UNIVERSIDAD DE GRANADA

Escuela Técnica Superior de Ingeniería Informática

Departamento de Ciencias de la Computación
e Inteligencia Artificial

SciMAT: Herramienta Software para el Análisis de la Evolución del Conocimiento Científico. Propuesta de una Metodología de Evaluación

Memoria de Tesis presentada por

D. Manuel Jesús Cobo Martín

para optar al grado de Doctor en Informática

Granada Septiembre de 2011

Editor: Editorial de la Universidad de Granada
Autor: Manuel Jesús Cobo Martín
D.L.: GR 1060-2012
ISBN: 978-84-695-1069-8

UNIVERSIDAD DE GRANADA

Escuela Técnica Superior de Ingeniería Informática

Departamento de Ciencias de la Computación
e Inteligencia Artificial

ugr | Universidad de Granada

SciMAT: Herramienta Software para el Análisis de la Evolución del Conocimiento Científico. Propuesta de una Metodología de Evaluación

Memoria de Tesis presentada por

D. Manuel Jesús Cobo Martín

para optar al grado de Doctor en Informática

Dirigida por

Dr. Antonio Gabriel López Herrera

Dr. Enrique Herrera Viedma

Dr. Francisco Herrera Triguero

Granada Septiembre de 2011

La memoria titulada **SciMAT: Herramienta Software para el Análisis de la Evolución del Conocimiento Científico. Propuesta de una Metodología de Evaluación**, que presenta **D. Manuel Jesús Cobo Martín** para optar al grado de Doctor en Informática, ha sido realizada en el **Departamento de Ciencias de la Computación e Inteligencia Artificial** de la Universidad de Granada bajo la dirección de los Doctores **Antonio Gabriel López Herrera**, **Enrique Herrera Viedma** y **Francisco Herrera Triguero**.

D. Manuel Jesús Cobo Martín
Doctorando

Dr. Enrique Herrera Viedma
Director

Dr. Antonio Gabriel López Herrera
Director

Dr. Francisco Herrera Triguero
Director

8 de Septiembre de 2011

Agradecimientos

Quisiera dedicar esta memoria de tesis a aquellas personas que me han ayudado durante todo este largo periodo que finalmente termina de la mejor manera que podíamos esperar.

Ante todo a mis padres, a mi hermana y al resto de mi familia porque todo lo que he conseguido ha sido gracias a ellos, a su apoyo, cariño y compresión en las dificultades por las que he pasado desde que empecé en este proceso. Estoy muy orgulloso de ellos por brindarme las oportunidades que me han dado.

Asimismo, quiero expresar mi más sentido agradecimiento a mi directores de tesis, Antonio Gabriel López Herrera, Enrique Herrera Viedma y Francisco Herrera Triguero, ya que sin cuya dedicación, esfuerzo, entusiasmo y confianza, esta memoria jamás habría visto la luz. Gracias por todos los valiosos consejos que me habéis dado, me dais, y a buen seguro me seguiréis dando.

A todos los miembros del grupo de investigación *Soft Computing y Sistemas De Información Inteligentes*, en especial a Nacho, Javi, Julián, Alberto, Álvaro, Isaac, Joaquín y Victoria, y a los *seniors* del grupo como Jesús y Rafa Alcalá, Carlos Porcel y Manolo Lozano. Tampoco puedo olvidarme de mencionar a los compañeros de estancia con los que he convivido como Pietro, Albert, Yosuke y Josean y muchos más.

Asimismo, también quisiera agradecérselo a mis amigos de toda la vida como Juanda,

Jorge, Luisa, Edu, Mati, Silvia, Pili, Tomi y Camacho. En especial quisiera agradecérselo a Salvi, que aunque puede considerarse como compañero de trabajo, es más un amigo que otra cosa. Gracias, por tus consejos y largas charlas.

Y finalmente, una y última gran mención a María con la que llevo compartiendo ocho años de mi vida, y que sin duda han sido los más felices. Gracias por tu comprensión y por acompañarme en este largo viaje.

Índice general

Índice de Tablas

Índice de Figuras

Planteamiento, Justificación y Objetivos

Hoy en día gracias a las nuevas tecnologías y, en concreto, a Internet, tenemos acceso a gran cantidad de información, algo que tan sólo hace unos años no era viable. En particular, la información científica disponible se ha visto enormemente incrementada, e Internet ha facilitado a los científicos el acceso a millones de documentos a los que antes no podían tener acceso. Gran parte de esta mejora es debida a las bases de datos bibliográficas o bibliométricas disponibles en Internet, entre las que destacan *ISI Web of Science*, *Scopus* y *Google Scholar*, por ser las más ampliamente conocidas y utilizadas. De igual modo, existe una gran cantidad de páginas web de investigadores, grupos de investigación y universidades donde se puede acceder a su producción científica.

Sin embargo, aunque cada vez es mayor la cantidad de información a la que se tiene acceso, la capacidad humana para interpretar, analizarla, comprenderla y, finalmente, generar nuevo conocimiento a partir de ella, no ha variado con el tiempo. De hecho, el exceso de información, dificulta enormemente su interpretación y análisis. Debido a ello, los investigadores, y en general, la comunidad científica, tienen dificultades a la hora de tomar decisiones adecuadas.

Como consecuencia de este hecho, surge la necesidad de simplificar y transformar la información en general, y científica en particular, en conocimiento, con el objetivo de facilitar la toma de decisiones. Por este motivo, es necesaria la creación y desarrollo de

nuevas metodologías, técnicas y herramientas capaces de filtrar la información y generar conocimiento.

Desde el punto de vista de la política científica, mantenerse actualizado en relación a los temas *calientes*, o que están teniendo una mayor atención por parte de la comunidad científica, ayudaría a detectar cuáles son los frentes donde interesa invertir recursos, tales como, dinero, personal, infraestructuras, etc.

Por este motivo, en los últimos años han surgido nuevas y diversas técnicas que facilitan la gestión de grandes cantidades de información. Así, técnicas basadas en inteligencia artificial y minería de textos, como los sistemas de recuperación de información, sistemas de clasificación de documentos, sistemas de recomendaciones y sistemas de filtrado, logran que el usuario sea capaz de detectar la información más relevante para él.

Particularmente, para el tratamiento de la información científica han surgido técnicas especializadas que nos permiten extraer conocimiento de dicha clase de información. Principalmente hablamos de la *bibliometría* y *cienciometría*.

La *cienciometría* [61, 80] puede definirse formalmente como [61]:

> "*El análisis estadístico y sociométrico de la bibliografía científica mediante el uso de modelos matemáticos, y cuyos objetivos se basan en el estudio del tamaño, crecimiento y distribución de la bibliografía científica y en el estudio de la estructura y dinámica social que la producen y utilizan.*"

Es decir, la cienciometería tiene por objetivo analizar y estudiar la ciencia y en particular la calidad de ésta. Por otro lado, la bibliometría [20, 29, 43, 83, 109] puede entenderse como la aplicación práctica de la cienciometría al análisis de las publicaciones científicas.

En la actualidad, la cienciometría se considera un campo científico de pleno derecho ya que tiene sus propias leyes y métodos, como la ley de Zipf [118] de la frecuencia de

las palabras en un texto, la ley de Lotka [62] sobre la productividad de los autores, la ley de Bradford [16] sobre la productividad de las revistas, etc.

Dentro de la bibliometría existen dos procedimientos fundamentales: por un lado la evaluación y análisis del rendimiento y de la producción científica a través de indicadores bibliométricos, esto es, la actividad científica y, por otro, la creación y análisis de mapas científicos.

- La evaluación del rendimiento y de la producción científica, trata de evaluar grupos de *actores científicos* tales como, países, universidades, departamentos o investigadores, así como el impacto de sus actividades investigadoras [72, 109], a partir de datos bibliográficos.

- Los mapas científicos, mapas bibliométricos o cienciogramas (*Science Mapping* en la literatura anglosajona), tratan de mostrar los aspectos estructurales y dinámicos de la información científica [12, 67, 71]. En particular, intentan encontrar una representación de las conexiones intelectuales y sus evoluciones dentro del conocimiento científico [95].

Los mapas científicos pueden englobarse en un marco longitudinal o temporal [39, 81] para, de este modo, analizar los cambios estructurales que se han dado en la información científica a lo largo del tiempo. Es decir, son especialmente útiles en el estudio de la evolución intelectual, conceptual y social de un área científica determinada. Por ejemplo, a través de los mapas científicos, se pueden analizar las temáticas tratadas dentro de un campo científico concreto, así como su evolución a lo largo del tiempo. Esto nos permitiría comprender la evolución conceptual del campo y determinar cuáles han sido las temáticas más importantes y cuáles están siendo las temáticas que actualmente están recibiendo una mayor atención por parte de la comunidad científica. De forma similar,

se podría estudiar la evolución de la estructura social de una comunidad científica, analizando la evolución de las coautorías de las publicaciones científicas. De este modo se podrían comprender la evolución de las colaboraciones internacionales.

El desarrollo de los mapas científicos se ha podido realizar en gran medida gracias a la utilización de herramientas software. En este sentido, existen una gran variedad de herramientas bibliométricas que, aunque no fueron diseñadas para la realización de un análisis de mapas científicos, pueden utilizarse para esta tarea. Por ejemplo, "*Publish or Perish*" [40] es una herramienta que permite evaluar y medir el rendimiento y la producción científica. Además, es común el uso de herramientas pertenecientes a otros ámbitos, como por ejemplo: Pajek [8] o UCINET [11], que provienen del campo del Análisis de Redes Sociales (*SNA* por sus siglas en inglés); o Cytoscape [89], procedente del ámbito de la bioinformática.

Por otro lado, en la actualidad también es posible algunas diferentes herramientas diseñadas explícitamente para realizar un análisis mediante mapas científicos. Entre ellas, destacan: Bibexcel [73], CiteSpace II [23, 24], CoPalRed [6, 5], IN-SPIRE [114], Loet Leydesdorff's Software, Network Workbench Tool [13, 44], Science of Science Tool [88], VantagePoint [77] y VOSViewer [107]. Algunos datos básicos acerca de estas herramientas, pueden observarse en la Tabla 1.

Herramienta	Versión	Año	Desarrollada por
Bibexcel	2010-09-22	2010	Universidad de Umeå (Suecia)
CiteSpace	2.2.R9	2010	Universidad de Drexel (Estados Unidos)
CoPalRed	1.0 beta	2005	Universidad de Granada (España)
IN-SPIRE	5	2010	Laboratorio Nacional del Pacífico Noroeste (Estados Unidos)
Loet Leydesdorff's Software	N/D	N/D	Universidad de Amsterdam (Países Bajos)
Network Workbench Tool	1.0.0	2009	Universidad de Indiana (Estados Unidos)
Science of Science (Sci^2) Tool	0.0.3 alpha	2010	Universidad de Indiana (Estados Unidos)
VantagePoint	7	2010	Search Technology, Inc. (Estados Unidos)
VOSViewer	1.2.1	2010	Universidad de Leiden (Países Bajos)

Tabla 1: Información general de las herramienta existentes para el análisis de mapas científicos.

La presente memoria de tesis se encuadra dentro de éste ámbito bibliométrico, el análisis de mapas científicos.

Justificación

Una vez conocidos los principales conceptos a los que se refiere esta memoria de tesis, nos planteamos una serie de problemas abiertos que nos sitúan en el planteamiento y la justificación de la presente memoria de tesis.

- Existen diversas metodologías para la realización de análisis de mapas científicos, sin embargo, no existe ninguna que se enmarque en un contexto longitudinal y que integre medidas de rendimiento y de calidad en los mapas resultantes. A través de dichas medidas, los mapas científicos podrían enriquecerse para mostrar, además de los aspectos estructurales y dinámicos de un área científica, el rendimiento de los elementos que los componen y su calidad e impacto utilizando indicadores bibliométricos. Por tanto, sería de gran utilidad la creación de una metodología para el análisis de mapas científicos y su evolución, y que utilice medidas bibliométricas para cuantificar los resultados de los diferentes elementos.

- Aunque existen un gran número de herramientas diseñadas para realizar análisis de mapas científicos (Bibexcel, CiteSpace II, CoPalRed, IN-SPIRE, Loet Leydesdorff's Software, Network Workbench Tool, Science of Science Tool, VantagePoint y VOSViewer), estas contienen enfoques parciales del problema. Además, ninguna de ellas incluyen los indicadores bibliométricos como medida para cuantificar la calidad e impacto de los diversos resultados, o lo hacen de manera muy superficial. Por este motivo, está justificado el desarrollo de una herramienta que implemente un metodología de análisis de mapas científicos en un contexto longitudinal, complementada con indicadores bibliométricos y que incorpore un amplio conjunto de utilidades como: gestión completa de la información científica utilizada durante el análisis, posibilidad de configurar las técnicas y medidas utilizadas en la creación

del mapa, así como potentes técnicas de visualización que nos permitan interpretar los resultados de forma sencilla, para que de este modo, podamos extraer conocimiento útil.

Objetivos

La presente memoria de tesis se desarrolla en torno al análisis de mapas científicos, tratando de profundizar en las herramientas existentes, las metodologías basadas en medidas de calidad y rendimiento y en los estudios de áreas específicas de conocimiento. En concreto, los objetivos que persigue esta memoria de tesis son:

- Análisis del estado del arte de las herramientas diseñadas para el análisis de mapas científicos, tanto gratuitas como comerciales, estudiando los pros y contras de cada una de ellas
- Desarrollo de una metodología que integre medidas de rendimiento y de calidad en los mapas científicos resultantes del análisis y que permita realizar el análisis en un marco longitudinal.
- Creación de una nueva herramienta software que guíe al analista en todas las etapas de un análisis de mapas científicos, que incorpore un potente módulo de visualización y preprocesamiento, y que sea capaz de enriquecer los resultados con medidas de rendimiento, calidad e impacto.

Estructura de la Memoria de Tesis

Así, la presente memoria de tesis se divide en cuatro capítulos y dos Anexos y se estructura como sigue:

Capítulo 1 : en este capítulo describimos los aspectos fundamentales del análisis de mapas científicos y de aquellos indicadores bibliométricos que nos permitirán medir la actividad, calidad e impacto de los elementos de dichos mapas.

Capítulo 2 : en este capítulo describimos diversas herramientas diseñadas para realizar análisis de mapas científicos, como Bibexcel, CiteSpace II, CoPalRed, IN-SPIRE, Loet Leydesdorff's Software, Network Workbench Tool, Science of Science (Sci^2) Tool , VantagePoint, VOSViewer. Además, de describir dichas herramientas, realizamos un análisis comparativo profundo de ellas, para de este modo, determinar las ventajas, inconvenientes y diferencias más significativas de cada una de ellas.

Capítulo 3 : en este capítulo se propone una metodología para realizar análisis de mapas científicos en un marco longitudinal, y enriquecidos con indicadores bibliométricos de impacto y de calidad. La metodología nos permitirá analizar los aspectos conceptuales, sociales e intelectuales de un campo científico, así como analizar la evolución de éstos a los largo de diversos periodos de tiempo.

Capítulo 4 : en este capítulo presentamos una nueva herramienta para el análisis de mapas científicos llamada SciMAT. Dicha herramienta implementa la metodología presentada en el Capítulo 3, por lo que nos ofrecerá la oportunidad de identificar la evolución de la estructura social, intelectual y conceptual de un campo científico.

Anexo A : en este anexo mostramos la flexibilidad de SciMAT a través de la realización de un análisis mediante el uso de su API.

Anexo B : en este anexo se muestra el manual de usuario de SciMAT.

Finalmente, algunos comentarios, incluyendo conclusiones finales y trabajos futuros serán esbozados.

Capítulo 1

Introducción al Análisis de Mapas Científicos

El análisis de mapas científicos [12, 67, 71] es una técnica bibliométrica para la extracción de conocimiento a partir de la información científica. Para ello, muestra los aspectos estructurales (sociales, intelectuales y conceptuales) de un campo científico. Además, si dicho análisis se engloba en un marco longitudinal, es posible mostrar también los aspecto dinámicos de dicho campo.

Por otro lado, los indicadores bibliométricos miden aspectos cuantitativos y cualitativos de las publicaciones científicas. En especial, los indicadores basados en citas son especialmente útiles para medir el impacto en la comunidad científica y calidad de dichas publicaciones.

En este Capítulo se pretende introducir aquellos conceptos relacionados con el análisis de mapas científicos, y que son necesarios para la comprensión de la presente memoria de tesis. Además, se dará un breve resumen del estado del arte de los indicadores y medidas bibliométricas. De este modo, este Capítulo se estructura de la forma siguiente: en la Sección 1.1 daremos un breve resumen de los indicadores bibliométricos para medir la actividad científica, así como su calidad. Finalmente, en la Sección 1.2 detallaremos aspectos fundamentales sobre el análisis de mapas científicos, como las

fuentes de información bibliográfica, las unidades de análisis y los distintos tipos de redes bibliométricas, las técnicas de preprocesamiento que pueden emplearse, las diversas medidas que pueden utilizarse para normalizar las redes bibliométricas, las técnicas y algoritmos más comunes en la construcción del mapa bibliométrico, los diferentes métodos de análisis que pueden aplicarse y las técnicas de visualización más comunes.

1.1. Evaluación de la Actividad Científica: Indicadores Bibliométricos

Los indicadores bibliométricos [21, 103] son unas medidas para evaluar numéricamente, de un modo objetivo, las actividades que se desarrollan en el ámbito de la ciencia y la tecnología, y en especial, en la I+D+i, valiéndose del conocimiento científico publicado. Principalmente, este tipo de indicadores han tenido como objetivo principal el análisis de las salidas o resultados científicos desde un punta de vista cuantitativo y cualitativo.

Los indicadores bibliométricos pueden englobarse en tres grandes bloques: indicadores de producción, de visibilidad e impacto, e indicadores basados en el número de citas. A continuación resumiremos los diversos indicadores que se pueden encontrar dentro de estos bloques. Un detallado análisis de los diferentes tipos de indicadores puede consultarse en [103].

- Indicadores de producción: tienen como objetivo el recuento de las publicaciones de los distintos agentes envueltos en la investigación, considerándose como publicaciones los documentos propagados a través de canales formales y públicos. Este tipo de indicadores sirven para medir la cantidad de los resultados, ignorándose diversos aspectos como la calidad y el contenido. Entre los indicadores de pro-

ducción podemos encontrar: número de publicaciones, índice de especialización temática, porcentaje de trabajos indizados en algún repositorio o base de datos bibliográfica, distribución por idioma y tipos documentales, índice de transietoriedad, idiomas de publicación y nivel básico/aplicado [103].

- Indicadores de visibilidad e impacto: la mayor parte de los indicadores de visibilidad e impacto están basados en la contabilización de las citas recibidas por los documentos o por las revistas donde éstos fueron publicados. Este tipo de indicadores pueden dividirse en dos bloques: aquellos basados el *Impact Factor*[1] del *Journal Citation Report*[2] (JCR), o bien a partir de las citas recibidas por los propios documentos. Ambos tipos tienen significados diferentes. Por un lado, los basados en el *Impact Factor* miden la "calidad" de las revistas donde los documentos se han publicado. Por otro lado, los basados en el recuento directo de citas representan la influencia e impacto de documentos concretos o conjunto de documentos, como los publicados por un mismo autor, institución, etc. Por ejemplo, se podría analizar la influencia de un determinado autor en la comunidad científica, midiendo y comparando las citas de dicho autor con otros autores.

    - Indicadores basadas en el *Impact Factor*: este tipo de indicadores se calculan teniendo en cuenta la media de citas recibidas por los trabajos de un año concreto de una revista, en los dos siguientes años consecutivos. Dentro de este bloque podemos encontrar indicadores, como el factor de impacto esperado, factor de impacto ponderado, factor de impacto relativo, potencial investigador, distribución por cuartiles, posición decílica, posición normali-

[1] El factor de impacto es una medida de la frecuencia media con la que los artículos de una determinada revista han sido citados en un año o periodo concreto. Es una de las herramientas de evaluación ofrecidas por el JCR de Thomson Reuters

[2] http://science.thomsonreuters.com/es/productos/jcr/

zada, impacto potencial y número y porcentaje de publicaciones en revistas Top3[3] .

- Indicadores basados en el número de citas: número de citas, promedio de citas, porcentaje de documentos citados y no citados, tasa de citación relativa, índice de atracción, tasa de autocitación, trabajos altamente citados. Además, recientemente se han desarrollado indicadores bibliométricos complejos para el análisis y normalización de citas, como h-index [1, 46], g-index [34], hg-index [2], q^2-index [17] o el índice Crown [65].

- Indicadores de colaboración: para cuantificar la colaboración en la producción científica, se calculan medidas basadas en los autores o instituciones que firman los documentos. Entre este tipo de indicadores podemos encontrar: índice de coautoría, índice de coautoría institucional, patrones de colaboración (local, regional, nacional, internacional), medidas de similaridad, así como la tasa de citación relativa de las co-publicaciones internacionales [10].

Finalmente, existe otro tipo de indicadores denominados *indicadores relacionales* [71, 72], los cuales son un conjunto de técnicas de mapeo que generan representaciones gráficas de la ciencia a través del uso de información de carácter relacional. Este tipo de indicadores forman el segundo procedimiento fundamental de la cienciometría, los mapas científicos. En la siguiente sección detallaremos los aspectos fundamentales del análisis de mapas científicos.

[3] Revistas que ocupan algunas de las tres primeras posiciones en el ranking *Impact Factor* de las diferentes categorías JCR.

1.2. Análisis de Mapas Científicos

Los mapas científicos, también conocidos como mapas bibliométricos o cienciogramas, son una representación espacial de cómo las disciplinas, campos, especialidades científicas, y documentos individuales o autores se relacionan entre sí. Dichos mapas se centran en monitorizar un campo científico, delimitando las subáreas de investigación, para de este modo, comprender su estructura intelectual, social, conceptual y cognitiva, así como analizar su evolución estructural [12, 67, 71, 72]. Al proceso de creación, y análisis de un mapa científico se lo conoce como *Análisis de Mapas Científicos* (*Science Mapping Analyisis*, en la literatura anglosajona).

En la literatura se han diseñado diferentes técnicas para crear mapas científicos a partir de un conjunto de documentos, estableciendo diferentes tipos de relaciones entre ellos. En este sentido se puede analizar un campo científico de forma intelectual, social o conceptual.

