2	12	0.121
2	24	0.114
2	25	0.114
3	11	0.258
3	15	0.233
3	17	0.209
3	23	0.14
3	12	0.124
4	19	0.166
4	11	0.159
4	14	0.158
4	34	0.129
4	15	0.116
5	24	0.103
5	18	0.198
6	24	0.114
6	25	0.114
6	29	0.114
6	12	0.241
6	18	0.219
6	32	0.143
7	20	0.155
7	27	0.139
7	48	0.13
8	25	0.227
8	18	0.205
8	28	0.145

Image Interpretation, Computer Assisted : methods

Cadenas internas

D1	D2	e
16	37	0.333
16	38	0.333
16	47	0.333
16	31	0.25
16	33	0.25
16	58	0.222
16	72	0.222
16	73	0.222
16	75	0.222
31	37	0.75
31	47	0.75
31	33	0.562
31	58	0.5
31	72	0.5
31	73	0.5
31	75	0.5
31	38	0.333
33	58	0.5
33	72	0.5
33	73	0.5
33	75	0.5
33	38	0.333
37	47	1
37	58	0.667
37	72	0.667
37	73	0.667
37	75	0.667
37	38	0.444
38	47	0.444
47	58	0.667
47	72	0.667
47	73	0.667
47	75	0.667
58	72	1
58	73	1
58	75	1
72	73	1
72	75	1
73	75	1

Cadenas externas

1	16	0.095
2	37	0.068
2	47	0.068
2	16	0.063
2	31	0.051
2	58	0.045
2	72	0.045
2	73	0.045
2	75	0.045
3	37	0.07
3	38	0.07
3	47	0.07
3	31	0.052
3	33	0.052
3	16	0.209
3	58	0.047
3	72	0.047
3	73	0.047
3	75	0.047
4	33	0.073
4	37	0.043
4	38	0.043
4	47	0.043
4	16	0.129
7	16	0.077
7	38	0.058
9	16	0.174
10	37	0.083
10	47	0.083

Numerical Analysis, Computer - Assisted

Cadenas internas

D1	D2	e
11	30	0.077
11	15	0.277
11	17	0.214
11	23	0.115
12	15	0.033
15	23	0.6
15	19	0.229
15	30	0.225
15	57	0.2
17	30	0.25
17	23	0.167
23	30	0.375
23	57	0.333
12	19	0.107
15	17	0.1
11	19	0.099
19	23	0.095

Cadenas externas

12	29	0.417
12	24	0.267
12	25	0.067
12	18	0.167
19	77	0.286

Hydrostatic Pressure

Cadenas internas

D1	D2	e
13	36	0.3
13	44	0.3
13	45	0.3
13	60	0.2
13	71	0.2
36	44	1
36	45	1
36	60	0.667
36	71	0.667
44	45	1
44	60	0.667
44	71	0.667
45	60	0.667
45	71	0.667
60	71	1

Cadenas externas

N/A

Software

Enlaces intenos

D1	D2	e
18	24	0.4
18	25	0.225
18	29	0.225
18	32	0.125
24	29	0.36
24	25	0.16

Cadenas externas

25	35	0.267
26	32	0.2

Models, Chemical

Cadenas internas

D1	D2	e
14	40	0.3
14	21	0.057
14	34	0.225
14	52	0.2
14	53	0.2
14	65	0.2
14	67	0.2
14	76	0.2
14	48	0.133
40	52	0.667
40	76	0.667

8	35	0.136	10	31	0.062	48	53	0.667
9	15	0.306	10	33	0.062	48	65	0.667
9	23	0.26	10	16	0.174	48	67	0.667
9	11	0.236	11	16	0.419	53	65	1
9	57	0.125	11	37	0.231	53	67	1
10	48	0.188	11	38	0.231	65	67	1
10	11	0.173	11	47	0.231	**Cadenas externas**		
10	15	0.156	11	31	0.173	21	68	0.286
10	19	0.143	11	33	0.173	13	21	0.229
10	53	0.125	11	58	0.154	13	14	0.04
10	57	0.125	11	72	0.154			
10	65	0.125	11	73	0.154	**Adaptation, Physiological:physiology**		
10	67	0.125	11	75	0.154	**Cadenas internas**		
6	35	0.107	15	16	0.278	**D1**	**D2**	**e**
1	15	0.105	16	23	0.296	20	27	0.457
3	19	0.12	16	30	0.25	20	22	0.381
			16	57	0.222	22	78	0.333
			16	17	0.198	22	27	0.3
			26	33	0.2	**Cadenas externas**		
			2	33	0.091	N/A		