Las primeras técnicas que surgieron para realizar mapas científicos, analizaban los aspectos intelectuales de la ciencia. Para ello, establecían las relaciones basándose en las referencias de los documentos científicos, es decir, en su base intelectual. De este modo, Kessler propuso el emparejamiento bibliográfico en 1963 [50] basándose en los documentos que citaban a las mismas referencias. De forma similar, Small propuso en 1973 el análisis de co-citación [93], en el que se analizaban las referencias que se solían citar conjuntamente. Aunque Kessler fue el primero en proponer un tipo de mapa científico, fue sin embargo, el análisis de co-citación de Small el que obtuvo mayor acogida. De este modo, la co-citación se ha utilizado para delimitar un área científica [97], descubrir comunidades de conocimiento [48], nuevos frentes de investigación, [104], así como detectar colegios invisibles [70].

Diez años más tarde, Callon propuso analizar el contenido de los documentos me-

diante las relaciones de co-aparición de los términos contenidos en ellos. De este modo, surgió el análisis de co-palabras [19] como una técnica de análisis efectiva para crear mapas de la literatura científica y así mostrar sus aspectos conceptuales o cognitivos.

Otro tipo mapa científico es aquel que analiza los aspectos sociales de un área de investigación. En este sentido, el análisis de co-autores [37, 74] nos permite analizar la estructura social y las colaboraciones a nivel de investigadores, instituciones o países de un campo científico.

Finalmente, los mapas científicos pueden emplearse para diferentes propósitos, como por ejemplo:

- Analizar la evolución estructural del campo científico a través de diversos periodos de tiempo.

- Medir y cuantificar los resultados utilizando medidas de actividad y calidad basadas en indicadores bibliométricos [48, 71, 96, 97, 104, 109, 110].

En esta sección analizamos y describimos aspectos fundamentales en el análisis de mapas científicos, como: i) las fuentes de información, ii) las unidades de análisis y tipo de redes bibliométricas, iii) el preprocesamiento de la información, iv) las medidas de similitud que pueden emplearse para normalizar las relaciones establecidas entre las unidades de análisis, v) las técnicas que pueden emplearse para crear el mapa científico, vi) los tipos de análisis que se pueden realizar sobre el mapa para obtener conocimiento de éste, vii) las técnicas de visualización más comunes, y por último, viii) la interpretación de los resultados. Estos aspectos forman parte del flujo de trabajo habitual del análisis de mapas científicos [12] (Figura 1.1).

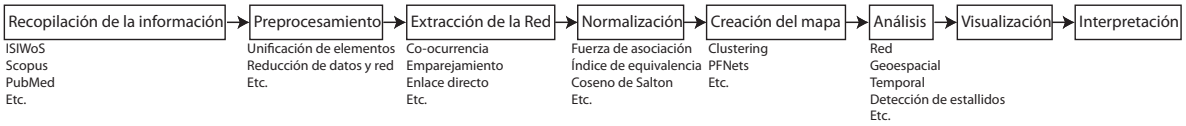


Figura 1.1: Flujo general de trabajo en un análisis de mapas científicos.

1.2.1. Fuentes de Información Bibliográfica

La parte básica y fundamental en un análisis de mapas científicos es la información que utilizaremos en el proceso. Por ello fuentes de información científica como bases de datos bibliográficas y bases de datos de patentes son de vital importancia.

Actualmente, existe una gran variedad de bases de datos bibliográficas, y también bibliométricas, disponibles en Internet que almacenan los trabajos y documentos científicos que los investigadores van desarrollando a lo largo de su carrera, así como las citaciones entre documentos. Las bases de datos de información bibliográfica nos permiten buscar, acceder y recopilar información sobre la mayoría de los campos científicos.

Sin ninguna duda, las base de datos bibliográficas más importantes son *ISI Web of Knowledge*[4] (ISIWoK), *Scopus*[5] , *Google Scholar*[6] y *NLM's MEDLINE*[7] .

La ISIWoS, probablemente la base de datos bibliográfica más famosa e importante, básicamente es un servicio académico ofrecido por *Thomson Reuters* (Estados Unidos) para la indexación de citas entre documentos científicos. La ISIWoK ofrece a sus suscriptores acceso a diferentes bases de datos bibliométricas: Web of Science (ISIWoS), Derwent Innovations Index, BIOSIS, MEDLINE, Zoological Record, entre otras.

De entre las bases de datos que da acceso ISIWoK, ISIWoS es quizás la más conocida de todas. De hecho, es la base de datos académica más importante en ciencias,

[4] http://www.webofknowledge.com
[5] http://www.scopus.com
[6] http://scholar.google.com
[7] http://www.ncbi.nlm.nih.gov/pubmed

ciencias sociales, arte y humanidades. En particular, tiene indexadas las 10.000 revista con mayor impacto mundial, así como 110000 artículos provenientes de conferencias. ISIWoS contiene varias bases de datos de citación: *Science Citation Index Expanded* (SCI-EXPANDED), *Social Sciences Citation Index* (SSCI), *Arts & Humanities Citation Index* (A&HCI), *Conference Proceedings Citation Index in Science* (CPCI-S) y *Social Science & Humanities* (CPCI-SSH). Además contiene dos bases de datos químicas: *Current Chemical Reactions* (CCR-EXPANDED) and *Index Chemicus* (IC).

Además, ISIWoK contiene otros recursos adicionales como el famoso *Journal Citation Report*, el cual cada año recopila métricas de la calidad de las revistas indexadas para de este modo, ofrecer un modo sistemático, objetivo y crítico de evaluar las revistas.

Scopus, recientemente SciVerse Scopus, es una base de datos de resúmenes y citaciones de artículos académicos procedentes tanto de revistas como de conferencias, desarrollada por la editorial Elsevier (Países Bajos). Contiene sobre unos 18000 títulos de más de 5000 editoriales, por lo que ofrece a los investigadores un recurso rápido, fácil y exhausto para abastecer las necesidades de áreas científicas, técnicas, de ciencias de la salud, de ciencias sociales, y de arte y humanidades. Además, Scopus contiene la herramienta *Journal Analyzer* la cual permite evaluar y comparar el rendimiento de las revistas indexadas.

Google Scholar es un motor de búsqueda desarrollado por Google para indexar el texto completo de la literatura académica accesible desde Internet. Proporciona una forma sencilla de buscar bibliografía especializada de muchas disciplinas: artículos, tesis, libros, resúmenes, etc. de editoriales académicas, sociedades profesionales, depósitos en línea, etc. Quizás, un inconveniente de esta base de datos, es que la descarga de la información no es una tarea trivial.

Es necesario comentar que ISIWoK, Scopus y Google Scholar no contienen indexadas las mismas revistas, por lo que su cobertura de los distintos campos científicos puede variar considerablemente. A consecuencia de esto, se han realizado diferentes estudios para comprobar y detectar la magnitud de este hecho [7, 36, 64].

Por otro lado, aunque las bases de datos anteriores son las más utilizadas y las más importantes, existen otras fuentes bibliográficas desde las que se puede recopilar información científica: arXiv[8] , CiteSeerX[9] , Digital Bibliography & Library Project[10] (DBPL), SAO/NASA Astrophysics Data System[11] (ADS), Science Direct[12] , etc.

Finalmente, el análisis de mapas científicos puede realizarse utilizando patentes o datos de financiación (proyectos de investigación, ayudas en I+D+i, etc.). Los datos de patentes pueden ser recopilados a través de las numerosas bases de datos de patentes que existen para tal efecto. Por ejemplo, se puede utilizar la Oficina de Patentes y Marcas Española[13] , la Oficina de Patentes Europea [14] , o la Oficina de Patentes y Marcas Estadounidense[15] . Por otro lado, los datos de financiación pueden descargarse, por ejemplo, a través de la *Fundación Nacional para las Ciencias* estadounidense[16] , aunque cabe destacar que estos datos de financiación no son de fácil localización.

1.2.2. Unidades de Análisis y Tipo de Redes Bibliométricas

El análisis de mapas científicos se realiza utilizando parte de la información contenida en los documentos recopilados de las bases de datos bibliográficas. Los diferentes tipos

[8] http://arxiv.org
[9] http://citeseerx.ist.psu.edu/
[10] http://dblp.uni-trier.de/
[11] http://adswww.harvard.edu/
[12] http://www.sciencedirect.com/
[13] http://www.oepm.es
[14] http://www.epo.org
[15] http://www.uspto.gov
[16] http://www.nsf.gov

de información que se pueden utilizar se conocen como unidades de análisis.

Las unidades de análisis más comunes en el análisis de mapas científicos son revistas, documentos, referencias citadas en los documentos, autores (incluyendo sus afiliaciones), así como términos a palabras descriptivas [12]. Los términos pueden seleccionarse del título del documento, de su resumen, del cuerpo completo de éste o utilizando cualquier combinación de los anteriores. Además, es posible utilizar las palabras clave añadidas por los autores del documento, así como aquellas añadidas por las bases de datos bibliográficas, como por ejemplo, los *ISI Keywords Plus* de ISIWoS.

Dependiendo de la unidad de análisis seleccionada para construir el mapa, se pueden estudiar diferentes aspectos de un campo científico. Por ejemplo, a través de los autores se puede analizar los aspectos sociales del área, utilizando los términos o palabras se puede estudiar los aspectos conceptuales, y mediante las referencias citadas se pueden observar los aspectos intelectuales.

Entre las unidades de análisis pueden establecerse diferentes tipos de relaciones que nos permitirán crear distintas redes bibliométricas. Fundamentalmente, tres son los tipos de relaciones entre unidades de análisis: co-ocurrencia, emparejamiento y enlace directo. La relación de co-ocurrencia se da entre dos elementos que aparecen conjuntamente en un documento, es decir, si el elemento i y j aparecen en el mismo documento, diremos que existe una relación de co-ocurrencia entre ellos. Además, esta relación puede cuantificarse, de modo que la relación represente el número de documentos en los que los mencionados elementos aparecen conjuntamente. La relación de emparejamiento se establece entre documentos, midiendo el grado de similitud de dos documentos en base a la unidad de análisis seleccionada. Por ejemplo, el emparejamiento entre los documentos d_1 y d_2 medirá el número de elementos que ambos tienen en común. Por último, la relación de enlace directo se establece normalmente entre documentos y las

referencias citadas en ellos, de modo que se existe un enlace directo entre un documento d_1 y todas sus referencias. Dado que las referencias representan documentos, a su vez una referencia puede tener un enlace directo con otro documento/referencia. De un modo similar, como veremos más adelante en esta sección, también podría establecerse redes de enlace directo entre otras unidades de análisis.

Además, las unidades de análisis pueden agregarse para formar una unidad de mayor nivel. En este caso, las relaciones se establecerían entre estas unidades superiores. Por ejemplo, en el caso del emparejamiento los documentos pueden agregarse por autores o revistas, estableciendo en este caso la relación entre autores o revistas. Por ejemplo, la relación de emparejamiento entre los autores a_1 y a_2 mostraría el número de elementos que ambos autores tienen en común en los documentos escritos por ellos.

Las relaciones entre las unidades de análisis pueden representarse como un grafo o una red, donde las unidades suelen ser los nodos y las relaciones entre ellas se representan mediante enlaces entre los nodos. Particularmente, en la relación de co-ocurrencia los nodos son las unidades de análisis, mientras que en el emparejamiento, los nodos son los documentos. Un caso especial es la relación de emparejamiento agregada, donde los nodos serían la unidad de nivel superior.

La Tabla 1.1 muestra una taxonomía de las redes bibliométricas más comunes en la literatura, de acuerdo a la unidad de análisis usada y al tipo de relación establecida entre éstas.

Técnica bibliométrica		Unidad de análisis	Tipo de relación
Emparejamiento bibliográfico	Autor	Obra del autor	Referencias comunes entre las obras de los autores
	Documento	Documento	Referencias comunes entre documentos
	Revista	Obra de la revista	Referencias comunes entre las obras de las revistas
Co-autor	Autor	Nombre del autor	Co-ocurrencia de autores
	País	País (extraído de la afiliación)	Co-ocurrencia de países
	Institución	Institución (extraído de la afiliación)	Co-ocurrencia de instituciones
Co-citación	Autor	Autor (extraído de la referencia)	Autores co-citados
	Documento	Referencia	Documentos co-citados
	Revista	Revista (extraído de la referencia)	Revistas co-citadas
Co-palabras		Palabra clave, o término extraído del título, resumen o cuerpo del documento	Co-ocurrencia de términos

Tabla 1.1: Taxonomía de tipos de redes bibliométricas.

Como mencionamos anteriormente, dependiendo del tipo de unidad de análisis elegida, se pueden observar diversos aspectos de un campo científico. De hecho, establecer diferentes relaciones entre la misma unidad puede descubrirnos varios lados del mismo aspecto. Por ejemplo, utilizando como unidad de análisis los autores (redes de co-autores, ACAA por sus siglas en inglés), se puede analizar la estructura social del campo científico analizado [37, 74]. De un modo similar, utilizando las afiliaciones de los autores se puede analizar la dimensión internacional del campo, por ejemplo, estableciendo relaciones de co-ocurrencia entre las instituciones de las afiliaciones (co-institución, ICAA), entre universidades (co-universidad, UCAA), o entre países (co-país, CCAA). Por otro lado, los términos o palabras suelen usarse mediante redes de co-palabras (CWA) para analizar la estructura conceptual de un campo científico, estudiando de este modo los principales temas de investigación tratados en él [19]. Por último, las referencias se utilizan para estudiar la estructura intelectual del campo, y normalmente se suelen establecer dos tipos de relaciones: co-ocurrencia (co-citación, DCA) y emparejamiento (emparejamiento bibliográfico, DBCA). La diferencia entre ambas es que el emparejamiento bibliográfico [50] es una relación fija y permanente debido a que se establece teniendo en cuenta las referencias contenidas en los documentos emparejados, mientras que la relación de co-citación [93] puede variar a lo largo del tiempo [47].

Las redes de emparejamiento bibliográfico pueden agregarse utilizando revistas y autores. Particularmente, el emparejamiento bibliográfico de autores (ABCA) [117] trata de descubrir las relaciones entre los autores que citan a la misma base intelectual (mismas referencias), mientras que el emparejamiento bibliográfico de revistas (JBCA) [38, 99] intenta descubrir las relaciones entre las revistas que citan a las mismas referencias.

De un modo similar, las redes de co-citación pueden extenderse utilizando revistas

y autores. Las referencias contienen diferentes trozos de información, como pueden ser los autores y la revista, utilizando dichos trozos se pueden construir tipos específicos de redes bibliométricas de co-citación. De este modo, la co-citación de autores (ACA) [113] muestra los autores que son citados frecuentemente de forma conjunta, mientras que la co-citación de revistas (JCA) [63] nos descubre aquellas revistas que suelen citarse conjuntamente. Además, las revistas pueden agregarse en un supra nivel, por ejemplo, basándonos en las categorías científicas (agrupación por campos) de áreas de las fuentes de información bibliográfica. Este tipo de redes, utilizando las categorías creadas por ISIWoS fue realizado en [69].

Finalmente, atendiendo a las relaciones por enlace directo, se pueden crear redes del tipo documento/documento (un documento cita a otro documento), autor-autor (un autor cita a otro autor), revista-revista (una revista cita a otra revista). Además, las redes de enlace directo pueden crearse utilizando unidades de análisis heterogéneas, como por ejemplo, autor-documento (un autor, consume/lee un documento).

1.2.3. Preprocesamiento

Los datos recopilados de las bases de datos bibliográficas normalmente contienen errores. Por ejemplo, los nombres de ciertos autores (especialmente, los españoles y portugueses) pueden estar mal escritos, o contener faltas ortográficas, lo mismo puede ocurrir con el nombre de la revista, o incluso que alguna referencia sea errónea. Además, algunas veces es necesario añadir o completar la información, por ejemplo, la afiliación de un autor puede estar incompleta, al igual que puede ocurrir con una referencia. Por esta razón, el análisis de mapas científicos no puede aplicarse directamente a los datos recogidos en bruto de las bases de datos bibliográficas. Es decir, es necesario un proceso de preprocesamiento para limpiar los datos recopilados y de este modo mejorar

su calidad. De hecho, el preprocesamiento es quizás uno de los pasos más importantes en el análisis de mapas científicos, ya que mejorando la calidad de las unidades de análisis, la calidad de los resultados obtenidos al final del análisis de mapas científicos se verá incrementada.

Existen diferentes tipos de preprocesamiento, y cada uno de ellos nos ayudan a limpiar la información y seleccionar aquella con la que queremos realizar el proceso de análisis.

- Unificación de elementos duplicados, similares o escritos erróneamente. Algunas veces, existen elementos que representan al mismo objeto o concepto, pero estos se encuentran escritos de forma diferente. Por ejemplo, el nombre de un autor puede escribirse de diversas formas (por ejemplo, Garfield, E.; Eugene Garfield), representando ambas al mismo autor. En otras ocasiones, un mismo concepto puede escribirse con diversas palabras, e incluso puede representarse con diversos acrónimos (por ejemplo, TSK ⇒ Takagi-Sugeno-Kang, o GDM ⇒ Group Decision Making). La detección de elementos duplicados o similares puede ayudar a solucionar esos errores.

- La división temporal es útil para partir o dividir los datos en diferentes periodos de tiempo, para así poder analizar la evolución temporal del campo científico bajo estudio. Este tipo de preprocesamiento es necesario sólo si el análisis de mapas científicos se está realizando bajo un marco longitudinal [39, 81].

- Reducción de datos para seleccionar la información más importante o relevante. Normalmente, disponemos de una gran cantidad de información, lo que dificulta la obtención de unos resultados adecuados y precisos mediante el análisis de mapas científicos. Por esta razón, el análisis se suele realizar con una porción de los

datos, que puede ser, por ejemplo, los autores más productivos, los documentos más citados, las revistas con mejores métricas, las palabras más frecuentes, etc.

- Reducción de datos en redes. Las redes bibliométricas también pueden filtrarse de un modo similar a la reducción de datos. La reducción de redes se emplea para seleccionar los enlaces más significativos de una red bibliométrica de acuerdo a diversas medidas. Por ejemplo, eliminado los enlaces con menor peso. Además, en la reducción de redes, también se pueden eliminar aquellos nodos que están aislados, es decir, aquellos que no están o que están marginalmente enlazados con el resto de la red.

Tanto la reducción de datos como la reducción de la red pueden entenderse como reducciones similares. Si suponemos una red construida sobre un conjunto de información sin filtrar, la reducción de datos afectaría a los nodos de dicha red, eliminado por ejemplo, aquellos nodos por debajo de una frecuencia determinada. Por otro lado, la reducción de red afectaría a los enlaces entre los nodos, utilizando medidas basadas en esos enlaces.

1.2.4. Normalización de las Redes Bibliométricas

Una vez que se ha construido la red bibliométrica estableciendo relaciones entre las unidades de análisis, es necesario aplicar una transformación sobre su grafo para obtener las similitudes entre sus nodos. Es decir, necesitamos normalizar la red bibliométrica [106]. La normalización nos permitirá relativizar las relaciones entre dos unidades de análisis, dando, por ejemplo, mayor importancia a aquellas unidades con una frecuencia baja pero con una gran frecuencia de co-aparición, frente a aquellas con frecuencia alta y una baja frecuencia de co-aparición.

En la literatura se han usado diversas medidas de similaridad para normalizar redes bibliométricas, de entre estas, las más populares o usadas con mayor frecuencia son: *Coseno de Salton* ($c_{ij}/\sqrt{e_i e_j}$)[17] [86], *Índice de Jaccard* ($c_{ij}/e_i + e_j - c_{ij}$) [75], *Índice de equivalencia* ($c_{ij}^2/e_i e_j$) [20], y *Fuerza de asociación* ($c_{ij}/e_i e_j$) [29, 105], también conocida como *Índice de proximidad* [75, 84], o *Índice de afinidad probabilística* [119].

Además, es común el uso de medidas de normalización para medir el peso de los términos en el conjunto de documentos analizados. Por ejemplo, si sobre una red de co-citación aplicamos un algoritmo de clustering o agrupamiento, como veremos más adelante, para construir el mapa, sería adecuado establecer una etiqueta a cada uno de los clusters o grupos detectados. Esta etiqueta podría ser el término con mayor importancia de entre los documentos asociados al cluster. La normalización de términos establece un peso a cada término equivalente a su importancia en todo el corpus (conjunto de documentos analizados). En este sentido, se pueden aplicar diferentes medidas de normalización: [3, 25, 86]: $tf \cdot idf$, *latent semantic analysis*, *log-likelihood ratio tests*, *log entropy*, *mutual information*, etc.

1.2.5. Creación del Mapa Científico

El proceso de mapeado o de creación del mapa científico, es sin duda el más importante. En sí mismo, es responsable de construir el mapa, aplicando diversos algoritmos sobre la red bibliométrica global formada mediante las relaciones entre las unidades de análisis.

La creación del mapa científico se suele realizar fundamentalmente mediante técnicas de reducción de la dimensionalidad y aplicando algoritmos de clustering [12].

Las técnicas de reducción de la dimensionalidad son utilizadas para transformar

[17] Donde, c_{ij} denota la co-ocurrencia de los elementos i y j, y e_i denota la frecuencia del elemento i.

las redes bibliométricas en espacios de baja dimensión, normalmente, espacios de dos dimensiones, de modo que sean más fácilmente entendibles y por tanto se más fácil su comprensión y análisis. Entre estás técnicas destacan la descomposición en autovalores y autovectores [32], análisis de la componente principal [102], escalado multidimensional [55], mapas auto-organizativos [49, 52, 76]. Un tipo especial de técnica de reducción de la dimensionalidad son las redes Pathfinder (PFNETs), las cuales se emplean para identificar la columna vertebral de la red bibliométrica [82, 87].

Las técnicas de clustering dividen un conjunto de elementos en diversos subconjuntos, los cuales deben cumplir la condición de tener una gran cohesión interna. Es decir, los elementos dentro de un mismo grupo deben tener una gran similitud entre sí, mientras que, por otro lado, deben ser bastante diferentes del resto de elementos que no están en el grupo. Los algoritmos de clustering aplicados a redes bibliométricas intentan descubrir las subredes que forman la red bibliométrica global, es decir, aquellos conjuntos de nodos que están fuertemente enlazados entre sí, pero pobremente enlazados con el resto de la red. Dentro de los algoritmos de clustering, existen algunos clásicos que se han solido aplicar para la creación de mapas científicos, como el algoritmo de los Centros Simples [29], o el algoritmo de Enlace simple, [100]. Además, recientemente, algunos autores han propuesto nuevos algoritmos de clustering para crear el mapa científico: *Streemer* [48], *spectral clustering* [25], *modularity maximization* [26] y un algoritmo de clustering significativo mediante *bootstrap resampling* [85], entre otros.

Por otro lado, para la creación del mapa científico se pueden emplear técnicas genéricas de minería de datos, específicamente de minería de grafos (*Graph mining*) [28, 91], o técnicas procedentes del Análisis de Redes Sociales (*Social Network Analysis*) [22, 112].

Lógicamente, la información obtenida y el tipo de mapa variará considerablemente dependiendo del tipo de técnica empleada para generar el mapa científico.

1.2.6. Métodos de Análisis

Una vez que se ha generado el mapa, se pueden realizar una serie de análisis para extraer conocimiento de él. Es decir, analizando el mapa con distintos métodos podremos observar diversas vistas de la información.

El *análisis de redes* [22, 28, 91, 112] nos permite realizar diferentes análisis estadísticos sobre los mapas generados, o incluso sobre la red bibliométrica global. Por ejemplo, se pueden establecer diversas medidas de red, como el número total de nodos, el numero de nodos aislados, el grado medio de la red, el número de componentes débilmente conectados, la densidad de la red, etc. Por otro lado, si se aplicó un algoritmo de clustering para construir el mapa, se puede medir la densidad y centralidad de los diversos clusters [20], o incluso establecer otras medidas a partir de las relaciones entre clusters. Además, se puede medir el solapamiento entre los clusters detectados utilizando medidas como en índice de Jaccard [75].

El *análisis temporal* nos permite identificar los patrones, tendencias, así como estacionalidades en las estructuras de las redes bibliométricas, en especial, en los mapas generados a partir de ellas. En otras palabras, el análisis temporal trata de analizar la evolución estructural de los diversos aspectos de un campo científico a través de diferentes periodos tiempo. Este tipo de análisis se suele realizar bajo un marco longitudinal [39, 81].

Un tipo especial de análisis temporal es la *detección de estallidos*, la cual trata de encontrar aquellas características que han tenido una fuerte intensidad en un instante de tiempo determinado. Por ejemplo, en [51] se propone un algoritmo específico para realizar este tipo de análisis.

Otro tipo de análisis es el *geoespacial* [9, 56, 98], el cual trata de responder a la pregunta de dónde suceden los acontecimientos y qué impacto provocan en sus áreas

vecinas. Para realizar un análisis geoespacial necesitamos elementos con información espacial, o elementos que puedan ser geolocalizados (como la universidad de los autores, o su dirección postal), por este motivo, este tipo de análisis se suele nutrir con los datos de afiliación de los autores.

Finalmente, tanto a la red bibliométrica global, como al mapa científico generado puede aplicársele un análisis basado en medidas de rendimiento, utilizando indicadores bibliométricos de producción, actividad, calidad e impacto como los vistos en la Sección 1.1, como h-index [1, 46], g-index [34], hg-index [2] o q^2-index [17]. Como sabemos, las redes bibliométricas están formadas por unidades de análisis; estas a su vez proceden de unos documentos determinados, es decir, pueden asociarse con un conjunto de documentos. De esto modo, los nodos de una red pueden enriquecerse contabilizando el número de documentos asociados a ellos. Además, si el mapa generado está basado en clusters, a cada uno de ellos se les puede asignar un conjunto de documentos (por ejemplo, aquellos que estén asociados con al menos un nodo de la red). Asimismo, es posible utilizar medidas de calidad basadas en el número de citas recibidas por los documentos, para de este modo determinar el impacto de los diferentes elementos del mapa y así detectar aquellos con una mayor visibilidad.

1.2.7. Técnicas de Visualización

Como se mostró en la sección anterior, la salida generada por cada tipo de análisis es diferente, por lo que necesitaremos emplear diversas técnicas de visualización para mostrar los resultados obtenidos a través de los mapas científicos. La técnica de visualización empleada es fundamental ya que, nos permitirá comprender y interpretar de forma adecuada los resultados.

Las redes bibliométricas, así como las posibles subredes detectadas en el proceso

de creación del mapa científico (Sección 1.2.5) pueden representarse utilizando *mapas heliocéntricos* [68], *modelos geométricos* [92] o redes temáticas [5] (detalladas en la Sección 3.1.2). Otro tipo de técnica de visualización para representar redes se encarga de posicionar los nodos en un espacio bi-dimensional, en donde la distancia entre dos nodos refleja el peso de la relación entre ambos [32, 35, 76, 107]. En este tipo de visualización, una menor distancia entre nodos implica una relación más importante [107].

Asimismo, si el mapa científico se construyó aplicando un algoritmo de clustering, las subredes detectadas pueden categorizarse en un diagrama estratégico, el cual, como veremos en la Sección 3.1.2, es un cartesiano en donde las subredes detectadas se posicionan de acuerdo a diversas medidas de redes, como puede ser la densidad y centralidad [20].

La evolución estructural de los mapas científicos a lo largo de periodos de tiempo (análisis temporal), puede visualizarse utilizando diversas técnicas: *Cluster string* [97, 101, 104], *rolling clustering* [48], *alluvial diagrams* [85], *ThemeRiver visualization* [41], así como *thematic areas* (Sección 3.1.3). Por otro lado, otros autores proponen posicionar la red de un periodo de tiempo teniendo en cuenta las redes de los periodos anteriores y posteriores [58], o incluso, resumiendo todos los cambios estructurales de los diversos periodos de tiempo en una única red [23, 25].

Por último, los resultados del análisis geoespacial suelen visualizarse sobre un mapa terrestre. Este tipo de mapas que aúnan información geográfica con otro tipo de información, son conocidos como mapas temáticos. Por ejemplo, si el análisis de mapas científicos se realizó empleando una red de co-autores, y se aplicó un algoritmo de clustering para detectar las diversas subredes de co-autores que suelen trabajar conjuntamente, dichas subredes pueden representarse sobre un mapa terrestre en donde cada

nodo se posiciona sobre el país al que pertenece el autor (extraido de su afiliación).

1.2.8. Interpretación

Una vez que el análisis de mapas científicos ha terminado, el analista tiene que interpretar los mapas y resultados obtenidos usando su propia experiencia y conocimiento, y a veces apoyándose en otros expertos, para así poder obtener conclusiones adecuadas sobre el campo científico analizado.

En la interpretación, el analista trata de descubrir, así como de extraer conocimiento útil y previamente desconocido que le ayude a tomar decisiones sobre, por ejemplo, que políticas científicas implementar, dónde invertir dándole mayores recursos, cuáles son los frentes de investigación, cuáles son los temas calientes, etc.

Capítulo 2

Herramientas para la Realización de Análisis de Mapas Científicos: Estado del Arte

Existe una gran cantidad de herramientas informáticas con las que se puede realizar análisis de mapas científicos. Algunas de ellas son genéricas, es decir, no fueron diseñadas específicamente para esa tarea. Entre ellas, podemos encontrar herramientas del ámbito de la bibliometría, como *Publish or Perish* [40], del campo de las redes sociales, como *Pajek* [8] o *UCINET* [11], o del campo de la bioinformática, como Cytoscape [89].

Además, podemos encontrar herramientas que han sido específicamente diseñadas para realizar análisis de mapas científicos. En este capítulo nos centramos en estas últimas, realizando una descripción y un análisis comparativo profundo de las herramientas más representativas, para de este modo, determinar las ventajas, inconvenientes y diferencias más destacadas de cada una de ellas. Particularmente, analizaremos las siguiente nueve herramientas: Bibexcel [73], CiteSpace II [23, 24], CoPalRed [5, 6], INSPIRE [114], Loet Leydesdorff's Software, Network Workbench Tool [13, 44], Science of Science (Sci^2) Tool [88], VantagePoint [77] y VOSViewer [107].

De este modo, el presente Capítulo se estructura de la siguiente forma: En la Sección

2.1 se describe cada una de las herramientas, dando detalles acerca de su funcionamiento, así como los métodos y técnicas que incorporan. En la Sección 2.2 se realiza una comparativa de las nueva herramientas basándose en cinco aspectos fundamentales del análisis de mapas científicos. En la Sección 2.3 se lleva a cabo un análisis de mapas científicos concreto utilizando diversas herramientas, para de este modo estudiar las características diferentes de los resultados obtenidos. Por último, en la Sección 2.4 se detallarán algunas conclusiones y lecciones aprendidas sobre las herramientas.

2.1. Descripción y Análisis de las Herramientas

En la actualidad se dispone de diversas herramientas informáticas diseñadas para la realización de mapas científicos, siendo las más representativas Bibexcel [73], CiteSpace II [23, 24], CoPalRed [5, 6], IN-SPIRE [114], Loet Leydesdorff's Software, Network Workbench Tool [13, 44], Science of Science (Sci^2) Tool [88], VantagePoint [77] y VOSViewer [107].

En esta sección se pretende realizar una descripción y análisis de cada una de las herramientas.

2.1.1. Bibexcel

Bibexcel[1] [73] es una herramienta bibliométrica desarrollada en la Universidad de Umeå (Suecia). Está diseñada para gestionar datos bibliométricos y construir con ellos mapas que pueden ser leídos por herramientas como Microsoft Excel, UCINET, Pajek, etc. Bibexcel puede ser descargada y utilizada de forma gratuita.

Bibexcel puede leer datos descargados desde diferentes fuentes de información bibliográficas, como ISIWoS y Scopus. Además, es capaz de leer el formato de exportación

[1] http://www.umu.se/inforsk/Bibexcel/

del gestor de referencias Procite.

Esta herramienta permite aplicar diferentes métodos de preprocesamiento sobre datos textuales. Por ejemplo, permite aplicar un lematizador (únicamente para Inglés), eliminar los documentos duplicados, etc. Además, Bibexcel posibilita la reducción de datos basándose en la frecuencia de los elementos (autores, términos, etc.), dejando los más frecuentes, así como la reducción de los enlaces de la red basándose en su peso (dejando los más fuertes).

Bibexcel es capaz de extraer diferentes redes bibliométricas, siendo las principales: co-citación, emparejamiento bibliográfico, co-autor, co-palabras. Además, permite extraer diferentes matrices de co-ocurrencia usando cualquier campo del documento o cualquier combinación de ellos. Las matrices o redes pueden ser normalizadas utilizando tres medidas de similitud diferentes: el coseno de Salton [86], el índice de Jaccard [75] o la medida de Vladutz y Cook.

Sobre los datos normalizados, el usuario puede aplicar un algoritmo de clustering o preparar una matriz para realizar un análisis de escalado multidimensional -MDS- (es necesario utilizar una herramienta externa).

Bibexcel no dispone de un módulo para la visualización, pero si permite exportar los resultados para de este modo poder realizar la visualización a través de herramientas, como Pajek, UCINET o SPSS. Además, las redes bibliométricas también pueden ser exportadas.

2.1.2. CiteSpace II

CiteSpace[2] [23, 24] ha sido desarollado en la Universidad de Drexel (EE.UU.). Es una herramienta para la detección, análisis y visualización de patrones y tendencias

[2] http://cluster.cis.drexel.edu/~cchen/citespace

en la literatura científica, siendo su principal objetivo facilitar el análisis de tendencias emergentes en dominios de conocimiento o campos científicos.

CiteSpace es capaz de leer diferentes formatos de fuentes bibliográficas, como ISI-WoS, PubMed, arXiv y SAO/NASA Astrophysics Data System (ADS). Además, puede leer datos de subvenciones/financiación (NSF Awards), y datos de patentes (Derwent Innovations Index).

Esta herramienta permite construir una gran variedad de redes bibliométricas: co-instituciones, co-países, co-subvenciones[3] , co-categorías, co-palabras, co-citación de documentos, co-citación de autores, co-citación de revistas y emparejamiento bibliográfico. Las redes, o grafos, pueden construirse para diferentes periodos de tiempo, para de este modo, analizar la evolución del dominio científico estudiado. Además, el analista puede filtrar los elementos que formarán la red, seleccionando así los elementos más importantes. La red puede ser normalizada utilizando el coseno de Salton, el índice de Dice o el índice de Jaccard.

Una vez que la red ha sido creada y normalizada, CiteSpace nos permite visualizarla y realizar diferentes análisis sobre ella. Además, CiteSpace permite al analista realizar un clustering espectral (Spectral Clustering, en la literatura anglosajona) y una detección de estallidos (Burst Detection, en la literatura anglosajona) utilizando las citas. Asimismo, CiteSpace posee distintos modos de visualización [24], como la vista de grupos, o la vista temporal.

Si el proceso de clustering detectó grupos, CiteSpace puede asignar etiquetas a cada uno de ellos. Para ello, elige el término más importante de entre los títulos, resúmenes y palabras clave de los documentos del grupo. La importancia de los términos puede establecerse a través de diversas medidas [25]: $tf \cdot idf$, *log-likelihood ratio tests*, o *mutual*

[3] Mediante el análisis del nombre de los patrocinadores de las subvenciones que aparecen en el campo *financiación* de los documentos.

information.

2.1.3. CoPalRed

CoPalRed[4] [5, 6] es una herramienta comercial desarrollada por el grupo de investigación EC^3 de la Universidad de Granada (España). Ha sido específicamente creada para realizar mapas científicos basados en redes de co-palabras, utilizando como unidades de análisis las palabras clave de los documentos científicos. CoPalRed ha sido descrito como un "*Sistema de Conocimiento*", que recoge la información contenida en las bases de datos bibliométicas y la transforma en *nuevo conocimiento* [4].

Esta herramienta sólo lee documentos en formato CSV generados a través del gestor de referencias Procite.

Uno de los puntos fuertes de CoPalRed es su módulo para realizar la unificación de elementos. Gracias a este módulo, el usuario puede unificar elementos que representen al mismo concepto. Posteriomente, una vez que las palabras clave han sido unificadas, CoPalRed construye una matriz de co-ocurrencia y la normaliza utilizando el índice de equivalencia propuesto por [20].

CoPalRed puede realizar tres tipos de análisis: estructural, estratégico y dinámico.

- **Análisis Estructural**: muestra el conocimiento en forma de redes temáticas, en donde se muestran las palabras clave y sus relaciones.

- **Análisis Estratégico**: sitúa cada red temática en una posición relativa dentro de la red global de acuerdo a sus valores de cohesión externa (centralidad) y cohesión interna (densidad).

- **Análisis Dinámico**: analiza las transformaciones de las redes temáticas a lo

[4] http://ec3.ugr.es/copalred/

largo del tiempo, permitiendo identificar unificaciones, bifurcaciones, apariciones o desapariciones de los temas detectados.

Finalmente, CoPalRed visualiza los resultados usando diagramas estratégicos, temas y redes temáticas [5, 6, 59, 60]. Cada tema tiene asignado una etiqueta que se corresponde con el nombre del nodo más central (palabra clave) de su red temática asociada. Además, cada tema se representa en un diagrama estratégico.

2.1.4. IN-SPIRE

IN-SPIRE[5] [114] es una herramienta comercial, que brinda al analista la posibilidad de descubrir relaciones, tendencias y temas ocultos bajo los datos objeto de estudio, para de este modo, poder obtener nuevo conocimiento y nuevas vistas de los datos. IN-SPIRE utiliza metáforas geográficas para ayudar al usuario a descubrir, de un modo sencillo, las relaciones entre documentos y conjuntos de documentos que son muy similares entre sí. Esta herramienta utiliza patrones estadísticos sobre las palabras, para caracterizar documentos basándose en su contexto [45]. IN-SPIRE surgió del proyecto *SPIRE*, el cual fue financiado por el *Departamento de Energía* y la *Agencia de Inteligencia* de Estados Unidos. El desarrollo de esta herramienta ha sido realizado en en *Laboratorio Nacional del Pacifico Noroeste*.

IN-SPIRE permite leer documentos no estructurados, o documentos estructurados como, HTML o XML. Incluso, ofrece la posibilidad de leer datos provenientes de documentos en formato Microsoft Excel, o CSV. La herramienta permite seleccionar los campos que el usuario desea utilizar para medir la similitud entre documentos. Además, permite seleccionar otros meta-campos tales como, el título del documento o su fecha asociada (normalmente, fecha de publicación).

[5] http://in-spire.pnl.gov

Al contrario que otras de las herramientas analizadas en este capítulo, IN-SPIRE no permite extraer redes bibliométricas a partir del campo seleccionado (unidad de análisis). Por contra, utiliza el campo o conjunto de campos para calcular la similitud entre documentos utilizando su propio motor textual [114]. De forma abreviada, dicho motor utiliza el modelo espacio-vectorial [86], representando cada documento como un vector de términos. De este modo, si las palabras clave han sido seleccionadas como campo a analizar, la similitud mostrará si dos documentos contienen palabras clave similares. Aunque IN-SPIRE es capaz de construir un mapa utilizando cualquier campo del documento, su motor funciona mejor con palabras (procedentes, por ejemplo, del título, resumen o cuerpo del documento), ya que necesita una gran cantidad de información para establecer correctamente las similitudes entre los documentos.

Una vez que se han medido las similitudes entre documentos, IN-SPIRE aplica un algoritmo de clustering llamadado *"Fast Divisive Clustering"* [114]. Al final del proceso, se generan diferentes temas (conjuntos de documentos). Cada tema tiene por nombre el conjunto de los términos más frecuentes (usando $tf \cdot idf$) de los documentos asociados a ellos.

IN-SPIRE contiene dos técnicas de visualización diferentes, las cuales son el estandarte de la herramienta: *Galaxies* y *ThemeScape*™. La técnica *Galaxies* emplea una metáfora de cielo estrellado, donde los documentos representan estrellas en un cielo nocturno. Por otro lado, la técnica de visualización *ThemeScape* se construye directamente a partir de la distribución de documentos en la vista *Galaxies*, representando los temas como capas de sedimentos que conjuntamente crean la apariencia de un paisaje natural. En la vista *ThemeScape*, la altura de los picos de las "montañas" se corresponde con la fuerza de los temas en esa localización; la extensión de los picos se corresponde al área y brillo de los temas en la vista *Galaxies*. En ambas vistas, la proximidad de dos

elementos (documentos) revela la similitud entre ellos.

Finalmente, IN-SPIRE contiene un conjunto de herramientas que ayudan al analista a descubrir el conocimiento intrínseco en el corpus de documentos estudiados:

- *Time slicer* nos permite visualizar como un tema concreto crece o decrece a lo largo del tiempo y como cambia la mezcla de temas a lo largo del tiempo en la vista *Galaxies*.
- La herramienta *Groups* define una colección de documentos en el corpus estudiado, ofreciendo la posibilidad de analizar su comportamiento y pertenencia a los distintos temas detectados.
- La herramienta *Facets* nos permite descubrir las relaciones entre los temas detectados, así como entre los grupos definidos por el usuario.
- Permite varios tipos de consulta, tanto búsqueda booleana, por proximidad de palabras o basada en frases y ejemplos.
- La herramienta *Correlation* nos permite descubrir relaciones entre grupos.

2.1.5. Loet Leydesdorff's Software

Loet Leydesdorff's software[6] es un conjunto de herramientas en línea de comandos que permiten realizar un análisis de mapas científicos a través de diferentes funciones de análisis. Fue desarrollada en la Universidad de Amsterdam (Países Bajos) y pueden usarse gratuitamente.

Estas herramientas permiten extraer diferentes redes bibliométricas: co-palabras, co-autor, emparejamiento bibliográfico agregado por autores y revistas, así como co-citación de autores. Además, se puede analizar la colaboración a nivel internacional,

[6] http://www.leydesdorff.net

institucional, o de ciudades. Las diferentes matrices o redes se normalizan utilizando el coseno de Salton.

Los resultados pueden visualizarse utilizando herramientas externas como, Pajek, UCINET, Network Workbench Tool (ver Sección 2.1.6) o Science of Science Tool (ver Sección 2.1.7). Asimismo, la visualización de las redes de colaboración puede realizase a través de Google Maps[7] .

De entre el conjunto de herramientas, algunas de las proporcionadas por Loet Leydesdorff's Software son específicas para organizar la información descargada de las fuentes de información (ISIWoS, Scopus, Google Scholar, Google) en una base de datos. Esta base de datos será el fichero de entrada de las herramientas restantes.

Una desventaja de estas herramientas, es que no permite el preprocesamiento de los datos, de modo que, por ejemplo, para realizar un análisis longitudinal, es necesario utilizar herramientas externas para dividir los datos en diferentes periodos.

2.1.6. Network Workbench Tool

Network Workbench Tool (NWB)[8] es una herramienta para el análisis, modelado y visualización de redes para investigadores en física, biomedicina y ciencias sociales [13, 44], desarrollada por el *Cyberinfrastructure for Network Science Center* de la Universidad de Indiana (Estados Unidos). La herramienta puede utilizarse de forma gratuita.

NWB contiene algoritmos específicos para tratar con datos de publicaciones científicas y construir redes y mapas bibliométricos. La herramienta es capaz de leer distintos formatos de datos bibliográficos, como ISIWoS, Scopus, Bibtex y el formato de expor-

[7] http://maps.google.com/
[8] http://nwb.slis.indiana.edu

tacion del gestor de referencias EndNote[9] . Además, permite leer datos de financiación provenientes de la *National Science Foundation* (NFS), así como otros datos académicos en formato CSV.

La herramienta permite aplicar diferentes preprocesamientos sobre los datos, construir distintas redes bibliométricas, realizar análisis de redes, análisis temporal, y finalmente visualizar los resultados.

- El preprocesamiento de los datos se realiza a través de la eliminación de documentos duplicados, división de los datos en diferentes periodos de tiempo y detección y unificación de elementos (por ejemplo, elementos que representan al mismo autor en un análisis de co-autoría, o términos que representan al mismo concepto en un análisis de co-palabras).