Conglomerados pertenecientes a la red en 2006

No.	Conglomerados	No.	Descriptores
3	**Algorithms**	16	Mechanics
2	Computer Simulation	21	Stochastic Processes
1	Nonlinear Dynamics	24	Diffusion
5	Models, Statistical	25	Equipment Failure Analysis:methods
4	Models, Biological	26	Electromagnetic Fields
7	Feedback	28	Periodicity
11	Systems Theory	30	Computer-Aided Design
6	Biological Clocks:physiology	31	Fuzzy Logic
12	Neural Networks(Computer)	33	Adaptation, Physiological:physiology
8	Oscillometry:methods	34	Lasers
		35	Numerical Analysis, Computer-Assisted
29	**Cell Culture Techniques:methods**	36	Optics
41	Bioreactors:microbiology	37	Reproducibility of Results
43	Cell Proliferation	38	Signal Processing, Computer-Assisted:instrumentation
61	Bacterial Physiology	39	Data Interpretation, Statistical
20	Feedback:physiology	40	Elasticity
		42	Fractals
18	**Motion**	44	Linear Models
19	Rheology:methods	45	Air Pollutants:analysis
62	Water	46	Models, Molecular
23	Water Movements	47	Models, Neurological
27	Energy Transfer	48	Monte Carlo Method
17	Models, Chemical	49	Movement:physiology
9	Kinetics	50	Nerve Net:physiology
		51	Neurons:physiology
15	**Pattern Recognition, Automated:methods**	52	Oceans and Seas
22	Artificial Intelligence	54	Quantum Theory
53	Cluster Analysis	55	Regression Analysis
32	Information Storage and Retrieval:methods	56	Sample Size
10	Models, Theoretical	57	Semiconductors
13	Signal Processing, Computer-Assisted	58	Sensitivity and Specificity
14	Time Factors	59	software
		60	Artifacts

Índices de equivalencia de los conglomerados pertenecientes a la red en 2006

Algorithms

Cadenas internas

D1	D2	e
1	2	0.757
1	3	0.692
1	4	0.308
1	5	0.28
1	6	0.187
1	7	0.187
1	8	0.178
1	11	0.159
1	12	0.15
2	3	0.705
2	4	0.383
2	5	0.37
2	6	0.247
2	7	0.247
2	8	0.235
2	11	0.186
2	12	0.111
3	5	0.305
3	4	0.256
3	7	0.244
3	11	0.23
3	6	0.219
3	12	0.19
3	8	0.182
4	6	0.606
4	5	0.292
5	6	0.202
5	12	0.052
5	7	0.015
7	11	0.576
7	8	0.379
7	12	0.078
8	11	0.375
11	12	0.033

Cadenas externas

D1	D2	e
1	9	0.159
1	10	0.159
1	13	0.121
1	14	0.121
1	15	0.093
1	16	0.084
1	17	0.084
1	18	0.084
1	19	0.075
2	9	0.142
2	13	0.115
2	16	0.111
2	19	0.099
2	10	0.088
2	18	0.088
2	14	0.077
2	22	0.074
3	13	0.15
3	9	0.134
3	16	0.122
3	15	0.109
3	10	0.096
3	22	0.081
3	21	0.069
4	20	0.156

Cadenas externas

D1	D2	e
4	24	0.097
4	29	0.091
5	14	0.092
7	16	0.45
7	9	0.424
7	25	0.09
7	15	0.08
7	10	0.074
8	16	0.474
8	9	0.375
8	38	0.158
8	28	0.118
8	52	0.105
8	57	0.105
8	60	0.105
9	11	0.419
10	12	0.18
11	16	0.529
11	31	0.078
12	13	0.236
12	15	0.225
12	14	0.173
12	22	0.167
12	32	0.083
12	35	0.083

Motion

Cadenas internas

D1	D2	e
9	17	0.163
9	18	0.026
17	18	0.049
18	19	0.222
18	62	0.222
18	23	0.2
18	27	0.111
19	23	0.4
19	62	0.25

Cadenas externas

D1	D2	e
9	16	0.529
9	25	0.106
17	26	0.111
19	52	0.25
23	52	0.4

Cell Culture Techniques:methods

Cadenas internas

D1	D2	e
20	29	0.429
20	41	0.286
20	43	0.286
20	61	0.286
29	41	0.667
29	43	0.667
29	61	0.667
41	43	1

Cadenas externas

N/A

Pattern Recognition, Automated:methods

Cadenas internas

D1	D2	e
10	22	0.088
10	15	0.053
10	13	0.041
10	14	0.041
13	14	0.148
13	15	0.123
13	32	0.103
14	15	0.031
15	22	0.267
15	53	0.2
15	32	0.133
22	53	0.333

Cadenas externas

D1	D2	e
10	40	0.118
14	25	0.062

Conglomerados pertenecientes a la red en 2007

No.	Conglomerados
4	**Models, Theoretical**
3	Algorithms
1	Nonlinear Dynamics
2	Computer Simulation
10	Feedback
15	Decision Support Techniques
6	Neural Networks(Computer)
13	Systems Theory
18	Mathematics
5	Models, Statistical
7	**Time Factors**
28	Oscillometry
52	Electrochemistry:methods
60	Action Potentials:physiology
61	Neurons:physiology
74	Synaptic Transmission:physiology
75	Time
26	Models, Neurological
11	Signal Processing, Computer-Assisted
36	Diffusion
33	**Scattering, Radiation**
58	Light
70	Radiation Dosage
43	Absorption
12	**Models, Chemical**
65	Colloids:chemistry
73	Solubility
37	Electrostatics
40	Models, Molecular
47	Catalysis
24	Kinetics
8	Models, Biological
9	**Artificial Intelligence**
14	Pattern Recognition, Automated:methods
23	Information Storage and Retrieval:methods
50	Computing Methodologies
62	Cluster Analysis
72	Robotics:methods