- NWB permite construir diferentes tipos de redes: co-citación de documentos, co-autor, co-palabras y emparejamiento bibliográfico de documentos. Además, la herramienta puede construir redes basadas en enlace directo; por ejemplo, puede construir redes autor-documento, o redes de citación directa.

- Para la realización del mapa científico y analizar las redes obtenidas, NWB permite utilizar diferentes algoritmos. Además, la herramienta permite realizar una detección de estallidos para identificar los incrementos en la frecuencia de los elementos analizados.

- La visualización de las redes generadas se realiza a través de componentes o complementos como, GUESS[10] o Jung[11] . Además, se pueden aplicar diferentes algoritmos de posicionamiento o dibujado de redes como, el algoritmo DrL, el

[9] http://www.endnote.com/
[10] http://graphexploration.cond.org/
[11] http://jung.sourceforge.net/

cual es el sucesor de código libre de VxOrd [33], que fue usado por la herramienta VxInsight [14, 32].

2.1.7. Science of Science Tool

Science of Science Tool (Sci^2) [12] es una herramienta modular para realizar estudios de la ciencia. Permite realizar análisis temporal, geoespacial, textual y de redes a niveles micro (individual), meso (local) y macro (global). Al igual que NWB (Sección 2.1.6), ha sido desarrollada por el *Cyberinfrastructure for Network Science Center* de la Universidad de Indiana, y de igual modo, es gratuita.

Sci^2 está centrada en estudios científicos y por tanto tiene algoritmos específicos para tratar con esta clases de análisis. La principal virtud de Sci^2 podría ser la incorporación de módulos para tratar con datos bibliométricos, pudiendo así prepararlos para el análisis posterior.

Al igual que NWB, Sci^2 es capaz de leer diferentes formatos de datos bibliográficos: ISIWoS, Scopus, Bibtex y el formato de exportación del gestor de referencias EndNote. Además, permite leer datos de financiación provenientes de la *National Science Foundation* (NFS), así como otros datos académicos en formato CSV.

Sci^2 permite preparar y preprocesar los datos, extrayendo diferentes tipos de redes, realizando distintos análisis (temporal, geoespacial, textual y de redes) y finalmente, visualizar los resultados a través de complementos y algoritmos de posicionamiento de redes (Sci^2 incluye el algoritmo de posicionamiento DrL).

Los preparación de los datos de Sci^2 se encarga de limpiar la información bibliográfica y crear diferentes redes y tablas que pueden usarse en las fases de preprocesamiento, análisis y visualización. Principalmente, Sci^2 permite extraer las siguientes redes

[12] http://sci.slis.indiana.edu

bibliométricas: co-autor, co-IP (Investigador Principal), co-citación de documentos, co-citación de revistas, co-citación de autores, emparejamiento bibliográfico, emparejamiento bibliográfico agregado por autores y emparejamiento bibliográfico agregado por revistas. Asimismo, la herramienta permite construir redes de enlace directo como, redes de autor-referencias, documento-referencias, revistas-referencias, y finalmente, autor-documentos.

Por último, Sci^2 contiene distintos algoritmos para realizar los mapas y aplicar análisis sobre ellos. La creación del mapa puede realizarse utilizando algoritmos de detección de comunidades o de identificación de la columna o base principal de la red (*backbone identification*). El análisis temporal se realiza dividiendo los datos en diferentes periodos de tiempo, o a través de una detección de estallidos. El análisis geoespacial se puede realizar a través de geocodificación y de mapas temáticos geográficos. El análisis textual se realiza aplicando una detección de estallidos sobre las palabras y a través de un análisis de co-palabras. Finalmente, el análisis de redes se realiza a través de medidas y algoritmos estadísticos.

2.1.8. VantagePoint

VantagePoint[13] [77] es una potente herramienta comercial de minería de textos que permite descubrir conocimiento a partir de documentos científicos, así como de patentes. Permite analizar grandes volúmenes de información estructurada, y descubrir patrones y relaciones, así como direccionar rápidamente el *dónde*, *cómo*, *cuándo* y *quién*. VantagePoint ha sido diseñado por la empresa *Search Technology Inc.* (Estados Unidos). Esta herramienta ha sido usada para realizar una gran cantidad de análisis de mapas científicos [66, 78, 79, 90].

[13] http://www.thevantagepoint.com/

VantagePoint contiene más de 180 filtros (por ejemplo, para ISIWoS, Scopus, la Oficina de patentes Americana y Europea, etc.) para importar información en diferentes formatos, lo que posibilita poder trabajar con la mayoría de las bases de datos bibliográficas y de patentes. Además, existen filtros para importar información contenida en ficheros de Microsoft Excel y Access, o en formato XML[14] . Incluso permite la creación de filtros definidos por el usuario.

Una vez que el corpus está cargado en la herramienta, VantagePoint muestra los diferentes campos incluidos en los documentos. Por ejemplo, si el corpus contiene información bibliográfica, los campos podrían ser el título, los autores, las afiliaciones, el resumen y las referencias de los documentos.

La interfaz gráfica de VantagePoint contiene tres partes bien diferenciadas: el espacio de trabajo, la vista *Título* y la vista de Detalle. El espacio de trabajo muestra todas las vistas de listas, matrices y mapas generadas por el usuario. La vista *Título* muestra el título de los registros en el corpus asociados con los elementos seleccionados. Finalmente, la vista de Detalle muestra la co-ocurrencia de los elementos seleccionados en un campo concreto, pudiendo ser seleccionado cualquier campo, a través de listas y diagramas.

La herramienta permite crear diferentes listas a partir de cualquier campo. Estas listas muestran todos los elementos en el corpus del campo escogido. La vista lista muestra el número de documentos y el número de instancias (número de veces que un elemento aparece en el corpus, teniendo en cuenta los elementos duplicados en el registro) para cada uno de los elementos contenidos en ella. Los elementos de una lista pueden asignarse a distintos grupos. Dichos grupos son útiles para definir una porción del corpus de modo que se pueda reducir la cantidad de datos usados en el posterior análisis. Por ejemplo, se puede construir un grupo que contenga los 30 autores más productivos. Además, un mismo elemento puede estar asociado con más de un grupo

[14] Existe un módulo capaz de generar un filtro a partir de la estructura de un fichero XML concreto.

simultáneamente.

Uno de los puntos fuertes de VantagePoint es su módulo de preprocesamiento y limpieza de datos. Una lista puede limpiarse y reducirse utilizando la función *Cleanup*, la cual trata de identificar los elementos que pueden ser equivalentes realizando búsquedas de proximidad difusas en un campo específico. Asimismo, una lista puede ser limpiada a través de la aplicación de un tesauro. Aunque VantagePoint contiene diversos tesauros predefinidos, el usuario puede definir sus propios tesauros o editar los existentes. Cualquier cambio realizado sobre una lista, generará una nueva lista, de modo que siempre guardaremos la información original.

VantagePoint nos permite construir diferentes clases de matrices, que muestran los registros en el corpus contenidos en dos listas específicas:

- Matriz de co-ocurrencia: muestra el número de registros en los que el elemento i (de la primera lista) y el elemento j (de la segunda lista) aparecen conjuntamente en el corpus.
- Matriz de Auto-correlación: muestra las correlaciones de los elementos de una lista.
- Matriz de correlación cruzada: muestra las correlaciones de los elementos de una lista basándose en los valores de otra lista.
- Matriz de factores: es el resultado de un Análisis de Componentes Principales. Esta matriz muestra los elementos como filas y los factores como columnas.

VantagePoint también permite construir diferentes redes bibliométricas, dependiendo del tipo de aspecto que queramos estudiar, las cuales pueden utilizarse en la creación del mapa. En concreto, VantegePoint es capaz de construir las siguientes matrices: co-autor (usando el nombre del autor, su afiliación o país), co-citación (usando las

referencias, los autores de la referencia, o la revista de la referencia), así como co-palabras (usando cualquier conjunto de términos). Incluso, permite construir matrices heterogéneas seleccionando dos listas diferentes. Por ejemplo, se podría construir una matriz de autores por año, para analizar la productividad de los autores. Las matrices pueden exportarse en ficheros de texto plano, o directamente puede copiarse una selección de la matriz y pegarse en Microsoft Excel.

VantagePoint incluye tres tipos de mapas que se corresponden con las tres últimas matrices mencionadas: mapa de correlación cruzada, mapa de auto-correlación y mapa de factores. Estos mapas son una representación gráfica de la matriz correspondiente. En el mapa de correlación cruzada, la similitud entre elementos se mide usando el coseno. En el mapa de factores y el mapa de auto-correlación, la medida de similitud usada es el coeficiente de correlación de Pearson. Las matrices de correlación se normalizan usando el coeficiente de correlación de Pearson, coseno de Salton, o *Max Proportional*. Además, la matriz de co-ocurrencia puede normalizarse usando la medida de similitud $tf \cdot idf$.

Finalmente, comentar que VantagePoint también permite ejecutar scripts desarrollados en Visual Basic por el usuario para realizar tareas repetitivas y complejas.

2.1.9. VOSViewer

VOSViewer[15] [107] es una herramienta diseñada especificamente para construir y visualizar mapas científicos, prestando especial atención a la representación gráfica de dichos mapas. Gracias a las capacidades de zoom, a algoritmos especiales de etiquetado y a las metáforas de densidad, VOSViewer es una herramienta apropiada para representar grandes mapas. La herramienta ha sido desarrollada por el *Centre for Science and*

[15] http://www.vosviewver.com

Technology Studies de la Universidad de Leiden (Países Bajos). Esta herramienta se puede utilizar gratuitamente.

Aunque VOSViewer se puede utilizar para construir y visualizar mapas científicos a partir de cualquier clase de datos de co-occurrencia, la herramienta no permite la creación de ninguna clase de red bibliométrica a partir de un conjunto de documentos científicos. Para crear dichas redes, es necesario la utilización de herramientas externas. Del mismo modo, la herramienta no posee capacidades de preprocesamiento, por lo que es necesario una herramienta externa para preparar los datos para el análisis posterior.

Para el posicionamiento de los elementos en el mapa, se utiliza la técnica de posicionamiento VOS [108], la cual construye una matriz de similitud a partir de una matriz de co-ocurrencia (el usuario tiene que crear dicha matriz y cargarla en la herramienta) utilizando la medida de similitud conocida como fuerza de asociación [105, 106]. La técnica VOS construye un mapa bidimensional en los que los elementos son posicionados de tal modo que las distancias entre cualquier par de elementos reflejan su grado de similitud, del modo más preciso posible.

Aunque VOSViewer implementa la técnica VOS, la herramienta puede visualizar cualquier otro mapa bidimensional construido con cualquier otra técnica.

VOSViewer nos permite realizar una detección de comunidades usando la técnica de clustering VOS, la cual está relacionada con la técnica de clustering basada en modularidad [111].

Una vez que el mapa ha sido construido, VOSViewer permite examinarlo a través de cuatro vistas diferentes:

- Vista de etiquetas. En esta vista cada elemento se representa por una etiqueta y un círculo. Cuanto más importante sea un elemento, mayor tamaño tendrá su etiqueta y mayor volumen tendrá su circulo asociado. Gracias a su algoritmo inteligente,

únicamente se muestran las etiquetas más importantes (más frecuentes) en cada nivel de zoom, por lo que se evitan los solapamientos. El color de los círculos representa el grupo al que pertenecen, de modo que dos círculos con idéntico color pertenecerán al mismo grupo. Dicho color es el mismo que el color de los grupos en la vista de grupos.

- Vista densidad. En esta vista cada elemento es representado, de un modo similar a la vista anterior, por una etiqueta. Cada punto en el mapa tiene un color que depende de la densidad de elementos en ese punto. La densidad de elementos en un punto del mapa dependerá tanto del número de elementos posicionados en ese punto, como del peso de estos. VOSViewer calcula la densidad de cada punto de acuerdo con la ecuación definida en [107], la cual usa función de nucleo Gausiano. La densidad es transformada en una escala de colores.

- Vista de grupos. Esta vista está disponible sólo si los elementos han sido previamente asignados a un grupo. La vista de grupos es similar a la vista de densidad, con la excepción de que la densidad de elementos en cada punto del mapa es calculada de forma separada para cada grupo.

- Vista de puntos. Esta es una vista simple en donde los elementos son representados exclusivamente por un círculo pequeño, sin ninguna etiqueta asociada.

2.2. Análisis Basado en Cinco Aspectos

Como se mencionó al principio de este capítulo, las herramientas anteriormente descritas han de ser comparadas para de este modo, poder resaltar sus principales diferencias y las sinergias existentes entre ellas. Para realizar dicho estudio comparativo,

en está sección analizaremos las nueve herramientas teniendo en cuenta cinco aspectos diferentes de ellas:

- Los métodos de preprocesamiento disponibles.
- Las redes bibliométricas que las herramientas son capaces de construir.
- Las medidas de normalización que cada herramienta permite aplicar sobre las redes o matrices.
- Los tipos de análisis que se permite realizar.
- Y finalmente, otros aspectos secundarios como la documentación y ayuda disponibles sobre la herramienta, la posibilidad de ejecutar la herramienta en diferentes plataformas o sistemas operativos, etc.

2.2.1. Métodos de Preprocesamiento

Una característica importante de las herramientas para análisis de mapas científicos es la disponibilidad de módulos de preprocesamiento. En la Tabla 2.1 se muestran los principales módulos para preprocesamiento disponible en cada herramienta. Para detalles concretos de cada uno de los métodos de preprocesamientos consultar la Sección 1.2.3.

Herramienta	Unificación	División temporal	Reducción de los datos	Reducción de la red
Bibexcel			x	x
CiteSpace		x	x	x
CoPalRed	x	x	x	
IN-SPIRE			x	
Loet Leydesdorff's Software				
Network Workbench Tool	x	x	x	x
Science of Science Tool	x	x	x	x
VantagePoint	x	x	x	
VOSViewer				

Tabla 2.1: Métodos de preprocesamiento.

La opción de división temporal es necesaria cuando el usuario quiere analizar la evolución temporal del dominio científico estudiado. Asimismo, el módulo para reducir

los datos es útil para filtrar la información, quedándonos con la más importante o significativa. Para más información sobre las redes bibliométricas consultar la Sección 1.2.2.

El módulo para unificar elementos es muy importante, por ejemplo, en un análisis basado en redes de co-palabras o co-autores. Con este módulo, el usuario podría decidir unir dos o más elementos que representen al mismo concepto o al mismo autor. Este módulo, no sólo une dos elementos en un único elemento, sino que además puede agregar algún valor de un atributo, como por ejemplo, el número de citas recibidas por un documento.

Finalmente, la reducción de la red es útil para filtrar los nodos y enlaces de la red (de un modo similar al módulo de reducción de datos), o aplicar un algoritmo de poda sobre la red.

Sólo las herramientas NWB y Sci^2 disponen de los cuatro módulos de preprocesamiento. Por contra, las herramientas Loet Leydesdorff's Software y VOSViewer no poseen ningún módulo de preprocesamiento, siendo esto una fuerte desventaja.

IN-SPIRE realiza la división temporal directamente sobre los datos, por lo que no es necesario preprocesar los datos para dividir el corpus en diferentes periodos de tiempo.

2.2.2. Redes Bibliométricas

Un aspecto importante a la hora de usar una herramienta para la realización de un análisis de mapas científicos es la posibilidad que ofrece de establecer diferentes relaciones entre las unidades de análisis. Dicho de otro modo, la capacidad que tiene la herramienta de extraer diferentes redes bibliométricas es de vital importancia.

En la Tabla 2.2 se muestran las diferentes redes bibliométricas que cada herramienta es capaz de generar. La columna "*otras*" quiere decir si la herramienta es capaz de

construir redes o matrices menos comunes o heterogéneas.

Aunque no hay una herramienta capaz de construir todas las variedades de redes bibliométricas, Bibexcel, CiteSpace, Loet Leydesdorff's Software, Sci^2 y VantagePoint son las herramientas que permiten construir la mayoría de ellas. Por el contrario, VOS-Viewer no es capaz de construir ninguna, ya que está centrado sólo en la visualización de datos. Por otro lado, CoPalRed se centra exclusivamente en un tipo de red. Finalmente, aunque IN-SPIRE es capaz de construir los mapas utilizando cualquier campo de los documentos, su modo de representar los documentos, usando el espacio modelo-vectorial, hace que sea difícil generar los mapas con otro campo, como por ejemplo, los autores.

Herramienta	Emparejamiento bibliográfico			Co-autor			Co-citación			Co-palabras	Enlace directo	Otras
	Autor (ABCA)	Documento (DBCA)	Revista (JBCA)	Autor (ACAA)	País (CCAA)	Institución (ICAA)	Autor (ACA)	Documento (DCA)	Revista (JCA)	(CWA)	(DL)	
Bibexcel		x		x	x	x	x	x	x	x		x
CiteSpace		x		x	x	x	x	x	x	x		x
CoPalRed										x		
IN-SPIRE										x		
Loet Leydesdorff's Software	x		x	x	x	x	x			x		
Network Workbench Tool		x		x				x		x	x	
Science of Science Tool	x	x	x	x			x	x	x	x	x	x
VantagePoint				x	x	x	x	x	x	x		x
VOSViewer												

Tabla 2.2: Redes bibliométricas.

Otras herramientas permiten la extracción de redes menos comunes, por ejemplo, la red de co-subvenciones disponible en CiteSpace, la red co-IP disponible en Sci^2, o las matrices particulares que se pueden extraer en Bibexcel y VantagePoint usando campos específicos de los documentos. Incluso, algunas herramientas como Bibexcel y VantagePoint son capaces de extraer redes heterogéneas combinando campos diferentes.

Finalmente, NWB y Sci^2 permiten extraer redes bibliométricas basadas en enlace directo.

2.2.3. Medidas de Normalización

Una vez que las redes bibliométricas han sido construidas, se puede realizar el proceso de normalización usando diferentes medidas de similitud. En la Tabla 2.3 se muestran las medidas usadas por cada herramienta. Una completa descripción de las medidas de normalización puede encontrarse en la Sección 1.2.4.

Herramienta	Measure
Bibexcel	Coseno de Salton, Índice de Jaccard, o las medidas de Vladutz y Cook
CiteSpace	Coseno de Salton, Índice de Dice o Índice de Jaccard
IN-SPIRE	Probabilidad condicional
CoPalRed	Índice de equivalencia
Loet Leydesdorff's Software	Coseno de Salton
Network Workbench Tool	Definida por el usuario
Science of Science Tool	Definida por el usuario
VantagePoint	Coeficiente de correlación de Pearson, el coseno de Salton, o la Máxima Proporcional
VOSViewer	Fuerza de asociación

Tabla 2.3: Medidas de normalización.

Tres de las herramientas analizadas utilizan el coseno de Salton como medida de similitud. Por el contrario, otras herramientas como NWB y Sci^2, permiten al usuario definir sus propias medidas.

2.2.4. Métodos de Análisis

Existen diversos tipos de análisis que pueden utilizarse para extraer conocimiento de los mapas científicos. En la Tabla 2.4 se muestran los distintos métodos disponibles en cada una de las herramientas. Una completa descripción de los métodos de análisis puede encontrarse en la Sección 1.2.6.

Herramienta	Detección de estallidos	Geoespacial	Redes	Temporal
Bibexcel			x	
CiteSpace	x	x	x	x
CoPalRed			x	x
IN-SPIRE	x		x	x
Loet Leydesdorff's Software				
Network Workbench Tool	x		x	x
Science of Science Tool	x	x	x	x
VantagePoint	x	x	x	x
VOSViewer			x	

Tabla 2.4: Métodos de análisis.

Sólo CiteSpace, Sci^2 y VantagePoint permiten realizar los cuatro tipo de análisis.

Por el contrario, Loet Leydesdorff's Software no dispone de ninguno de ellos.

CiteSpace y Sci^2 poseen capacidades de geolocalización. CiteSpace utiliza los servicios de geolocalización de Google y Yahoo! sobre los datos institucionales disponibles. Por otro lado, Sci^2 utiliza el servicio de Yahoo!'y un geolocalizador interno, sobre los datos que contienen información geográfica como, la dirección institucional, la localización de las conferencias, etc.

2.2.5. Otros Aspectos

En esta sección realizaremos una comparación de las herraientas atendiendo a otros aspectos, tales como, la documentación y ayuda disponible, el tipo de licencia de las herramientas (gratuita o comercial), la disponibilidad del código fuente de las herramientas, la posibilidad de instalar la herramienta en diferentes plataformas y sistemas operativos, y por último, las capacidades de extensión de las herramientas.

Las herramientas NWB y Sci^2 tienen una gran guía de usuario en donde se describen en profundidad. Además, dichas guías de usuario explican aspectos importantes del análisis de mapas científicos, siendo las únicas que realizan esta explicación. VantagePoint tiene también un buen manual de usuario y ayuda en-línea, además, su página web contiene una gran cantidad de vídeo-tutoriales. IN-SPIRE posee un gran sitio web, con diversos vídeo-tutoriales y ayuda en línea. VOSViewer dispone de un buen manual de usuario. CiteSpace dispone de una gran wiki donde se describen aspectos importantes de la herramienta. Loet Leydesdorff's Software tiene en su página web una buena descripción y manual de usuario para cada uno de los programas que lo componen.

Sólo tres de las nueve herramientas descritas son comerciales: CoPalRed, IN-SPIRE y VantagePoint. Las restantes se pueden usar de forma gratuita.

Teniendo en cuenta la disponibilidad del código fuente de las herramientas, sólo

NWB y Sci^2 lo tienen disponible para descarga.

CiteSpace, NWB, Sci^2 y VOSViewer han sido desarrollados utilizando el lenguaje de programación Java, por lo que pueden usarse en cualquier plataforma o sistema operativo (Windows, MacOS, Linux, etc.). Por otro lado, Bibexcel, CoPalRed, IN-SPIRE, Loet Leydesdorff's Software y VantagePoint sólo pueden utilizarse bajo el sistema operativo Windows.

Finalmente, teniendo en cuenta las posibilidades de extensión de las herramientas, NWB y Sci^2 han sido desarrolladas bajo Cyberinfrastructure Shell[16] (CIShell), por lo que pueden extenderse utilizando dicha plataforma. VantagePoint puede extenderse a través de scripts escritos en VisualBasic.

2.3. Análisis de los Mapas Generados: un con las Distintas Herramientas

La comparación realizada anteriormente se puede completar realizando un estudio cooperativo de las nueve herramientas estudiadas con un mismo conjunto de datos. Este estudio nos dará la oportunidad de descubrir las posibles sinergias positivas que podrían generarse al usar las herramientas de forma conjunta.

Para hacer una buena comparación entre las herramientas, realizaremos el mismo tipo de análisis de mapas científicos, sobre una unidad de análisis específica, con cada una de ellas. Como se mostró en la Tabla 2.2, las herramientas analizadas no son capaces de extraer las mismas redes bibliométricas, siendo la red de co-palabras la única disponible en todas las herramientas. Por esta razón, hemos seleccionado las palabras (concretamente, las palabras clave) como unidad de análisis para realizar el análisis de

[16] http://cishell.org/

mapas científicos.

Como ejemplo, estudiaremos la estructura conceptual del área de investigación de la Teoría de los Conjuntos Difusos (TCD) [115, 116] usando las publicaciones que han aparecido en el periodo de tiempo comprendido entre los años 2005 y 2009, en las dos revistas más importantes y prestigiosas del área, de acuerdo a su factor de impacto (JCR del 2009): *Fuzzy Sets and Systems* (FSS) y *IEEE Transactions on Fuzzy Systems* (IEEE-TFS). Un análisis profundo de estas dos revistas, analizando su evolución conceptual a lo largo de cinco periodos diferentes se realizará en el Capítulo 3.

En total se han analizado una cantidad de 1576 documentos, los cuales fueron descargados desde la ISIWoS[17] . Particularmente, 1086 documentos fueron publicados por la revista FSS y 490 por la revista IEEE-TFS.

Como unidades de análisis usamos las palabras clave de los documentos, concretamente las palabras clave dadas por los autores y las palabras clave dadas por la base de datos bibliográfica (ISI Keywords Plus) de donde se descargó la información. Después del proceso de unificación de elementos similares, que fue realizado mediante CoPalRed, existía un total de 5034 palabras clave en el corpus. CoPalRed nos permite exportar los documentos con sus elementos preprocesados en un fichero CSV, por lo que dicho fichero será la entrada para las herramientas restantes, siempre y cuando sea posible. La red global de palabras clave (sin reducir), construida a partir de las co-ocurrencias de las palabras clave en los documentos, contenía un total de 25075 enlaces o arcos.

A continuación, se mostrará los distintos resultados obtenidos por cada una de las herramientas. El estudio comparativo se ha realizado utilizando aquellas herramientas que permiten visualizar los resultados. Por esta razón, no hemos utilizado Bibexcel ni Loet Leydesdorff's software.

En primer lugar, realizamos un análisis de co-palabras con CiteSpace. Dado que

[17] La descarga se realizó el 15 de Enero de 2010.

CiteSpace no permite leer los datos en formato CSV, los documentos fueron cargados directamente de los ficheros descargados de la ISIWoS, por lo que los datos estaban en bruto, es decir, sin preprocesar. En la Figura 2.1 se muestra el mapa generado por CiteSpace. El mapa se creo utilizando las 200 palabras clave más frecuentes. Las líneas entre los nodos representan la similitud de ambos nodos, medida mediante el coseno de Salton. Los nodos sombreados representan los grupos detectados. Los nombre de los grupos se escogieron seleccionando las palabras clave más importantes de cada grupo de acuerdo a la medida $tf \cdot idf$.

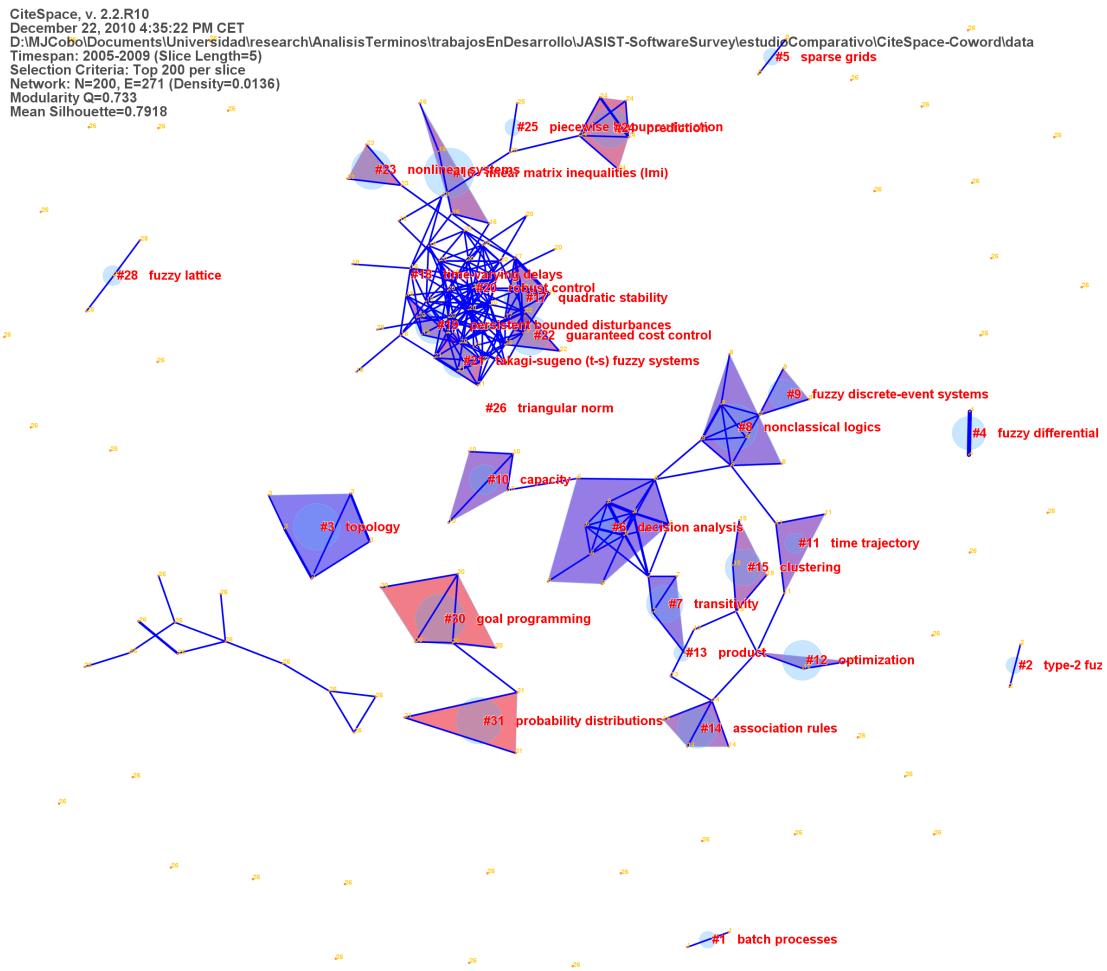
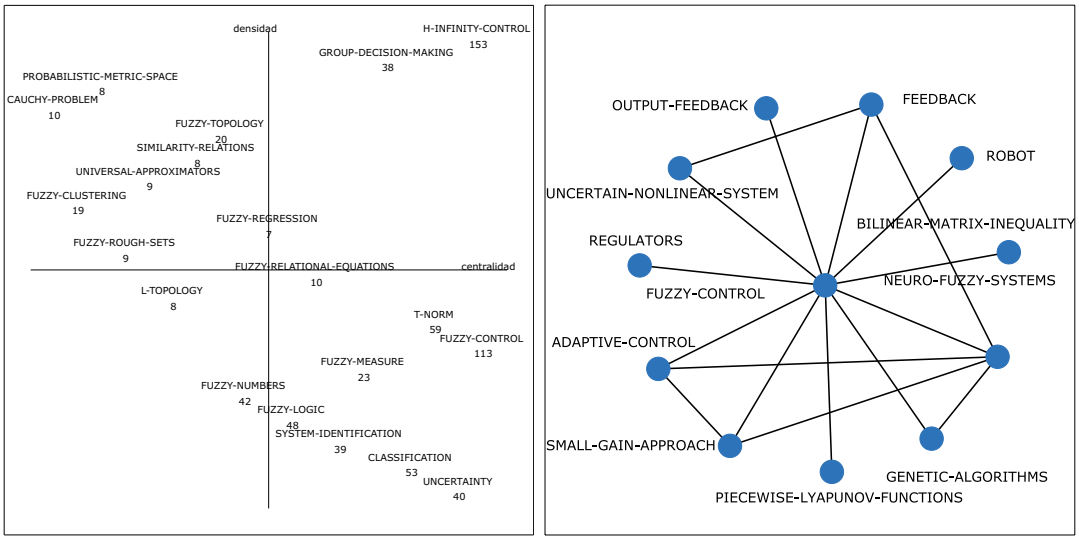


Figura 2.1: Mapa generado por CiteSpace.

Es necesario comentar que la versión impresa del mapa no refleja todo el poder de CiteSpace. Para realizar una buena interpretación de los resultados obtenidos, el analista debería interactuar con la interfaz gráfica de CiteSpace, la cual le permite realizar una variedad de análisis, aplicar diferentes algoritmos de posicionamiento, etc. Además, el analista puede ampliar o reducir una zona del mapa, para de este modo apreciar los detalles de esa zona.



(a) Diagrama estratégico. (b) Red temática.

Figura 2.2: Resultados de CoPalRed.

En segundo lugar, en la Figura 2.2 se muestran los resultados obtenidos por CoPalRed. En la Figura 2.2.a, se muestra el diagrama estratégico generado, y en la Figura 2.2.b, se muestra la red temática de un tema concreto: *FUZZY-CONTROL*. CoPalRed generó el mapa utilizando aquellas palabras clave con una frecuencia igual o superior a 5, y con una co-ocurrencia mayor o igual a 3. La red global, después del proceso de reducción de datos y red, contenía un total de 229 nodos y 432 enlaces entre ellos. Con esta reducción mantenemos las palabras clave más frecuentes e importantes del corpus. El diagrama estratégico muestra los principales temas estudiados por la comunidad

científica TCD, clasificándolos en cuatro clases diferentes de acuerdo a su densidad y centralidad[18] . Cada tema en el diagrama estratégico se representa a través de una esfera y una etiqueta. Las etiquetas representan a la palabra clave (nodo) más central de la red temática asociada al tema, en la cual, cada nodo se corresponde con una palabra clave.

En tercer lugar, el fichero en formato CSV exportado por CoPalRed fue cargado en la herramienta IN-SPIRE. Después de definir el corpus y seleccionar los términos a usar, IN-SPIRE crea dos mapas: mapa de galaxias (Figura 2.3) y mapa de temas (Figura 2.4).

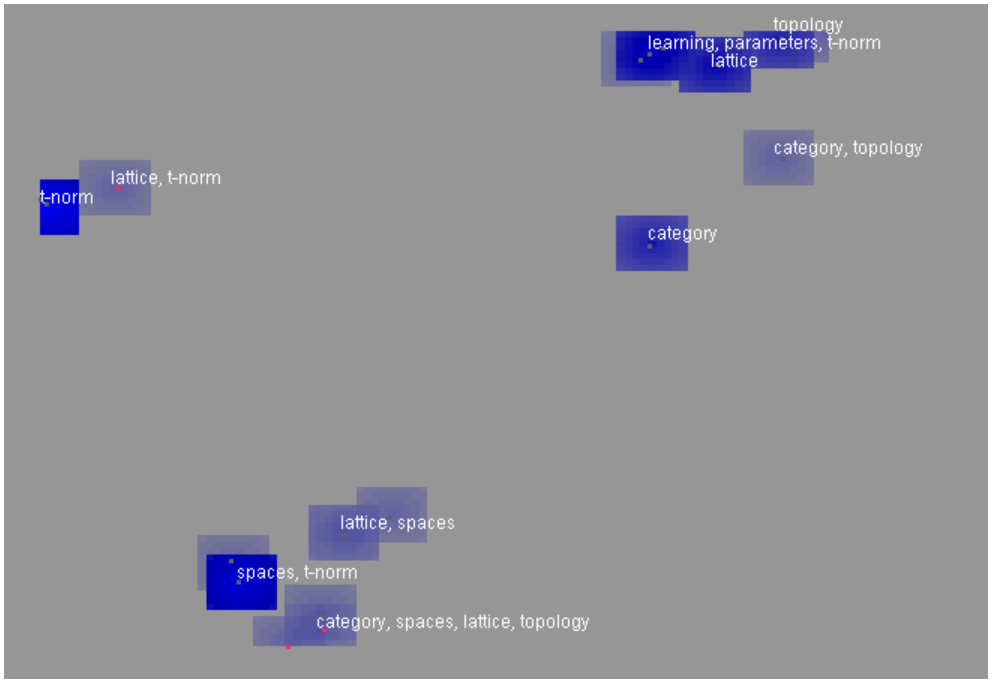


Figura 2.3: Vista Galaxy de IN-SPIRE'.

En el mapa de galaxias, las sombras representan grupos de documentos similares. Los nombres de dichos temas fueron generados usando las palabras clave más importantes de acuerdo a la medida $tf \cdot idf$. En el mapa de temas, la altura de cada pico representa

[18] Como medidas de densidad y centralidad se usaron las propuestas por Callon [20].

la fuerza del tema en una posición determinada, y su extensión se corresponde con el área y brillo del correspondiente tema en el mapa de galaxias.

Como podemos observar en ambos mapas, IN-SPIRE no detecta demasiados temas. Esto es debido al modo en que la herramienta interpreta la información. Al contrario que las otras herramientas estudiadas, IN-SPIRE utiliza el modelo de espacio-vectorial para representar los documentos, por lo que necesita una gran cantidad de términos para poder detectar correctamente los temas. En nuestro conjunto de datos, los documentos no contienen el número de palabras clave necesario, por lo que IN-SPIRE no puede determinar correctamente la similitud entre los documentos. Probablemente, si hubiésemos utilizado el resumen de los documentos o el texto completo de estos, IN-SPIRE habría obtenido mejores resultados.

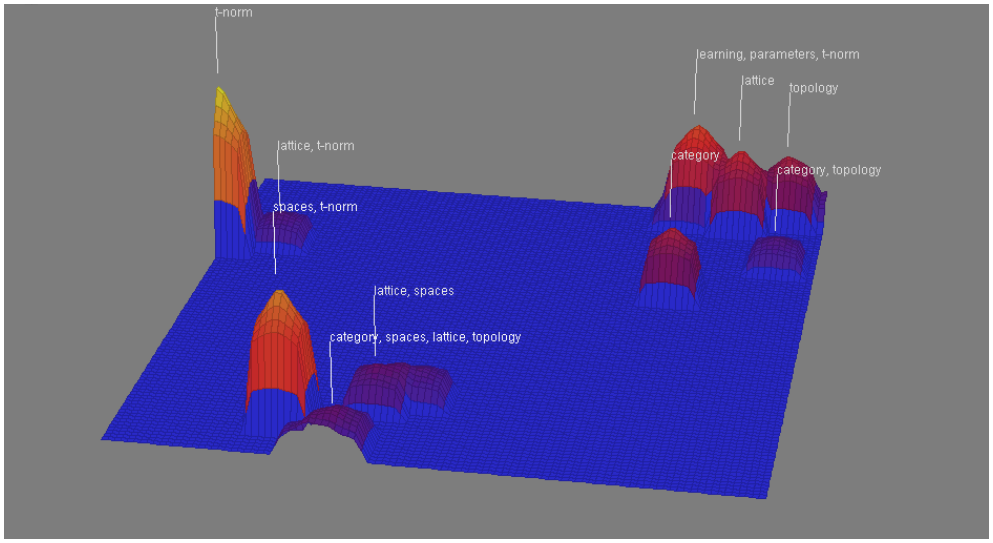


Figura 2.4: Vista Temas de IN-SPIRE'.

En cuarto lugar, el fichero CSV fue importado en la herramienta Sci^2, creando posteriormente una red de co-palabras. Después de eliminar las palabras clave con una frecuencia menor que 5 y los enlaces con una co-ocurrencia menor que 3 (la red global es la misma que generó CoPalRed), aplicamos un algoritmo de detección de

componentes débiles. La componente más grande detectada se muestra en la Figura 2.5, en ella, el tamaño de los nodos es proporcional a la frecuencia de la palabra clave correspondiente, y el grosor de las líneas representan la co-ocurrencia (sin normalizar) de los nodos enlazados. Únicamente se muestran las 50 palabras clave más frecuentes. El color de los nodos varía de un modo linear desde gris a negro dependiendo de su frecuencia, y el color de los enlaces varía linealmente desde verde a negro dependiendo de su valor de co-ocurrencia. La red se dibujo utilizando el complemento GUESS.

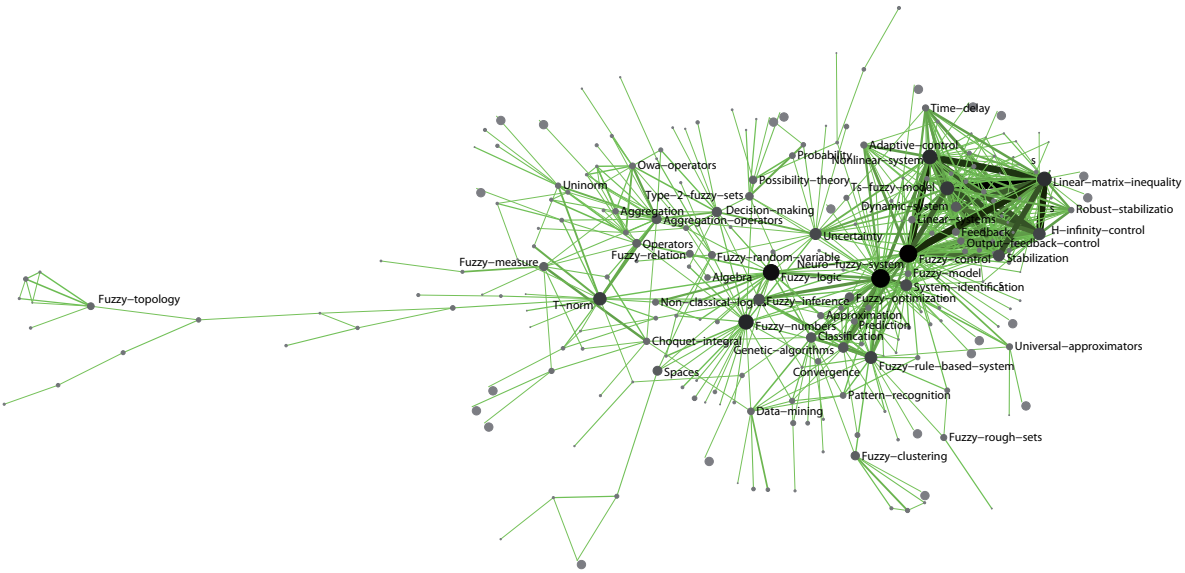


Figura 2.5: Mapa generado por Sci^2.

Es necesario comentar que en este ejemplo concreto, las herramientas Sci^2 y NWB obtienen los mismos resultados.

En quinto lugar, se construyó un mapa de factores utilizando VantagePoint (Figura 2.6), utilizando aquellas palabras clave con una frecuencia igual o mayor a 5. Cada nodo representa un grupo de términos. La etiqueta de cada tema se eligió de acuerdo a su término más importante. El tamaño de los nodos es proporcional al número de documentos, y el grosor de las líneas entre los nodos representan la similitud (coeficiente

de correlación de Pearson) entre los factores (nodos).

Finalmente, la matriz de co-ocurrencia generada por CoPalRed se transformó, mediante un script diseñado ad-hoc, al formato de VOSViewer, para poder visualizar los resultados con él. En la Figura 2.7 se muestra la vista de grupos. Podemos observar como las diferentes palabras clave se posicionan a lo largo de una línea horizontal. Esto significa que las palabras clave situadas a la izquierda del mapa y las situadas a la derecha tienen un grado de similitud muy bajo entre si. El tamaño de las etiquetas de las palabras clave es proporcional a su frecuencia; además, VOSViewer sólo visualiza las más importantes (más frecuentes) en el nivel más alto de zoom. VOSViewer selecciona un color diferente y aleatorio para cada grupo. Dentro de cada grupo, la fuerza del color en un punto particular, representa la densidad en ese punto. La densidad se mide utilizando funciones de núcleo Gaussiano [107].

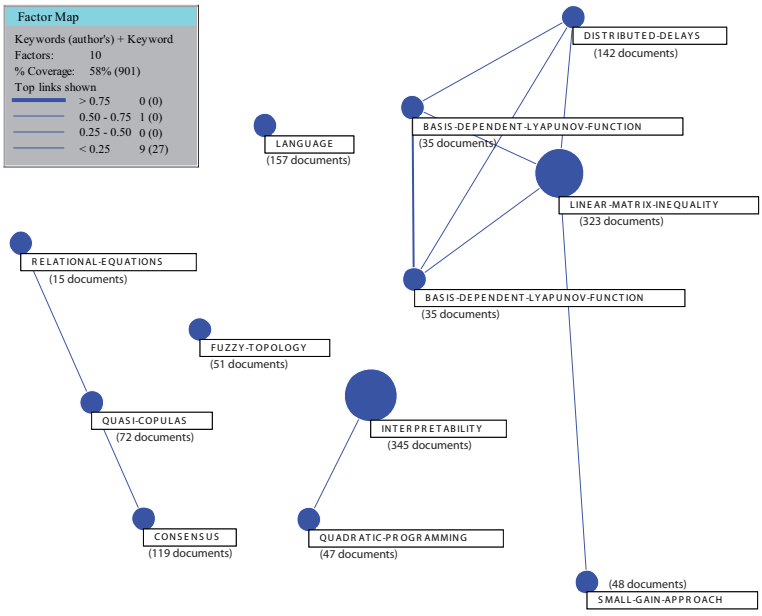


Figura 2.6: Mapa generado por VantagePoint.

De un modo similar a CiteSpace, la versión impresa de VOSViever no muestra todo

el poder de su interfaz gráfica. En cada vista, el usuario puede ampliar o centrar una zona específica, para de este modo descubrir los elementos escondidos detrás de aquellos más importantes. Como ejemplo, en la Figura 2.8 se muestra una vista ampliada de la vista de grupos, centrada en las palabras clave *FUZZY-TOPOLOGY* y *T-NORM*.