No.	Conglomerados
34	**Vibration**
59	Materials Testing:methods
22	Elasticity
31	Stress, Mechanical
46	Reproducibility of Results
49	Ultrasonics
32	Equipment Failure Analysis
25	Linear Models
19	Motion
No.	**Descriptores**
16	Biological Clocks:physiology
17	Feedback:physiology
20	Numerical Analysis, Computer-Assisted
21	Data Interpretation, Statistical
27	Optics
29	Oscillometry:physilogy
30	Physics:methods
35	Acoustics
38	Entropy
39	Equipment Design
41	Nerve Net:physiology
42	Noise
44	Biophysics:methods
45	Principal Component Analysis
48	Stochastic Processes
51	Artifacts
53	Electrons
54	Fuzzy Logic
55	History, 19th Century
56	Biological Clocks
57	Least-Squares Analysis
63	Nonlinear Dynamics:history
64	Bayes Theorem
66	Pattern Recognition, Automated
67	Protein Binding
68	Quantitative Structure-Activity Relationship
69	Quantum Dots
71	Rheology:methods
76	Computer-Aided Design
77	Water

Índices de equivalencia de los conglomerados pertenecientes a la red en 2007

Vibration

Cadenas internas			Cadenas internas		
D1	**D2**	**e**	**D1**	**D2**	**e**
19	46	0.167	31	46	0.267
19	49	0.167	31	49	0.267
19	34	0.125	31	32	0.2
19	22	0.1	32	59	0.5
19	31	0.1	32	46	0.333
22	34	0.45	32	49	0.333
22	59	0.4	32	34	0.25
22	31	0.36	34	59	0.5
22	46	0.267	34	46	0.333
22	49	0.267	34	49	0.333
22	32	0.2	46	49	1
25	34	0.2	46	59	0.667
31	34	0.45	49	59	0.667
31	59	0.4	**Cadenas externas**		
			19	35	0.167

Scattering, Radiation

Cadenas internas		
D1	**D2**	**e**
33	58	0.5
33	70	0.5
33	43	0.333
Cadenas externas		
N/A		

Models. Theoretical

Cadenas internas

D1	D2	e
1	2	0.611
1	3	0.549
1	4	0.345
1	5	0.257
1	6	0.212
1	10	0.142
1	13	0.097
1	15	0.08
1	18	0.071
2	3	0.561
2	4	0.357
2	18	0.045
2	5	0.22
2	10	0.178
2	6	0.155
2	15	0.13
2	13	0.065
3	4	0.258
3	10	0.198
3	6	0.151
3	15	0.145
3	5	0.142
3	13	0.023
4	10	0.231
4	15	0.231
4	6	0.209
4	13	0.189
4	18	0.157
4	5	0.149
5	13	0.254
5	18	0.155
5	10	0.019
5	6	0.006
6	10	0.44
6	15	0.296
6	18	0.021
10	15	0.444
10	13	0.023
13	18	0.102

Cadenas externas

1	7	0.204
1	8	0.177
1	9	0.142
1	11	0.133
1	12	0.097
1	14	0.088
2	9	0.178
2	14	0.145
2	11	0.139
2	8	0.122
2	7	0.076
3	8	0.206
3	9	0.198
3	11	0.182
3	14	0.131
3	16	0.129
3	17	0.129

Cadenas externas

D1	D2	e
3	23	0.081
4	30	0.128
4	7	0.09
4	11	0.084
4	28	0.082
4	9	0.079
4	38	0.077
4	44	0.077
5	7	0.121
5	30	0.11
5	38	0.103
5	44	0.103
5	48	0.103
5	11	0.083
6	9	0.167
6	50	0.083
6	54	0.083
6	66	0.083
6	23	0.075
6	26	0.075
9	10	0.098
10	72	0.125
10	11	0.104
10	14	0.1
13	28	0.164
13	30	0.164
13	36	0.121
13	38	0.121
13	44	0.121
13	47	0.121
18	30	0.4
18	75	0.25
18	44	0.167

Models, Chemical

Cadenas internas

D1	D2	e
8	12	0.018
12	65	0.182
12	73	0.182
12	37	0.121
12	40	0.121
12	47	0.121
12	24	0.073
24	47	0.267

Cadenas externas

8	16	0.306
8	17	0.306
8	21	0.04
8	56	0.1
8	48	0.067
20	24	0.1

Time factors

Cadenas internas

D1	D2	e
7	28	0.139
7	52	0.087
7	60	0.087
7	61	0.087
7	74	0.087
7	75	0.087
7	26	0.078
7	11	0.072
7	36	0.058
11	26	0.053
26	60	0.4
26	61	0.4
26	74	0.4
28	36	0.267
60	61	1
60	74	1
61	74	1