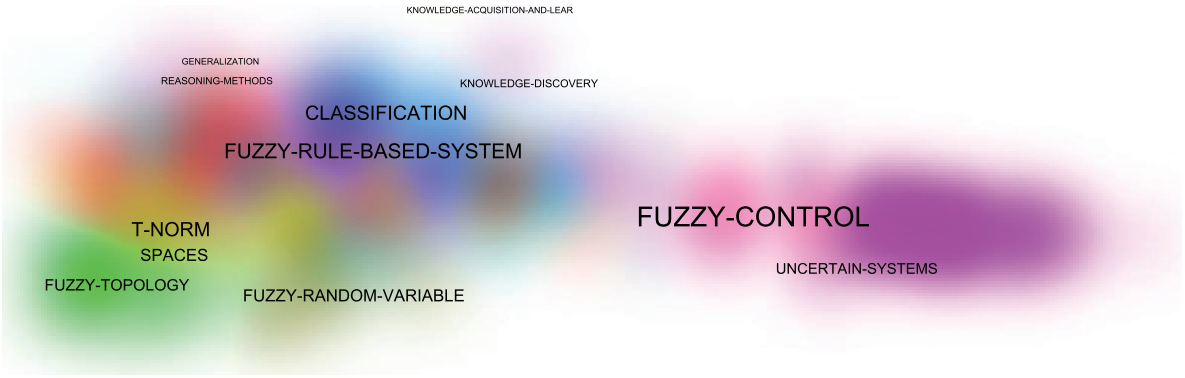


Figura 2.7: Vista de grupos de VOSViewer.

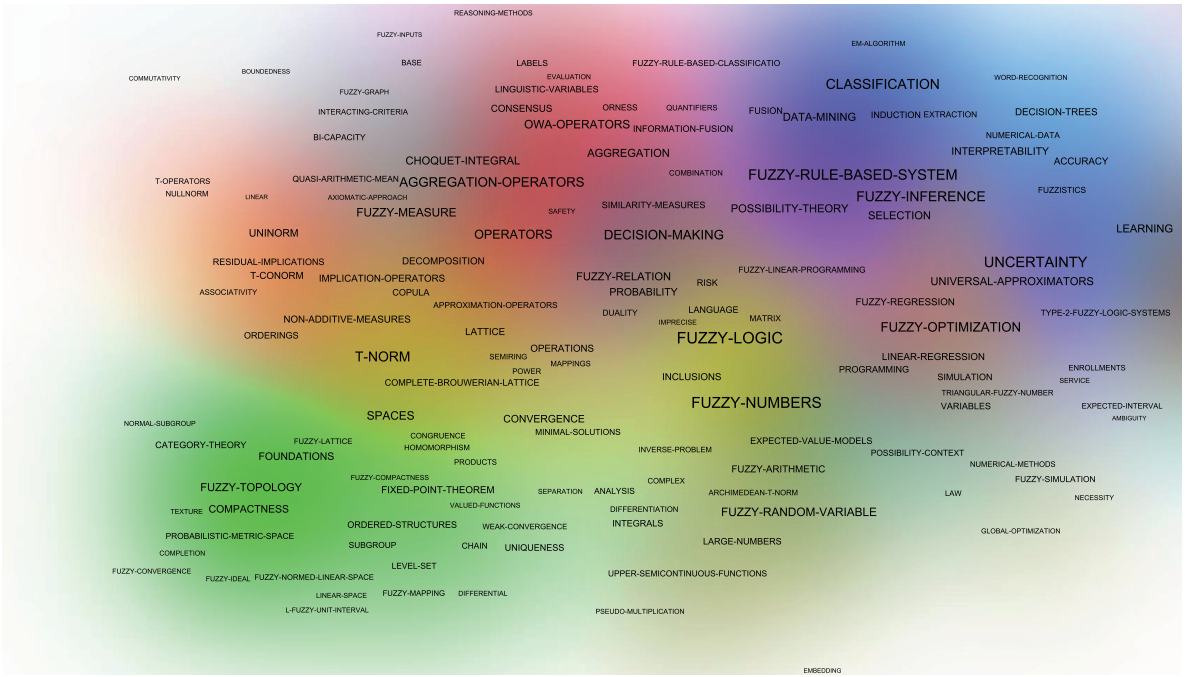


Figura 2.8: Vista de grupos ampliada de VOSViewer.

2.4. Lecciones Aprendidas

Como se ha mostrado en las Secciones 2.2 y 2.3, cada herramienta posee características diferentes. Muchas de las herramientas contienen potentes herramientas de preprocesamiento, otras permiten la generación de una gran cantidad de redes bibliométricas, y otras en cambio sólo permiten extraer un tipo de red. Finalmente, no todos los pasos del análisis pueden llevarse acabo en cada una de las herramientas. Por este motivo, un análisis profundo y completo de un campo de estudio utilizando mapas científicos requiere la utilización de varias de estas herramientas.

Teniendo en cuenta las capacidades de preprocesamiento, VantagePoint es la herramienta más potente. Incorpora una gran cantidad de filtros de importación que permiten leer la información exportada de la gran mayoría de bases de datos bibliográficas. Además, las funciones de limpieza de listas y la posibilidad de realizar dicha tarea mediante la aplicación de tesauros, facilitan el preprocesamiento, en particular la unificación de elementos similares. VantagePoint, nos permite exportar sus resultados en un fichero CSV. De este modo, otras herramientas pueden leer los datos preprocesados por él y realizar así otras etapas del análisis.

CoPalRed contiene también un buen módulo para la unificación de elementos, pero la herramienta está centrada en una única unidad de análisis: las palabras clave. Las herramientas NWB y Sci^2 tienen un módulo de unificación de elementos, pero este necesita realizarse a través de un proceso externo utilizando una herramienta adicional. No obstante, ambas herramientas tienen buenas funciones para la reducción de redes.

Atendiendo a las redes bibliométricas, las herramientas analizadas permiten generar una gran cantidad de redes, pero como se mostró en la Tabla 2.2 no existe ninguna herramienta capaz de extraer todas las clases de redes bibliométricas.

Teniendo en cuenta los mapas generados, y las visualizaciones de estos realizadas

por cada herramienta, existen bastantes diferencias, como se mostró en la Sección 2.3:

- CiteSpace es capaz de visualizar las redes utilizando diferentes algoritmos de posicionamiento. El nombre de los grupos detectados se puede asignar utilizando métricas diferentes. Finalmente, su interfaz gráfica nos permite interactuar con la red para de este modo, realizar una buena exploración de esta.

- CoPalRed agrupa los elementos (palabras clave) en temas, categorizándolos en un diagrama estratégico de acuerdo a sus medidas de centralidad y densidad. Esta categorización nos permite detectar los *temas motores* del campo científico estudiado. Por cada tema, CoPalRed genera una red temática que muestra las relaciones entre las palabras clave que forman el tema.

- IN-SPIRE permite la visualización de dos tipos de mapas, siempre y cuando se le nutra con suficientes datos. En la vista de temas, el analista puede detectar las zonas más importantes del mapa, es decir, aquellas donde se localizan una mayor concentración de documentos. La vista de galaxias nos permite detectar fácilmente los documentos similares basándonos en su contendido.

- Las herramientas NWB y Sci^2 generan unas visualizaciones similares. Ambas permiten visualizar las redes usando diferentes complementos, aplicando diversos algoritmos de posicionamiento, o scripts que permiten adaptar la visualización a las necesidades del usuario. Además, la herramienta Sci^2 incorpora mapas temáticos, en donde la información se muestra sobre un mapa terrestre o geográfico.

- VantagePoint tiene tres clases de mapas, lo que permite al usuario crear diferentes vistas de los datos. En la vista de mapa, VantagePoint muestra una leyenda para dar información acerca del grosor de las líneas, siendo la única herramienta que ofrece esta clase de leyenda. Quizás, un de los puntos fuertes de VantagePoint es

su interfaz gráfica, la cual permite al usuario seleccionar un conjunto de elementos del mapa, y mostrar sus documentos asociados con sus elementos, así como otro tipo de información la vista de detalle.

- VOSViewer tiene una potente interfaz gráfica que nos permite examinar los mapas de un modo muy intuitivo. Por otro lado, detectar, visualmente, los temas más importantes no es una tarea sencilla. Además, en la vista de grupos es difícil determinar a qué grupo pertenece los elementos que están en la frontera entre dos grupos.

De acuerdo con los métodos de análisis disponibles, existen también diferencias entre las herramientas analizadas. Por ejemplo, el análisis geoespacial sólo está disponible en CiteSpace, Sci^2 y VantagePoint, teniendo sólo los dos primeros capacidades de geolocalización, lo que permitiría representar las redes sobre un mapa terrestre (usando Google Maps, Yahoo! Maps, etc).

Herramientas	Preprocesamiento	Redes	Normalización	Análisis
Bibexcel	Reducción de datos y redes	DBCA, ACAA, CCAA, ICAA, ACA, DCA, JCA, CWA, Otros	Coseno de Salton, índice de Jaccard, o las medidas de Vladutz y Cook	Red
CiteSpace	División temporal, y reducción de datos y redes	DBCA, ACAA, CCAA, ICAA, ACA, DCA, JCA, CWA, Otros	Coseno de Salton, Índice de Dice o Índice de Jaccard	Detección de estallidos, geoespacial, red, temporal
CoPalRed	Unificación, división temporal y reducción de datos	CWA	Índice de equivalencia	red, temporal
IN-SPIRE	Reducción de datos	CWA	Probabilidad condicional	Detección de estallidos, red, temporal
Loet Leydesdorff's Software		ABCA, JBCA, ACAA, CCAA, ICAA, ACA, CWA	Coseno de Salton	
Red Workbench Tool	Unificación, división temporal y reducción de datos y redes	DBCA, ACAA, DCA, CWA, DL	Definida por el usuario	Detección de estallidos, red, temporal
Science of Science Tool	Unificación, división temporal y reducción de datos y redes	ABCA, DBCA, JBCA, ACAA, ACA, DCA, JCA, CWA DL, Otros	Definida por el usuario	Detección de estallidos, geoespacial, red, temporal
VantagePoint	Unificación, división temporal y reducción de datos	ACAA, CCAA, ICAA, ACA, DCA, JCA, CWA, Otros	Coeficiente de correlación de Pearson, el coseno de Salton, o la Máxima Proporcional	Detección de estallidos, geoespacial, red, temporal
VOSViewer			Fuerza de asociación	Red

Tabla 2.5: Resumen de características.

En la Tabla 2.5 se muestra un resumen de las características de las nueve herramientas de acuerdo a los cuatro aspectos considerados en la Sección 2.2. Como podemos observar, las herramientas CiteSpace, IN-SPIRE, NWB, Sci^2 y VantagePoint pueden identificarse como las más completas.

Es necesario comentar que las herramientas NWB y Sci^2 tienen partes comunes ya que comparten algoritmos. NWB está especializada en el análisis, modelado y visualización de redes, mientras que Sci^2 ha sido diseñada para realizar estudios de la ciencia. Ambas han sido desarrolladas por el *Cyberinfrastructure for Network Science Center*, compartiendo bastantes algotritmos y métodos. De todos modos, algunas características como, la geolocalización son exclusivas de Sci^2.

Algunas veces, la importación de un conjunto de datos con formato específico es bastante compleja en las nueve herramientas analizadas. Otras veces, es difícil modificar las herramientas o incorporar nuevas medidas, algoritmos y visualizaciones. Por este motivo, la capacidades de extensión como las que proveen NWB, Sci^2 y VantagePoint son muy útiles.

Como se mencionó anteriormente, cada herramienta tiene diferentes características e implementan diferentes técnicas que son llevadas a cabo mediante diversos algoritmos. Consecuentemente, cada herramienta ofrece su vista particular del campo científico analizado. El uso combinado de las diferentes herramientas podría permitirnos desarrollar un completo análisis de mapas científicos. De hecho, pensamos que la cooperación entre las herramientas ofrece una sinergia positiva que nos brinda la posibilidad de extraer el conocimiento desconocido que de otro modo quedaría sin descubrir.

Finalmente, comentar que el análisis realizado en este capítulo no incorpora todas las herramientas para realizar análisis de mapas científicas existentes. Esto es debido a que los investigadores usualmente utilizan sus propias herramientas y algoritmos desarrollados a medida, quizás motivados por la falta de flexibilidad de las herramientas existentes. Aunque estas herramientas podrían tener características similares a las herramientas analizadas en este capítulo, no han sido publicadas en revistas científicas. Algunas veces, estas herramientas no se publican debido a que fueron desarrolladas para

realizar un estudio concreto, quedando su publicación en un segundo plano.

Capítulo 3

Una Metodología para Detectar, Cuantificar y Visualizar la Evolución de un Área Científica

La construcción de mapas científicos a partir de información bibliográfica [39] es una técnica comúnmente utilizada para extraer los diferentes aspectos sociales, intelectuales y conceptuales de un comunidad científica a partir de sus publicaciones. Además, el análisis de mapas científicos puede englobarse dentro de un marco temporal o longitudinal, para de este modo, estudiar los cambios en la estructura conceptual, social o intelectual de la comunidad científica a lo largo del tiempo. Por otro lado, los indicadores bibliométricos se utilizan para medir aspectos cualitativos y cuantitativos de las publicaciones científicas. De hecho, son especialmente útiles para medir el impacto y calidad de éstas, así como el rendimiento de una comunidad científica.

Ambas técnicas, aunque son muy utilizadas y conocidas no suelen combinarse. En este sentido, el enriquecimiento de los mapas científicos con indicadores bibliométricos de calidad (como por ejemplo el muy conocido índice h) sería de gran utilidad, ya que nos ayudaría a determinar qué aspectos del mapa científico son de mayor calidad, cuáles han tenido un mayor impacto en la comunidad científica y cuáles han sido más

productivos.

En este Capítulo se presenta una metodología para la realización del análisis de mapas científicos bajo un marco longitudinal y enriquecidos con indicadores bibliométricos para medir la calidad y rendimiento. Aunque la metodología diseñada puede emplearse para analizar cualquier aspecto de un campo científicos, nos hemos centrado en un tipo particular de red bibliométrica para de este modo, facilitar la compresión de la metodología propuesta. Así, la metodología se centrará en las redes de co-palabras, analizando los aspectos conceptuales de un campo científico. Como veremos al final del Capítulo, esta metodología es fácilmente ampliable/adaptable para utilizar cualquier tipo de unidad de análisis y red bibliométrica.

El Capítulo se estructura de la siguiente forma: en la Sección 3.1 se describe la metodología. En la Sección 3.2, la metodología será validada mediante su utilización en el estudio de un campo científico concreto: el área científica de la Teoría de los Conjuntos Difusos. Finalmente, en la Sección 3.3 se describe cómo la metodología puede extenderse y generalizarse para utilizar otras unidades de análisis, y por tanto, analizar diferentes aspectos de un campo científico.

3.1. Descripción de la Metodología

Formalmente, la base metodológica del análisis de co-palabras se basa en la idea de que la co-ocurrencia de términos puede describir el contenido de un conjunto de documentos [20]. De acuerdo con [54], esta técnica muestra las asociaciones entre términos, a través de la construcción de múltiples redes que resaltan las asociaciones entre dichos términos y entre las redes. En esta metodología, estas redes serán asociadas con temas.

Cada publicación o documento, en el campo científico bajo estudio, puede caracterizarse por un conjunto de términos clave. Dichos términos pueden interpretarse como

la *huella dactilar*, o el ADN de una publicación [12]. De este modo, la similitud de un par de documentos puede medirse a través de la comparación de las huellas dactilares formadas por sus palabras clave. Cuantas más palabras clave tengan en común dos documentos, más similares serán, y por lo tanto, será más probable que pertenezcan a la misma especialidad dentro de un campo de investigación particular. Siguiendo la metáfora del ADN, si las huellas de dos publicaciones son suficientemente similares, provendrán de la misma especie [54].

Con la lista de los términos importantes, o palabras clave del campo científico, se puede construir un grafo o una red (red bibliométrica de co-palabras). En esta red, los nodos representan las palabras clave, y los enlaces entre ellos representan sus relaciones. Dos nodos (palabras clave) estarán conectados si ambos aparecen en los mismos documentos. Además, podemos añadir un peso a los enlaces, de modo que represente cómo es de importante esa relación en el seno del corpus analizado.

Como resultado del análisis de co-palabras, para cada uno de los periodos de tiempo estudiados se obtiene un conjunto de temas. Estos resultados pueden visualizarse aplicando diferentes técnicas. En concreto, la metodología propuesta emplea diagramas estratégicos [18, 20, 29, 31, 42] para visualizar y categorizar los temas detectados, redes temáticas [5] para mostrar las palabras clave de cada tema y las relaciones entre ellas, y por último, áreas temáticas para mostrar la evolución conceptual del campo analizado.

En resumen, las diferentes etapas o fases realizadas por la metodología son las siguientes:

1. Detectar los temas tratados por el campo científico a través del análisis de co-palabras para cada uno de los periodos definidos.

2. Plasmar en un espacio de baja dimensionalidad los resultados de la primera fase (temas).

3. Analizar la evolución de los temas detectados a lo largo de los periodos de tiempo definidos, para de este modo, identificar las principales áreas temáticas del campo científico analizado, sus orígenes, y sus relaciones.

4. Medir el rendimiento, producción, calidad e impacto de los diferentes periodos, temas, y áreas temáticas, a través de medidas cuantitativas y bibliométricas.

Cada una de estas etapas tienen que desarrollarse de un modo secuencial, tal y como se muestra en la Figura 3.1.

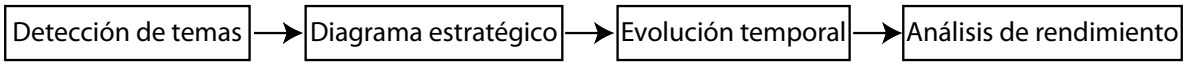


Figura 3.1: Flujo de trabajo de las etapas de la metodología.

En las siguientes secciones se describe en detalle cada una de estas etapas.

3.1.1. El Proceso de Detección de Temas

El proceso de delimitación estructural y conceptual un campo científico, se suele dividir en una serie de pasos consecutivos [12, 20, 25, 29, 30, 31, 48, 57, 84, 94, 97, 101, 104]. Es decir, la detección de los temas tratados por una comunidad científica en un periodo de tiempo determinado suele comprender una serie consecutiva de pasos. En nuestra metodología, el proceso de detección de temas se divide en cinco pasos:

1. Recopilación de los datos.

2. Selección de la unidad de análisis a utilizar.

3. Extracción de la red bibliométrica.

4. Cálculo de las relaciones de similitud entre los elementos de la red bibliométrica.

5. Aplicación de un algoritmo de agrupamiento o *clustering* para detectar los temas.

El primer paso es la recopilación de los datos con los que vamos a trabajar. De este modo, por ejemplo, para analizar un campo científico los datos que necesitamos recopilar son los documentos publicados por dicho campo. Como comentamos en capítulos anteriores, la recopilación de los documentos científicos puede realizarse a través de fuentes de información bibliográfica, como ISIWoS, Scopus, o Google Scholar, entre otras. La recuperación o recopilación de los documentos relevantes se realiza a través de una consulta con los términos descriptivos del campo científico que queremos analizar, intentando siempre cubrir el mayor ámbito posible. Una vez que hemos recogido los datos, estos pueden dividirse en diferentes periodos de tiempo para poder así analizar la evolución temporal del campo científico. Los diferentes periodos han de construirse utilizando un conjunto de años consecutivos.

El segundo paso consiste en la selección del tipo de unidad de análisis a utilizar. Como se describió en [12] y se recoge en la Sección 1.2.2, como unidad de análisis se suelen utilizar, las revistas, documentos, autores, y términos descriptivos. En nuestro caso, usaremos como unidad de análisis las palabras clave presentes en los documentos científicos. En concreto, como palabras clave se pueden utilizar, las dadas por los autores, por la revista, por la base de datos bibliográfica (como por ejemplo ISI Keywords Plus), o cualquier combinación de estas.

El tercer paso en el proceso, es la extracción de la información relevante a partir de los documentos recopilados en el primer paso, esto es, la construcción de la red bibliométrica a partir de las relaciones de las unidades de análisis contenidas en los documentos recopilados. En nuestra metodología, la información relevante consistirá en la frecuencia de co-aparición de palabras clave. Es decir, la red bibliométrica se construirá a partir de la co-ocurrencia de palabras clave, generando por tanto una red de

co-palabras. La co-ocurrencia de dos palabras clave del corpus analizado se establece mediante el conteo de documentos en los que ambas palabras clave aparecen juntas.

Una vez que se ha construido la red bibliométrica de co-palabras, el cuarto paso se basa en el cálculo de la similitud entre los elementos de la red. Esta similitud se calcula a partir de los valores de co-ocurrencia de las palabras clave en el corpus. El cálculo de la similitud entre dos palabras clave tiene como base, la normalización de los valores de co-occurrencia. De este modo, se les dará mayor valor a una palabra clave con una frecuencia baja pero que suele co-aparecer siempre con la misma palabra clave. Por el contrario, una palabra clave con una alta frecuencia y que co-aparece con un gran número de palabras clave tendrá un peso menor. Es decir, con la normalización se intenta potenciar las parejas de palabras clave que representan de un modo adecuado al corpus analizado. Como vimos en la Sección 1.2.4, en la literatura existen un gran número de medidas de similitud que pueden utilizarse para normalizar un red bibliométrica, como por ejemplo, el coseno de Salton [86], o el índice de Jaccard [75], o el índice de equivalencia [20], entre otros. Aunque en el resto del capítulo usaremos el indice de equivalencia [20], la metodología es flexible es este aspecto, por lo que permite utilizar cualquier medida de similitud para normalizar las redes bibliométricas.

El quinto paso está basado en un proceso de agrupamiento (aplicando un algoritmo de clustering), para localizar los grupos de palabras clave que están fuertemente relacionadas entre sí, y que se corresponden con los centros de interés o los problemas de investigación, en los que los investigadores han estado centrados [20]. Existe un gran número de algoritmos que pueden emplearse para particionar una red bibliométrica en diferentes subredes o subgrafos. Recientemente, algunos autores han propuesto algoritmos de clustering para realizar esta tarea: *Streemer* [48], *spectral clustering* [25], *modularity maximization* [26] y *bootstrap resampling* basado en clustering significativo

[85]. Al igual que en el paso anterior, la metodología que proponemos es flexible en cuanto al algoritmo de clustering a utilizar, siempre y cuando este puede aplicarse sobre una matriz de similitud y devuelva un conjunto de grupos etiquetados (cada grupo debe tener un nombre). Si el algoritmo de clustering no asigna ninguna etiqueta a los grupos, esta tarea se puede realizar mediante cualquier post-proceso automático o manual. Como ejemplo, proponemos el uso de un algoritmo de clustering basado en centros simples [29]. Este algoritmo es sencillo y bien conocido en el contexto del análisis de co-palabras. Además, ha sido usado en múltiples estudios basados en redes de co-palabras [5, 6, 29, 30, 31, 42, 59, 60]. Una ventaja de este algoritmo, es que los grupos devueltos tienen una etiqueta asignada, la cual se corresponde con el nodo más central del grupo. Por lo tanto, no es necesario ningún proceso a parte para la asignación de etiquetas a los grupos.

Tal y como fue descrito en [29], el algoritmo de los centros simples utiliza dos etapas para producir las redes deseadas. En la primera etapa, (Paso-1) se construyen las redes que representan las asociaciones más fuertes. Los enlaces añadidos en esta etapa se llamarán *enlaces internos*. La segunda etapa (Paso-2), consiste en añadir a las redes anteriores, enlaces que formen asociaciones entre ellas. Los enlaces añadidos durante esta segunda etapa se llamarán *enlaces externos*. El pseudo-código del algoritmo de los centros simples contiene los siguientes pasos [29]:

1. Seleccionar un umbral mínimo para la co-ocurrencia, c_{ij}, de las palabras clave i y j, seleccionar un valor máximo para el número de enlaces en el Paso-1, y seleccionar un valor máximo para el total de los enlaces del Paso-1 y Paso-2.
2. Empezar Paso-1;
3. Generar el mayor índice de equivalencia e_{ij} de entre todas las parejas de palabras clave posible para comenzar una red en el Paso-1;
4. Con este enlace, buscar otros enlaces mediante una búsqueda en anchura hasta que no haya más

enlaces posibles, debido al valor mínimo de co-ocurrencia, al valor máximo de enlaces en el Paso-1, o al número máximo de nodos. Eliminar todas las palabras clave incorporadas de la siguiente lista de palabras clave disponibles para el Paso-1;

5. Repetir los pasos 3 y 4 hasta que se hayan formado todas las redes posibles del Paso-1; por ejemplo, hasta que no quede ningún par de palabras clave con la co-ocurrencia suficiente para formar una red;
6. Empezar Paso-2;
7. Incorporar todas las palabras clave del Paso-1 a la lista de palabras clave disponibles.
8. Empezar con la primera red detectada en el Paso-1;
9. Encontrar todos los enlaces entre los nodos de la red del Paso-1 actual a cualquier nodo de otra red del Paso-1, que tengan al menos el valor mínimo de co-ocurrencia, en orden descendente del valor e_ij; parar cuando no quede ninguna pareja de palabras clave con un valor de co-ocurrencia superior al mínimo, o cuando se haya alcanzado el máximo número de enlaces. No eliminar ninguna palabra clave de la lista de palabras clave disponibles;
10. Seleccionar la siguiente red del Paso-1 y continuar con el paso 9.

Como se comentó en [29], dos palabras clave poco frecuentes en el corpus, pero que siempre aparecen juntas, tendrán un mayor peso que palabras clave con frecuencia muy alta y que aparecen conjuntamente con muchas de las palabras clave restantes. Por este motivo, asociaciones débiles o irrelevantes pueden dominar la red. El algoritmo de los centros simples soluciona este problema usando diferentes parámetros: umbral mínimo de frecuencia y de co-ocurrencia. Sólo las parejas de palabras clave que superen esos umbrales serán consideradas como enlaces potenciales en la construcción de las redes durante el Paso-1. Por otro lado, el algoritmo tiene dos parámetros más para limitar el tamaño de las redes detectadas: el mínimo y máximo tamaño de una red (tema).

Aunque el algoritmo de los centros simples tiene sólo cuatro parámetros, las subredes o temas detectados son muy dependientes de ellos. Por esta razón, se necesita un proceso

para determinar los parámetros que mejor se adaptan a nuestras necesidades. Esto hace que suela ser de utilidad que un grupo de expertos en la materia que estamos analizando nos ayuden y validen los resultados obtenidos, para de este modo, poder determinar la mejor configuración de parámetros que nos permita detectar los temas principales tratados por el campo científico.

Finalmente, se pueden utilizar dos medidas para representar las redes detectadas: la centralidad y densidad de Callon [20].

La *centralidad de Callon*, centralidad de aquí en adelante, mide el grado de interacción de una red con respecto a otras redes [20]. Puede definirse como: $c = 10 * \sum e_{kh}$, siendo k una palabra clave perteneciente al tema y h una palabra clave perteneciente a otro tema. La centralidad mide el grado de fuerza de los enlaces externos del tema con otros temas. Esta medida se puede interpretar como la importancia de un tema en el desarrollo global de campo científico analizado, o como el grado de cohesión externa del tema.

La *densidad de Callon*, densidad de aquí en adelante, mide la fuerza interna de una red [20]. Puede definirse como: $d = 100\frac{\sum e_{ij}}{w}$, donde i y j son palabras clave pertenecientes al tema y w el número de palabras clave (nodos) que forman el tema. La densidad mide la fuerza interna de todos los enlaces entre las palabras clave que describen al tema, o dicho de otro modo, el grado de cohesión interna del tema.

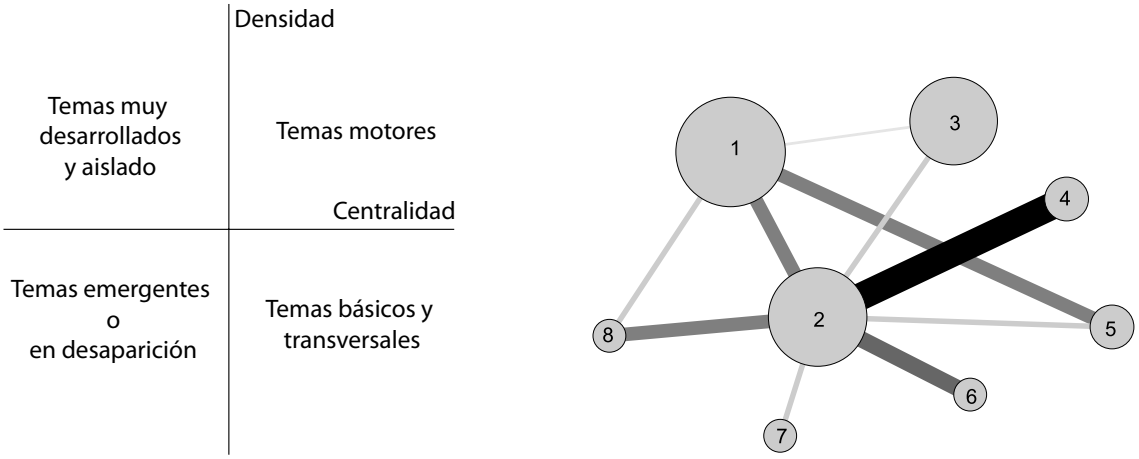
3.1.2. Visualización de Temas y Redes Temáticas

Como resultado de un análisis científico basado en un red de co-palabras, se obtienen un conjunto de grupos de palabras clave y sus inter-conexiones. Estos grupos de palabras clave serán llamados *temas*.

Cada tema obtenido en el proceso puede caracterizarse por dos medidas: densidad

y centralidad. Tanto los valores medios, como la mediana de ambas medidas pueden usarse para clasificar los temas en cuatro clases diferentes [18, 20, 29, 31, 42]. De este modo, un campo científico puede representarse como un conjunto de temas clasificados en cuatro categorías y posicionados sobre un espacio bidimensional.

Un Diagrama Estratégico es un espacio bidimensional construido mediante la colocación de los temas en él de acuerdo a sus rangos (si usamos la mediana para clasificar los grupos) o valores (si usamos la media) de centralidad y densidad, a lo largo de dos ejes: la centralidad en el eje X, y la densidad en el eje Y. Debido a su legibilidad, los diagramas estratégicos basados en rangos se utilizan de forma más frecuente que aquellos basados en valores [18]. Como ejemplo, en la Figura 3.2.a se muestra un diagrama estratégico.



(a) Diagrama estratégico. (b) Ejemplo de una red temática.

Figura 3.2: Diagrama estratégico y red temática.

El diagrama estratégico se divide en cuadrantes, de modo que podemos encontrar cuatro tipos de temas [18, 20, 29, 31, 42], dependiendo del cuadrante en el que se sitúen cada uno de ellos:

- En el cuadrante superior-derecho se encuentran los temas bien desarrollados e importantes para la construcción del campo científico. Estos temas son conocidos

como los *temas motor* de la especialidad, dado que presentan una fuerte centralidad y una alta densidad. El emplazamiento de un tema en éste cuadrante implica que este está muy relacionado externamente con conceptos aplicables a otros temas.

- Los temas del cuadrante superior-izquierdo poseen unos enlaces internos muy bien desarrollados, por lo que tienen una importancia marginal en el campo científico. Estos temas se caracterizan por estar muy especializados y ser muy periféricos.

- En el cuadrante inferior-izquierdo se situan los temas muy poco desarrollados y marginales. Los temas en este cuadrante tienen una densidad y centralidad baja. Principalmente representan temas emergentes o en desaparición.

- Los temas en el cuadrante inferior-derecho son importantes para el campo científico pero no están bien desarrollados. En este cuadrante se encuentran los temas transversales y genéricos, es decir, los temas básicos del campo científico.

En un tema, las palabras clave y sus interrelaciones forman una red o un grafo, a la cual llamaremos *red temática*. Cada red temática se etiqueta usando el nombre de la palabra clave más significativa del tema asociado (normalmente, es la palabra clave más central del tema). En la Figura 3.2.b, se muestra un ejemplo de una red temática. En ella, se muestran un conjunto de palabras clave diferentes e interconectadas entre sí. El volumen de las esferas es proporcional al número de documentos asociados a las palabras clave del tema. Por último, el grosor de las líneas entre dos palabras clave i y j, es proporcional a su índice de equivalencia e_{ij}.

Junto a la red global de temas y palabras clave interconectadas, un segunda red puede construirse, basándonos en los documentos asociados con cada red temática. En esta segunda red, a cada red temática se le asignan los documentos que comparten alguna

palabra clave con la red. De esto modo, podemos considerar dos tipos de redes: *documentos principales* y *documentos secundarios*. Dada una red temática, un documento será llamado "documento principal" si contiene al menos dos palabras clave de la red temática. Por otro lado, si el documento sólo contiene una palabra clave asociada con la red temática, se llamará "documento secundario". Tanto los documentos principales, como los secundarios pueden pertenecer a más de una red temática, y por lo tanto, a más de un tema.

Además, el diagrama estratégico puede enriquecerse añadiendo una tercera dimensión a los elementos representados en él, de modo que se ofrezca una mayor cantidad de información. De este modo, los temas pueden representarse como una esfera, en donde su volumen sea proporcional a diferentes medidas cuantitativas o cualitativas. Por ejemplo, como tercera dimensión podría utilizarse: i) el número de documentos asociados a un tema (documentos principales + documentos secundarios); ii) el número de citas recibidas por los documentos asociados a cada tema; iii) el número de autores[1] investigando en el campo asociado al tema.

3.1.3. Áreas Temáticas: la Evolución de los Temas

En esta sección describiremos qué son las áreas temáticas y cómo pueden detectase y visualizarse.

Si el corpus de documentos se divide en diversos grupos de años consecutivos (por ejemplo, en periodos de tiempo), se puede analizar la evolución del campo científico bajo estudio.

Sea T^t el conjunto de los temas detectados en el periodo de tiempo t, donde $U \in T^t$ representa cada uno de los tema detectados en el periodo t. Sea $V \in T^{t+1}$ el conjunto

[1] Un autor está asociado con un tema, si éste ha publicado algún documento relacionado con el tema.

de los temas detectados en el siguiente periodo de tiempo $t+1$. Diremos que hay una evolución temática desde el tema U al tema V si y sólo si las redes temáticas de ambos temas comparten al menos una palabra clave. De este modo, V puede considerarse como un tema que ha evolucionado de U. Las palabras clave $k \in U \cap V$ se considerarán como el "nexo temático" o el "nexo conceptual" de la evolución. Así, los mapas científicos de evolución pueden construirse enlazando temas del periodo T^t con temas del periodo T^{t+1} a través de los nexos conceptuales.

Las áreas temáticas pueden considerarse como un grafo bipartito. Un grafo bipartito es aquel en el que sus vértices están divididos en dos conjuntos disjuntos U y V, y los arcos o enlaces del grafo sólo pueden conectar elementos del conjunto U con elementos del conjunto V.

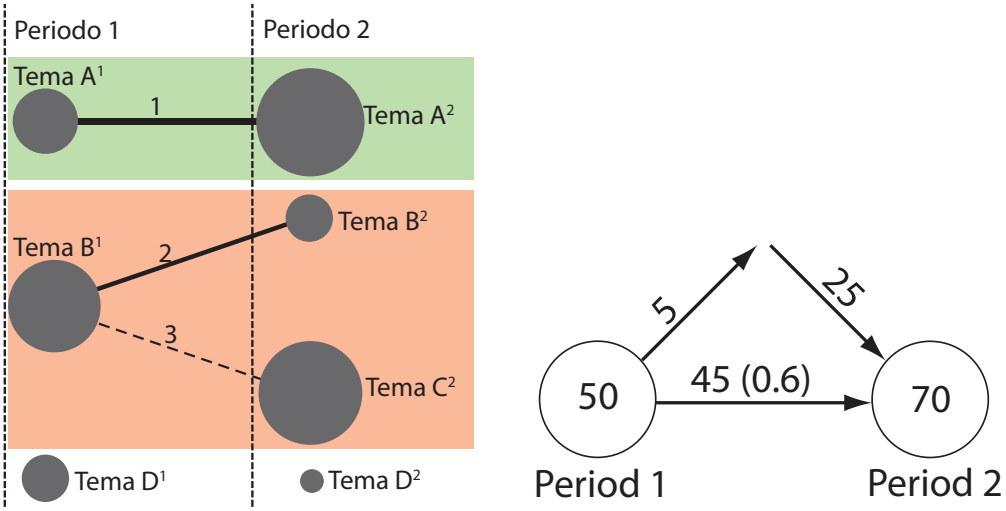
De este modo, habrá un enlace desde los temas del periodo t a los temas del periodo $t+1$ si existe un nexo conceptual entre ellos. Dicho de otro modo, si los temas contienen algún elemento en común.

La importancia de un nexo temático puede medirse a través de los elementos que ambos temas tienen en común. En nuestra metodología, utilizaremos el índice de inclusión [84] para realizar está tarea.

Aunque el peso de un nexo temático puede medirse a través de otras medidas de similitud, como por ejemplo, el índice de Jaccard, o el coseno de Salton, el índice de inclusión tiene la ventaja de ser más adecuado para medir la similitud entre conjuntos, en comparación con las restantes medidas de similitud (ver más detalles en la Sección 1.2.4). Esto es debido a que el índice de inclusión no está influenciado por el número de elementos, al contrario que las otras. De hecho, el índice de inclusión, ha sido usado como medida del grado de solapamiento en el campo de la recuperación de la información [106]. Además, el índice de inclusión es igual a 1 en el caso de que las palabras clave

del tema V estén completamente contenidas en el tema U. Por esta razón, y debido al hecho de que el peso de los nexos temáticos representan el grado de solapamiento entre los temas que lo componen, hemos elegido el índice de inclusión como medida para establecer el grado de importancia, o fuerza de los nexos temáticos. Aunque, al igual que en la selección de la medida de similitud usada para normalizar la red y en la selección del algoritmo de clustering, la metodología permite elegir cualquier otra medida de solapamiento para establecer el peso de un nexo conceptual.

Por lo tanto, un área temática es definida como un grupo de temas que han evolucionado a lo largo de diversos periodos de tiempo consecutivos. Cabe resaltar, que dependiendo de las interconexiones entre los temas, un mismo tema podría pertenecer a dos o más áreas temáticas diferentes, e incluso, no pertenecer a ningún área.



(a) Thematic evolution. (b) Stability between periods.

Figura 3.3: Ejemplos de evolución temporal.

Por ejemplo, supongamos que tenemos dos periodos de tiempo (periodo 1 y periodo 2), con tres temas detectados en el primero y cuatro en el segundo (junto con sus redes temáticas asociadas). En la Figura 3.3.a se muestra como ejemplo un mapa bibliométrico de evolución. En él, las líneas sólidas (líneas 1 y 2) significan que los temas enlazados

comparten el mismo nombre: ambos temas fueron etiquetados con la misma palabra clave, o la etiqueta de alguno de los temas es parte del otro tema (nombre del tema $\in$ { en el nexo conceptual }). Una línea discontinua (línea 3) significa que los temas comparten elementos que no son el nombre de los temas (nombre del tema $\notin$ { nexos conceptual }). El grosor de los enlaces entre dos temas es proporcional al valor del índice de inclusión de ambos temas. El volumen de las esferas es proporcional al número de documentos asociados con el tema. De hecho, al igual que en el diagrama estratégico, se pueden emplear otras medidas cuantitativas o cualitativas para establecer el volumen de las esferas (citas, índice h, etc.). Por último, las líneas verticales separan los periodos de tiempo.

En la Figura 3.3.a se pueden observar dos áreas temáticas distintas delimitadas por sombreados con un color diferente. Una de ellas está compuesta por los temas $TemaA^1$ y $TemaA^2$, y la otra compuesta por los temas $TemaB^1$, $TemaB^2$ y $TemaC^2$. $TemaD^1$ es un tema que no ha sido continuado, y $TemaD^2$ puede ser considerado como el comienzo de una nueva área temática.

Al igual que los temas tienen asociados un conjunto de documentos (documentos principales, documentos secundarios, o la unión de ambos), las áreas temáticas también pueden tener asociadas una colección de documentos. En este caso, los documentos asociados con un área temática serán determinados mediante la unión de los documentos asociados con el conjunto de temas pertenecientes a cada área temática.

Finalmente, el grado de solapamiento general entre los elementos de dos periodos consecutivos, puede medirse a través del índice de estabilidad [94], el cual tiene una ecuación similar al índice de Jaccard ($\frac{items_{ij}}{items_i + items_j - items_{ij}}$) para el caso de dos periodos de tiempo consecutivos [15]. El grado de solapamiento general mide el número de elementos compartidos entre dos periodos consecutivos. Para representar de un modo gráfico la

estabilidad a lo largo de diferentes periodos de tiempo, se ha utilizado una imagen similar a la presentada en [81].

Siguiendo el ejemplo anterior, en la Figura 3.3.b, se muestra la estabilidad entre dos periodos de tiempo consecutivos. Los círculos representa los periodos, y el número en su interior, el total de palabras clave del periodo. Las flechas horizontales representan el número de palabras clave (también puede ser autores, referencias, etc.) compartidas por ambos periodos, y entre paréntesis, se encuentra el índice de estabilidad. Las flechas entrantes representan el número de palabras clave del periodo, y las flechas salientes representan las palabras clave que están presentes en el periodo 1, pero no en el periodo 2

3.1.4. Análisis del Rendimiento

En las secciones previas, mostramos el proceso de detección de temas y áreas temáticas. Dichos resultados, pueden enriquecerse realizando un análisis del rendimiento a través de diversas medidas. Estas medidas pueden dividirse en dos categorías: cuantitativas y cualitativas. A través de medidas cuantitativas se puede medir la productividad o producción de los temas y áreas temáticas, mientas que a través de las medidas cualitativas se puede medir la calidad de los temas y áreas temáticas basándonos en medidas bibliométricas.

- Medidas cuantitativas: número de documentos, autores, revistas y universidades, instituciones, países.

- Medidas cualitativas o de impacto: número de citas recibidas por los documentos, media, máximo y mínimo de citas, o medidas más complejas basadas en índices bibliométricos, como el índice h [1, 46], índice g [34], índice hg [2] o índice q^2 [17].

Es necesario remarcar que ambas medidas pueden aplicarse a diferentes niveles para de este modo ayudar a analizar los temas, áreas temáticas y los periodos definidos.

3.2. El Área Científica de la Teoría de los Conjuntos Difusos: Aplicación de la Metodología

En esta sección aplicamos la metodología, presentada anteriormente, para analizar el campo científico de la Teoría de los Conjuntos Difusos (TCD) [115, 116], utilizando las publicaciones que han aparecido en las revistas más importantes del área: *Fuzzy Sets and Systems* (FSS) y *IEEE Transactions on Fuzzy Systems* (IEEE-TFS). La primera de ellas es la publicación oficial de la *International Fuzzy Systems Association* (IFSA), y la segunda la publicación de la *IEEE Computational Intelligence Society* para sistemas difusos. En comparación con otras revistas del área TCD, ambas presentan el mayor factor de impacto (FI) y el mayor número de documentos publicados (ver Tabla 3.1). Además, ambas son las más longevas del área.

Revista	FI 2008	FI 2007	FI 2006	Documentos	Año de comienzo
IEEE TRANSACTIONS ON FUZZY SYSTEMS	3.624	2.137	1.803	1243	1993
FUZZY SETS AND SYSTEMS	1.833	1.373	1.181	6309	1978
INTERNATIONAL JOURNAL OF UNCERTAINTY FUZZINESS AND KNOWLEDGE-BASED SYSTEMS	1.000	0.376	0.406	756	1993
JOURNAL OF INTELLIGENT & FUZZY SYSTEMS	0.649	0.221	0.283	503	1993

Tabla 3.1: Datos básicos a cerca de las revista del Área de Teoría y Conjuntos Difusos.

La revista FSS es la más antigua, empezando en 1978. La base de datos bibliográfica ISIWoS indexa sus publicaciones desde el año 1980. Por otro lado, IEEE-TFS comenzó en el año 1993, y ISIWoS indexa sus publicaciones desde el año 1994.