Cadenas externas

7	19	0.049
7	23	0.035
7	29	0.035
7	14	0.017
7	12	0.016
7	9	0.011
7	41	0.058
7	47	0.058
7	8	0.009
9	11	0.067
11	19	0.033
11	14	0.107
11	20	0.075
11	29	0.053
12	36	0.121
26	41	0.6
28	47	0.267
28	30	0.16
36	47	0.444
41	60	0.667
41	61	0.667
41	74	0.667

Artificial Intelligence

Cadenas internas

D1	D2	e
9	14	0.625
9	23	0.2
9	50	0.125
9	62	0.125
9	72	0.125
14	23	0.32
14	72	0.2

Cadenas externas

N/A

Revistas del Núcleo

Revista	Frecuencia
Chaos	42
IEEE Trans Neural Netw	29
Phys Rev E Stat Nonlin Soft Matter Phys	12
Neural Netw	6
IEEE Trans Syst Man Cybern B Cybern	5
J Zhejiang Univ Sci	4
Biophys J	2
IEEE Trans Image Process	2
ISA Trans	2
Philos Transact A Math Phys Eng Sci	2
Phys Rev Lett	2
Biotechnol Prog	1
BMC Bioinformatics	1
Neural Comput	1
Comput Methods Programs Biomed	1
J Pharmacokinet Pharmacodyn	1

Países con Publicaciones en el Núcleo

País	Frecuencia
USA	27
China	24
UK	11
Singapore	7
Taiwan	5
France	5
Germany	4
Canada	4
Italy	3
Mexico	3
Spain	3
Brazil	2
Hong Kong	2
Belgium	2
Turkey	2
India	1
Japan	1
Nigeria	1
Slovenia	1
Greece	1
Poland	1
Hungary	1
Jordan	1

Apéndice D

Se presentan los dendrogramas para los corpus C1, C2, C3 y C4. Se utilizo el método jerárquico de conglomerado denominado *"Vecino más Cercano"* con una distancia euclidiana.

Ilustración: Dendrograma del Corpus C1

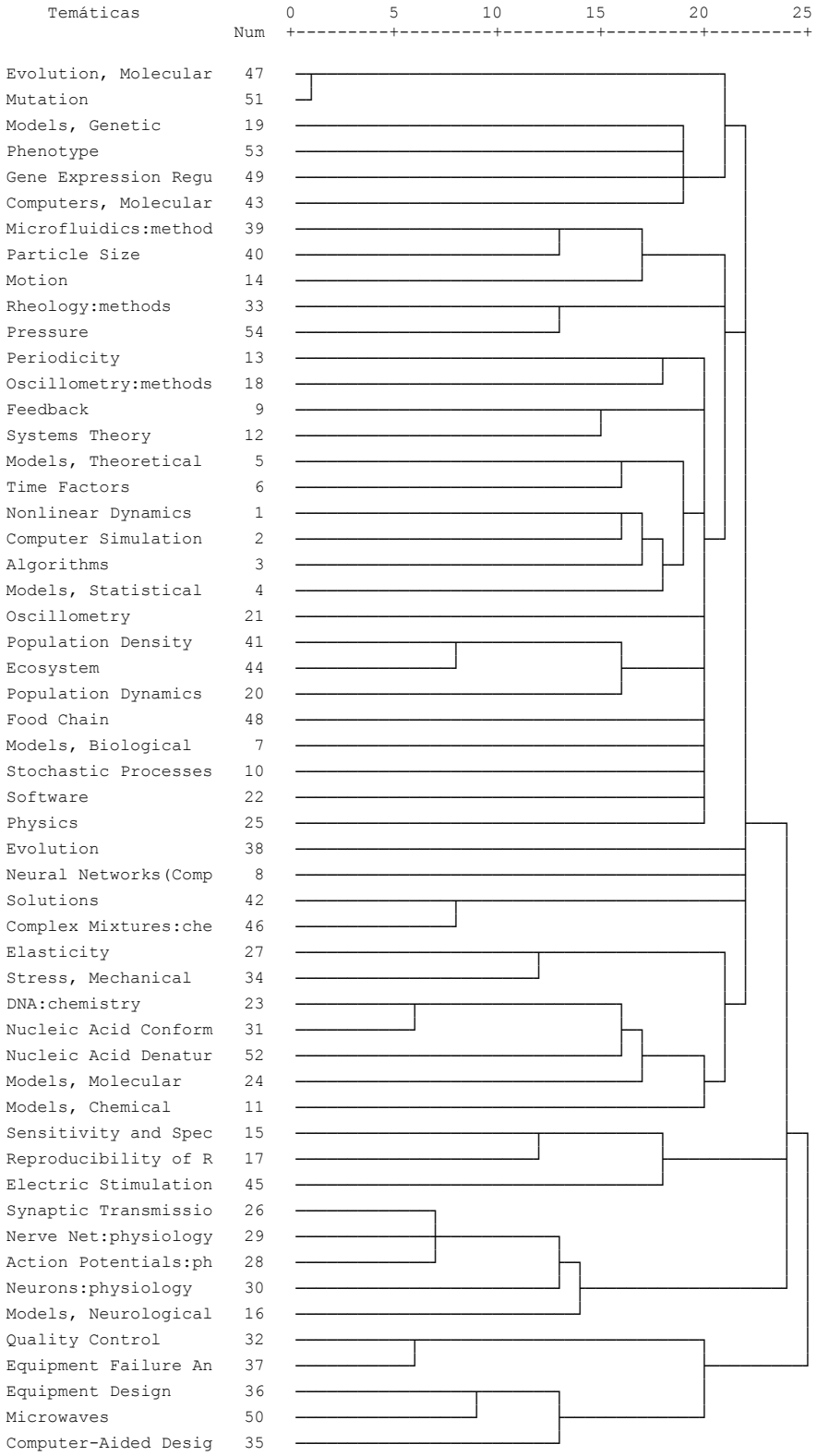


Ilustración: Dendrograma del Corpus C2

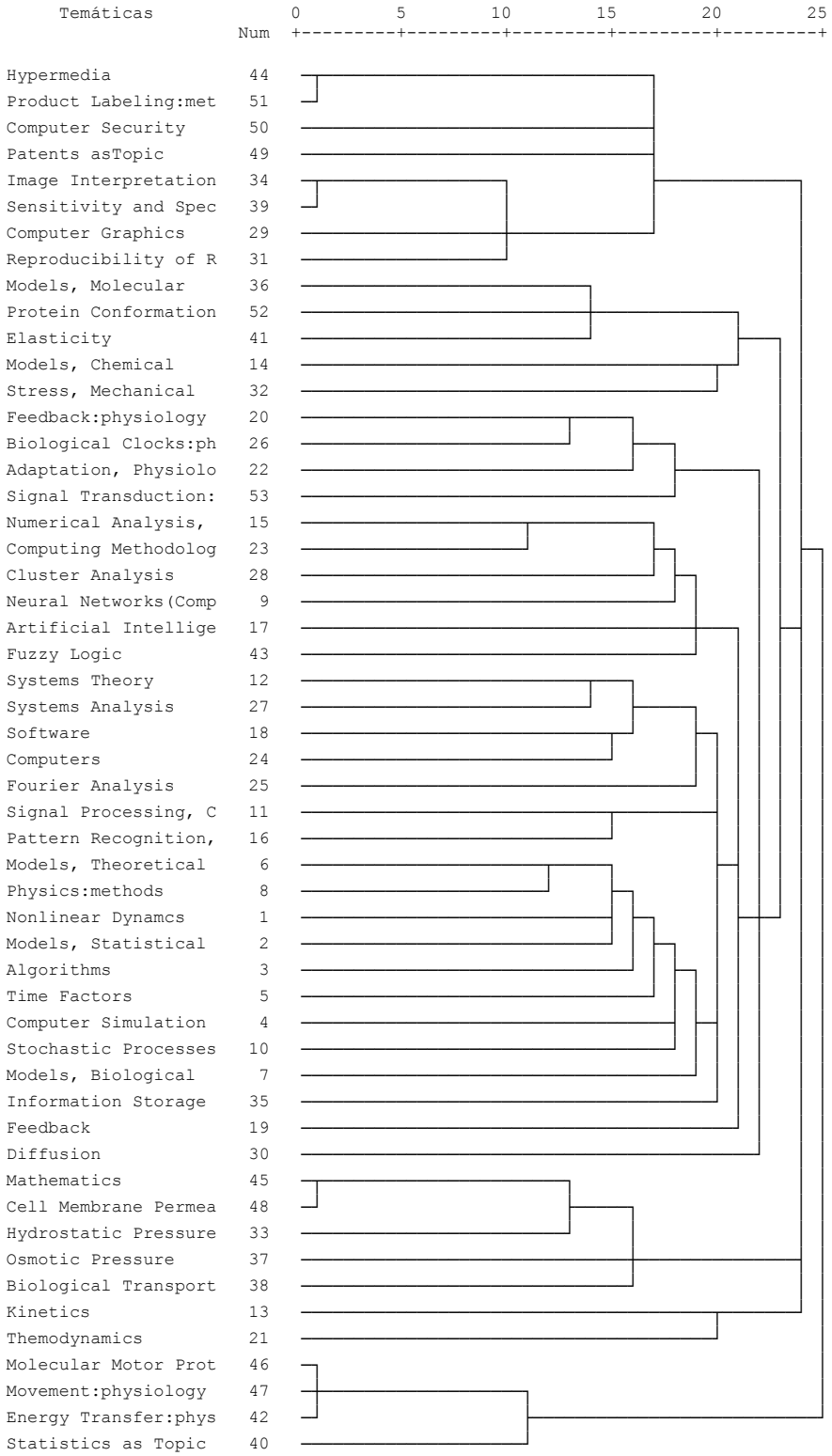


Ilustración: Dendrograma del Corpus C3

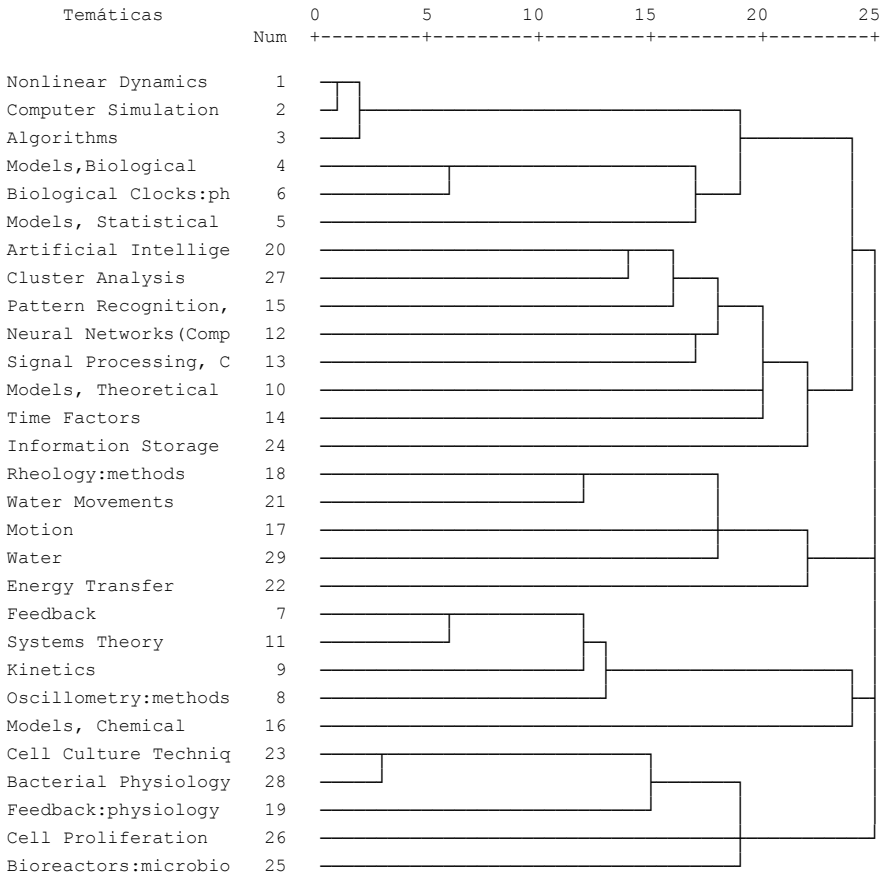
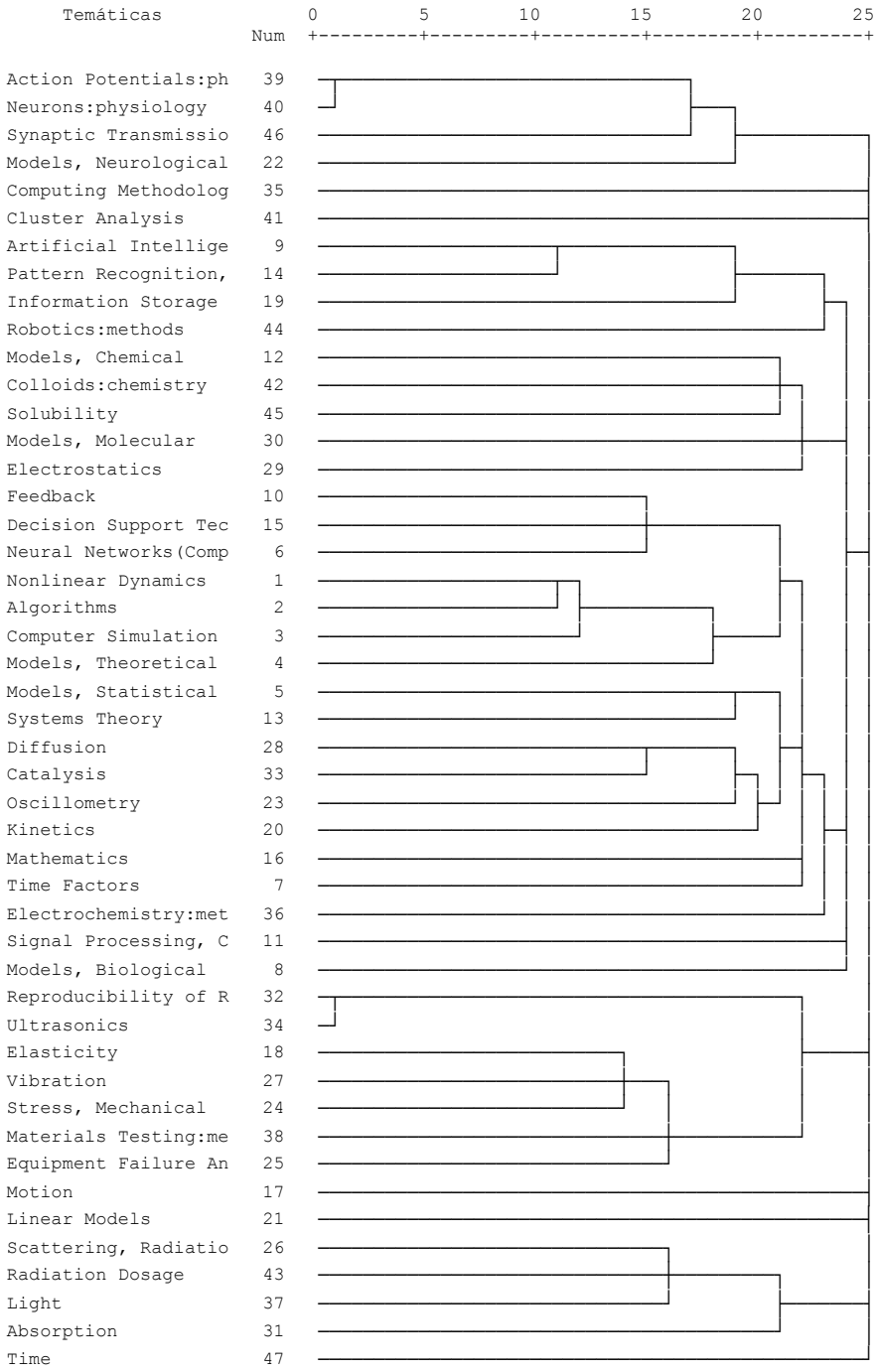


Ilustración: Dendrograma del Corpus C4



Bibliografía

[Agrawal et al., 1994]. Agrawal R; Srikant R; "Fast Algorithms for Mining Association Rules". In Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, 1994.

[Bailón-Moreno et al., 2005]. Bailón-Moreno R; Jurado-Alameda E; Ruiz-Baños R; Courtial J; "The Unified Scientometric Model. Fractality and Transfractality". Scientometrics, Vol, 63. No, 2 (2005), 231-257.

[Barry, 2001]. Barry de Ville, "Microsoft Data Mining: Integrated Bussines Intelligence for e-Commerce and Knowledge Management", Digital Press, 2001.

[Calvelo, 2000]. Calvelo-Ríos M; "El Papel de las Tecnologías de Información y Comunicación en el Desarrollo Rural y la Seguridad Alimentaría". 2000. Disponible en *http://www.fao.org/sd/CDdirect/CDre0055e.htm*, Acceso 07/06/08.

[Callon et al., 1995]. Callon M; Pierre J; "Cienciometría: la medicion de la actividad cientifica: de la bibliometría a la vigilancia tecnologica". Gijon, Trea, 1995.

[Courtial et al., 1991]. Callon M; Courtial J. P; Laville F; "Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry", Scientometrics, 22(1), 155-205. 1991

[Courtial el al., 1984]. Courtial J. P; Rip A; "Co-Word Maps of Biotechnology: an Example of Cognitive Scientometrics". Scientometrics, 6(6), 381-400. 1984.

[Courtial, 1990]. Courtial J. P; (1990). Introduction a la scientométrie: de la bibliométrie a la veille techonologique. Paris, Anthropos.

[Courtial, 1986]. Courtial J. P; "Mapping the dynamics of science and technology: Sociology of science in the real world". London: MacMillan Press LTD. 1986.

[Delmater et al., 2001]. Delmater, R. Hancock, M. "Data mining Explained: A Manager´s Guide to Customer-Centric Business Intelligence". Digital Press, 2001.

[Dunkel et al., 1997]. Dunkel B; Soparkar N; Szaro J; Uthurusamy R; "Systems for KDD: From Concepts to Practice". Future Generation Computer Systems, 13(2-3):231--242, November 1997.

[Fayyad et al., 1996A]. Fayyad U; Haussler D; Stolorz P; "KDD for Science Data Analysis: Issues and Examples". In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 50--56. Menlo Park, CA: AAAI Press. 1996

[Fayyad et al., 1996B]. Fayyad U; Piatetsky-Shapiro G; Smyth; "Knowledge Discovery and Data Mining toward a Unifying Framework". In Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining, pages 82--88, 1996.

[Fayyad et al., 1996C]. Fayyad U; Piatetsky-Shapiro G; "The KDD Process for Extracting Useful Knowledge from Volumenes of Data". Communications of the ACM, 39[11]: 27-34, 1996.

[Fayyad et al., 1996D]. Fayyad U; Piatetsky-Shapiro G; "From Data Mining to Knowledge Discovery: An Overview". Advantedge in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.

[Fayyad et al., 2002]. Fayyad U; Grintein G; Wierse A; "Information Visualization in Data Mining and Kowledge Discovery". Morgan Kaufmann, 2002.

[Frawley et al., 1992]. Frawley W. J; Piatetsky-Shapiro G; Matheus C. J; "Knowledge Discovery in Databases: An Overview". Knowledge Discovery in Databases, pages 1-27. AAAI/MIT Press, 1991.

[Gurjeva, 1992]. Gurjeva G; "Early Soviet Scientometrics and Scientometricians", Universiteit van Amsterdam. Thesis for the degree of MSc in Science Dynamics. 1992.

[Guzman, 2001]. Guzmán M; "Patentometría: Herramienta para el Análisis de Oportunidades Tecnológicas". Master Gerencia de información en las Organizaciones, Cátedra UNESCO.

[He, 1999]. He Qin; "Knowledge Discovery Through Co-Word Analysis" The Free Library. 1999.

[Hebb, 1949]. Hebb D. O; "The organization of behavior :A neuropsychological theory". New York, J. Wiley, 335 p. 1964c, 1949.

[Hegland, 2003]. Hegland M; "Data Mining: Challenges, Models, Methods and Algorithms". 2003.

[Hilderman et al., 1999]. 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.

[Himberg et al., 2001]. Himberg J; Ahola J; Alhoniemi E; Vesanto, J., Simula, O. "The Self-Organizing Map as a Tool in Knowledge Engineering". Helsinki University of Technology. 2001.

[Holsheimer et al., 1991]. Holsheimer M; Siebes A; "Data Mining: The Search for Knowledge in Databases", Report CS-R9406, ISSN 0169-118X, Amersterdam, The Netherlands 1991.

[Jolliffe, 1986]. Jollife I. T; "Principal Component Analysis". Springer Series in Statistics, 1986.

[Jorge, 2004]. Jorge S; "El PC No Distingue: Solo Ve Datos, El Aumento de los Datos Digitales". 2004. Disponible en *http://weblogs.cfired.org.ar/blog/archives/000226.php*, Acceso 07/06/08.

[Jurado-Alameda et al., 2002]. Jurado-Alameda J; Bailón-Moreno R; "Evaluación a través del Análisis de las Palabras Asociadas (1), Aplicación a la Evaluación de la Investigación Científica y Técnica". Biblioteca de Andalucia. 2002

[Kamber, 2001]. Kamber M; "Data Mining: Concepts and Techiques". Morgan Kaufmman Publishers, 2001.

[Kaski, 1997]. Kaski S; "Data Exploration Using Self-Organizing Maps", Ph. D. Thesis, Helsinki University of Technology, Finland, 1997.

[Kohonen, 2001]. Kohonen T; "The Self-Organizing Map", 3ra. Edición Verlag, Springer, 2001.

[Kohonen, 1998]. Kohonen T; "The Self-Organizing Map", Neurocomputing, 21, 1D6. 1998

[Matheus et al., 1993]. Matheus; Chan; Piatetsky-Shapiro G; "Systems for Knowledge Discovery in Databases," IEEE Transactions on Knowledge and Data Engineering, vol. 5, pp. 903-913, 1993.

[Macias-Chapula, 1998]. 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, 1998.

[Michelet, 1988]. Michelet B; "L'analyse des associations", Thèse de doctorat, Université de Paris VII, UFR de Chimie, Paris, 26 Octobre 1988. Spécialité: Information Scientifique et Technique.

[Michie et al., 1994]. Michie D. J; Spiegelhalter C; "Machine Learning, Neural and Statistical Classification". Ellis Horwood, 1994.

[Oscar, 2005]. Oscar J; "La Difusión de la Actividad Científica mediante Publicaciones". Instituto de Historia de la Ciencia y Documentación. 2005

[Pubmed, 2008]. Toda la Informacion disponible en: *http://www.ncbi.nlm.nih.gov*, Acceso 10/06/08.

[Ruiz-Baños et al., 1998]. Ruiz-Baños R; Bailón-Moreno R; "El Método de las Palabras Asociadas (1): La Estructura de las Redes Científicas". Boletín de la Asociación Andaluza de Bibliotecarios, No. 53, pp. 43-60, 1998.

[Ruiz-Baños et al., 1999]. Ruiz-Baños R; Bailón-Moreno R; "El Método de las Palabras Asociadas (2): Los Ciclos de Vida de los temas de investigación". Boletín de la Asociación Andaluza de Bibliotecarios, No. 54, pp. 59-71, 1999.

[Sánchez-Paus, 2002]. 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. 2002.

[Sancho, 1990]. Sancho R; "Indicadores Bibliométricos utilizados en la Evaluación de la Ciencia y la Tecnología". Revisión bibliográfica; Rev. Esp. Doc. Cienti.13, 3-4, 1990.

[Stegmann, 2003]. Stegmann J; Grohmann G; "Hypothesis Generation by Co-Word Clustering", Scientometrics, 56(1), 111-135, 2003.

[Silberschatz et al., 1996]. 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.

[Sotolongo, 2002]. 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.

[Spinak, 2001]. Spinak E; "Indicadores Cienciométricos", ACIMED v.9 n.s supl.4, 2001.

[Tijssen et al., 1989]. Tijssen R. J. W; Van Raan A. F. J; "Mapping Co-Word Structure: A Comparison of Multidimensional Scaling and Leximappe", Scientometrics, Vol. 15, Nos 3-4 (1989), 283-295.

[Vanti, 2000]. Vanti N; "Métodos Cuantitativos de Evaluación de la Ciencia: Bibliometría, Cienciometría e Informetría". Investigación Bibliotecológica, v14, No. 29, 2000.

[Venegas, 2003]. Vanegas A. N; "Inventario Breve de Índices e Indicadores de Ciencia y Tecnología", Universidad de Antioquia, 2003.

[Villaseñor, 2004]. Villaseñor E; "Análisis Inteligente de Datos con Redes Neuronales Artificiales". UNAM, Facultad de Ciencias, Tesis de Licenciatura. 2004.

[Williams et al., 1996]. 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.

[Wikipedia]. http://es.wikipedia.org/wiki/Base_de_datos. Acceso el 5/02/09