El estudio que realizamos en esta sección comprende desde el año 1978, hasta el año 2009 (ambos inclusive). En la Figura 3.4 se muestra la distribución de documentos (particularmente, hemos considereado los Artículos, Cartas, Artículos de congresos y

Revisiones como documentos a analizar) por año. En este periodo de tiempo, FSS publicó un total de 5724 documentos, mientas que IEEE-TFS publicó un total de 1169 documentos.

Para la recopilación de los datos se utilizaron tres bases de datos: ISIWoS, Scopus y Science Direct. Desde la ISIWoS recopilamos todos los documentos de ambas revistas para los años que tiene indexados. Para los años no indexados por la ISIWoS, utilizamos Scopus (para el año 1993 de la revista IEEE-TFS), y Science Direct (para los años 1978 y 1979 de la revista FSS).

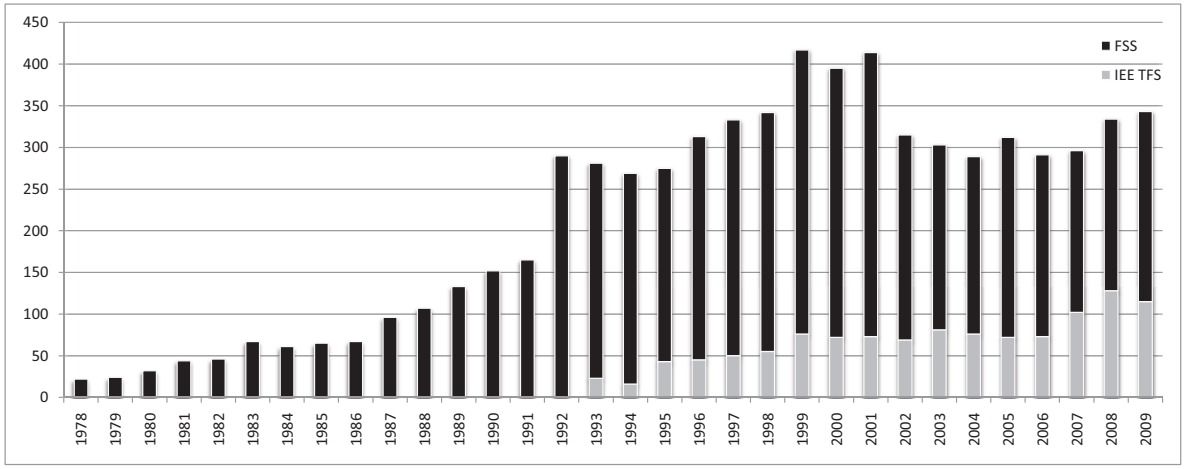


Figura 3.4: Documentos publicados en el campo del área de Teoría y Conjuntos Difusos desde 1978 a 2009.

En este estudio también se utilizan las citas recibidas por los documentos. En este sentido, se considerarán las citas recibidas hasta el 15 de Enero del 2010, fecha en la que se descargaron los datos. Además, exclusivamente se tendrán en cuenta las citas procedentes ISIWoS.

Los datos se dividieron en cinco periodos de tiempo consecutivos: 1978-1989, 1990-1994, 1995-1999, 2000-2004 y 2005-2009. En la Figura 3.5 se muestra la distribución de documentos por periodo.

Normalmente, la mejor opción a la hora de crear los periodos de tiempo es que

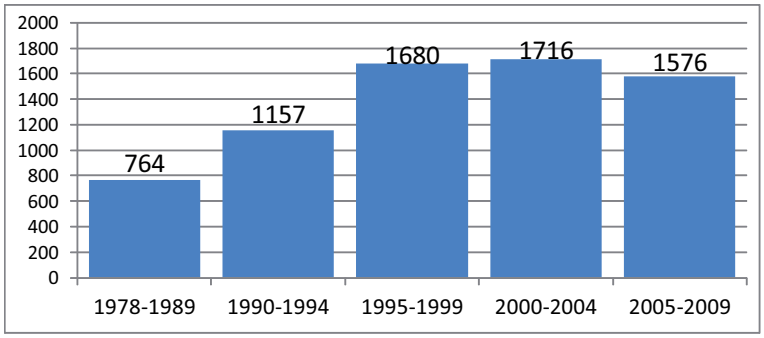


Figura 3.5: Documentos publicados por periodo.

estos contengan sólo un año, de esto modo se evita el suavizado (smooth, en la literatura anglosajona) de los datos. En el caso del estudio del área científica TCD, los periodos de un año de duración no contienen la cantidad suficiente de datos para que el proceso de detección de temas a través del análisis de co-palabras pueda realizarse satisfactoriamente. Por esta razón, en nuestro estudio los periodos contienen más de un año. Adicionalmente, aunque lo común es utilizar periodos de la misma longitud temporal, nuestro primer periodo contiene un total de doce años (1978-1989). De este modo, nutrimos con datos suficientes al proceso de detección de temas, para de este modo, detectar correctamente los temas principales. En los primeros años del área TCD, había muy pocos investigadores y por lo tanto pocas publicaciones. Además, observamos como la comunidad científica de este área tendía a utilizar un número de palabras clave extremadamente bajo para describir los documentos publicados (el número medio de palabras por documento en este periodo es de 1, de hecho, hay 117 documentos con menos de dos palabras clave). Por este motivo, los primeros doce años nos ofrecen un número adecuado de documentos y palabras clave para procesar. Por otro lado, observamos como en el siguiente periodo, el campo científico TCD empezó a consolidarse como disciplina, por lo que periodos de cinco años de duración son apropiados para proveer una buena entrada al proceso.

El análisis de co-palabras se realizó mediante la herramienta CoPalRed, el cual fue descrito en la Sección 2.1.3. CoPalRed se basa en el algoritmo de los centros simples para detectar los temas a través de diferentes periodos de tiempo. Tanto la representación de los diagramas estratégicos, redes temáticas y áreas temáticas se realizó utilizando herramientas desarrolladas ad-hoc.

Como se mencionó en el segundo paso de la metodología en la Sección 3.1.1, se utilizarán las palabras clave como unidad de análisis. Debido a que los datos fueron descargados de ISIWoS, se utilizaron tanto las palabras clave introducidas por los autores, como aquellas asignadas por la base de datos (ISI Keywords Plus). Ambos conjuntos de palabras clave se utilizaron de forma conjunta.

Primeramente, se realizó un proceso de unificación de las palabras clave, de modo que se unificaron las formas del singular y plural de las palabras clave (FUZZY-NUMBER, FUZZY-NUMBERS), así como los acrónimos (GDM ⇒ GROUP-DECISION-MAKING) con sus respectivas palabras clave. Además, se unificaron aquellas palabras clave que representaban al mismo concepto.

Para medir el rendimiento y la calidad de los temas y de las áreas temáticas detectadas, se realizó un análisis cuantitativo y de impacto en cada periodo de tiempo analizado. Para estudiar el rendimiento, se utilizó el número de documentos asociados a cada tema y área temática (documentos principales + documentos secundarios). Por otro lado, la calidad e impacto se realizó contabilizando el número de citas recibidas por los documentos asociados a los temas y áreas temáticas, así como el índice bibliométrico índice h.

Es necesario comentar que esta propuesta metodológica también se ha aplicado para analizar diversos ámbitos, pero por motivos de espacio los resultados obtenidos no se muestran en esta memoria de tesis. Particularmente, se han estudiado los siguientes

ámbitos científicos:

- El campo de los sistemas inteligentes de transporte, analizando las publicaciones de la revista *IEEE Transactions on Intelligent Transportation Systems* [27].
- Las hibridaciones entre la TCD y otras técnicas de inteligencia computacional como las redes neuronales o los algoritmos genéticos [59].
- El campo científico de la TCD en la comunidad española y en la comunidad científica internacional, detectando las similitudes y diferencias de ambas comunidades [60].

Centrándonos en el análisis que nos ocupa en esta sección, el área TCD estudiada a través de dos de sus revistas más importantes, a continuación, mostraremos la evolución conceptual de sus temas y redes temáticas, así como la evolución de dichos temas y el análisis del rendimiento de ambos resultados.

3.2.1. Dectección y Visualización de los Temas del Área de Teoría y Conjuntos Difusos

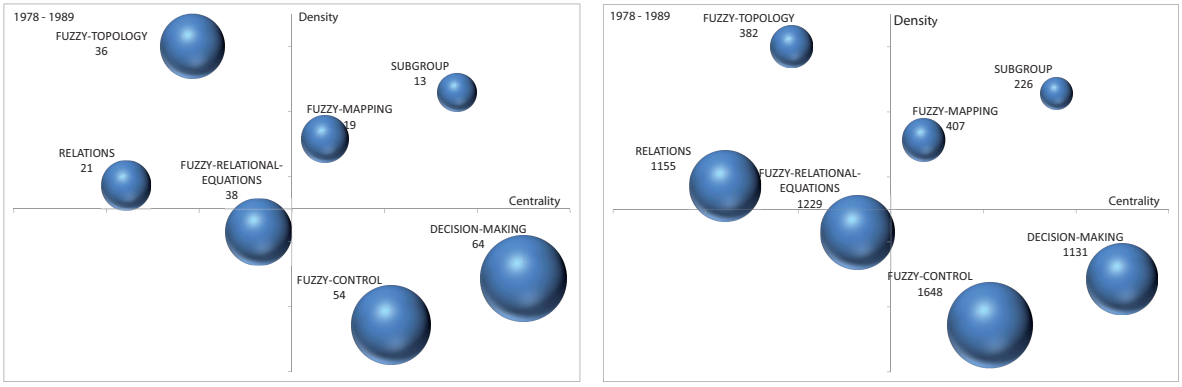
La visualización de los temas detectados en el campo científico de la TCD, a lo largo de los cinco periodos de tiempo analizados, se ha realizado a través de dos tipos de diagramas estratégicos. En el primero de ellos, el volumen de las esferas es proporcional al número de documentos publicados (documentos principales + documentos secundarios)[2] asociados con cada tema. En el segundo tipo de diagrama estratégico, el volumen de las esferas es proporcional al número total de citas recibidas por los documentos asociados a cada tema. Ambos diagramas estratégicos nos permitirán mostrar y resaltar los temas más importantes del área.

[2] El uso de los documentos secundarios y los principales implica que un mismo documento puede pertenecer a más de un tema.

A continuación, se mostrarán los diagramas estratégicos de cada periodo de tiempo, así como algunas tablas con medidas cuantitativas y de impacto, para poder de este modo, analizar cada periodo de tiempo de forma detallada.

En el periodo 1978-1989, aquel con mayor cantidad de años, se consideraron un total de 764 documentos, provenientes en exclusiva de la revista FSS.

De acuerdo con los diagramas estratégicos, (Figura 3.6) y a las medidas cuantitativas y de impacto (Tabla 3.2) podemosobservar que: i) los temas motores, *SUBGROUP* y *FUZZY-MAPPING* son muy poco citados y no han tenido una gran repercusión posterior a tenor de su bajo índice h; ii) los temas básicos y transversales, *FUZZY-CONTROL* y *DECISION-MAKING*, recibieron una gran cantidad de citas y tuvieron un gran impacto en la comunidad en los siguientes años; iii) el tema específico, *FUZZY-RELATIONAL-EQUATIONS* también tuvo una gran cantidad de citas y un fuerte impacto.



(a) Diagrama estratégico basado en el número de documentos publicados.

(b) Diagrama estratégico basado en el número de citas.

Figura 3.6: Diagramas estratégicos del periodo 1978-1989.

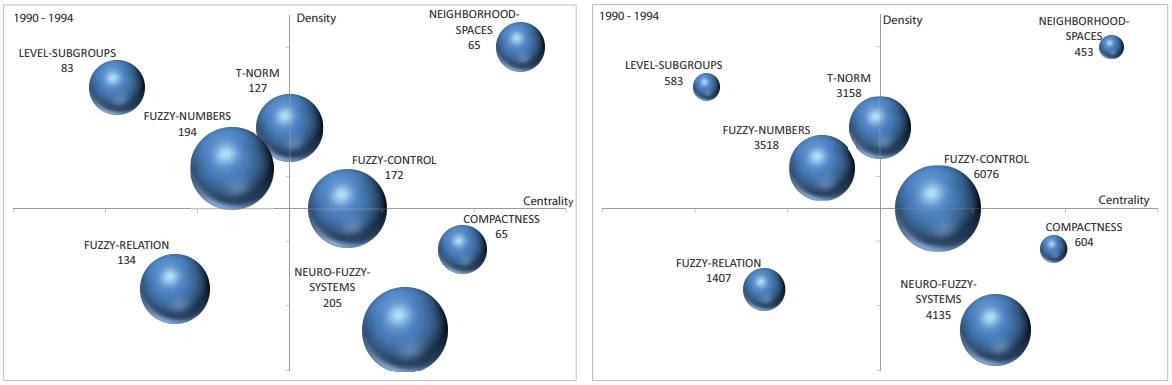
Debemos señalar que en este periodo sólo 230 documentos (en torno al 30 % de los documentos publicados en aquellos años) pudieron asociarse con algún tema. Esto es consecuencia directa del número bajo de palabras clave por documento, lo que hace

que el análisis de co-palabras y la asociación de documentos a los temas sea una ardua tarea.

Nombre del tema	Número de documentos	Número de citas	Media de citas	Índice h
DECISION-MAKING	64	1131	17.67	14
FUZZY-CONTROL	54	1648	30.52	18
FUZZY-RELATIONAL-EQUATIONS	38	1229	32.34	19
FUZZY-TOPOLOGY	36	382	10.61	13
RELATIONS	21	1155	55.00	7
FUZZY-MAPPING	19	407	21.42	11
SUBGROUP	13	226	17.38	6

Tabla 3.2: Medidas de rendimiento para los temas del periodo 1978-1989.

En el periodo 1990-1994 se publicaron un total de 1157 documentos en el campo científico TCD. En estos años, la revista IEEE-TFS comenzó a publicar, por lo que en este periodo de tiempo los documentos analizados pertenecen a ambas revistas.



(a) Diagrama estratégico basado en el número de documentos publicados.

(b) Diagrama estratégico basado en el número de citas.

Figura 3.7: Diagramas estratégicos del periodo 1990-1994.

A la vista de los resultados mostrados en la Figura 3.7 y en la Tabla 3.3, cabe resaltar que: i) el tema motor *NEIGHBORHOOD-SPACES* es el de menor impacto y menos citado; ii) los temas básicos *FUZZY-CONTROL* y *NEURO-FUZZY-SYSTEMS* son los temas más citados, y por tanto presentan el mayor impacto; iii) los temas específicos *FUZZY-NUMBERS* y *T-NORM* también muestran unos valores altos de citación e impacto.

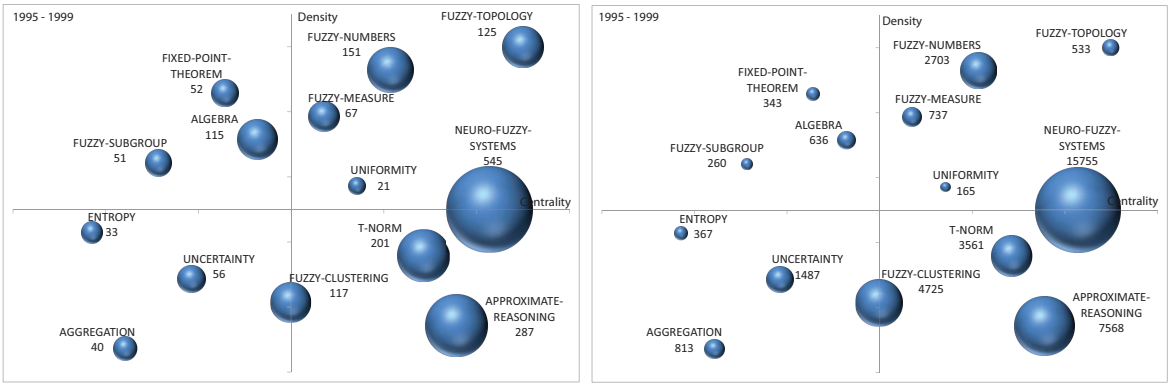
Por otro lado, del total de 1157 documentos analizados en este periodo de tiempo,

el 67 % de ellos (774) pudieron asociarse con alguno de los temas detectados. Este incremento en el número de documentos asociados es debido a que en este periodo los documentos contienen un mayor número de palabras clave describiendo su contenido.

Nombre del tema	Número de documentos	Número de citas	Media de citas	Índice h
NEURO-FUZZY-SYSTEMS	205	4135	20.17	34
FUZZY-NUMBERS	194	3518	18.13	31
FUZZY-CONTROL	172	6076	35.33	40
FUZZY-RELATION	134	1407	10.50	21
T-NORM	127	3158	24.87	30
LEVEL-SUBGROUPS	83	583	7.02	12
NEIGHBORHOOD-SPACES	65	453	6.97	8
COMPACTNESS	65	604	9.29	11

Tabla 3.3: Medidas de rendimiento para los temas del periodo 1990-1994.

En el siguiente periodo, 1995-1999, observamos un mayor número de temas, consecuencia directa del aumento en el número de documentos analizados, los cuales fueron 1680. A la vista de la Figura 3.8 y de la Tabla 3.4, debemos resaltar que los temas motores y los temas básicos son los que alcanzaron unos mayores niveles de citación e impacto.



(a) Diagrama estratégico basado en el número de documentos publicados.

(b) Diagrama estratégico basado en el número de citas.

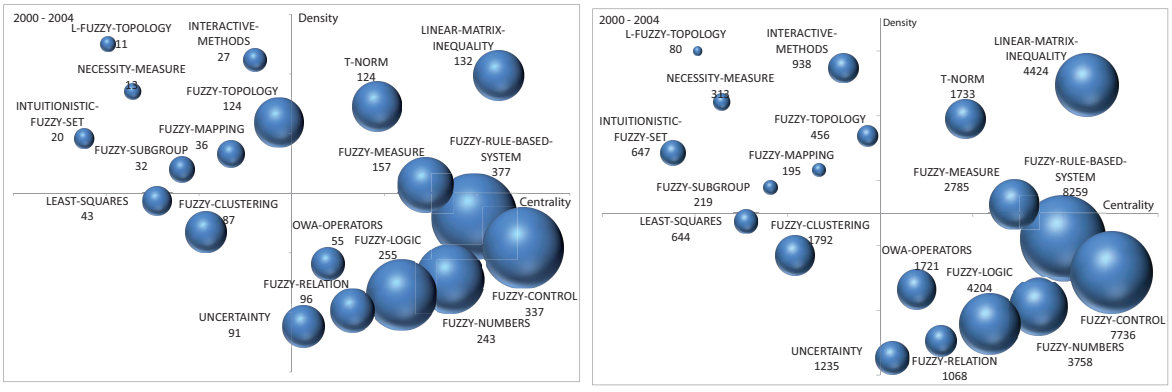
Figura 3.8: Diagramas estratégicos del periodo 1995-1999.

En este periodo de tiempo se consiguieron asociar con algún tema el 76 % (1281) de los documentos del periodo.

Nombre del tema	Número de documentos	Número de citas	Media de citas	Índice h
NEURO-FUZZY-SYSTEMS	545	15755	28.91	60
APPROXIMATE-REASONING	287	7568	26.37	45
T-NORM	201	3561	17.72	31
FUZZY-NUMBERS	151	2703	17.90	28
FUZZY-TOPOLOGY	125	533	4.26	11
FUZZY-CLUSTERING	117	4725	40.38	39
ALGEBRA	115	636	5.53	14
FUZZY-MEASURE	67	737	11.00	14
UNCERTAINTY	56	1487	26.55	20
FIXED-POINT-THEOREM	52	343	6.60	11
FUZZY-SUBGROUP	51	260	5.10	8
AGGREGATION	40	813	20.33	14
ENTROPY	33	367	11.12	10
UNIFORMITY	21	165	7.86	8

Tabla 3.4: Medidas de rendimiento para los temas del periodo 1995-1999.

En el periodo 2000-2004 (ver Figura 3.9 y Tabla 3.5) podemos observar un gran número de temas, estando situados la mayor parte de ellos en el cuadrante inferior-derecho (temas básicos y transversales). Es decir, hay una gran cantidad de temas básicos en este periodo. Además, en comparación con el periodo anterior esta cantidad es notablemente superior. De un modo similar al periodo anterior, los temas básicos y temas motores de este periodo son los más citados y con mayor impacto. Además, del total de documentos del periodo se consiguió asociar el 81 % de ellos.



(a) Diagrama estratégico basado en el número de documentos publicados.

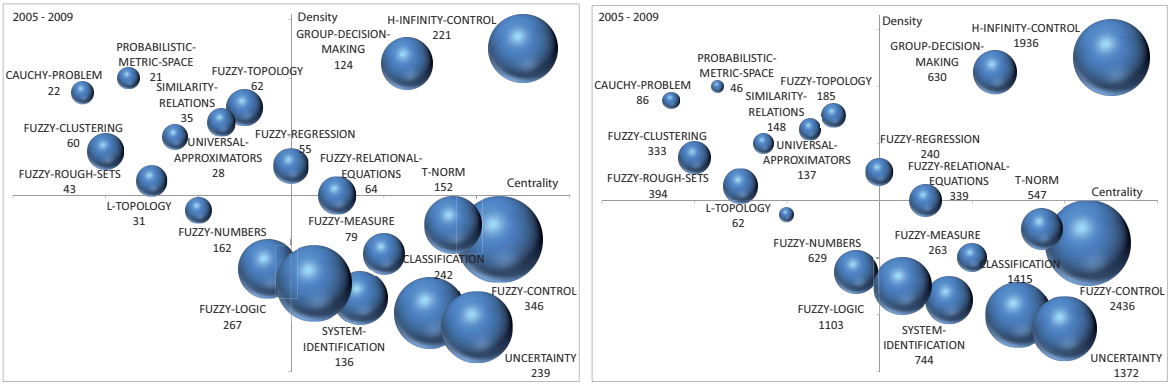
(b) Diagrama estratégico basado en el número de citas.

Figura 3.9: Diagramas estratégicos del periodo 2000-2004.

Nombre del tema	Número de documentos	Número de citas	Media de citas	Índice h
FUZZY-RULE-BASED-SYSTEM	377	8259	21.91	43
FUZZY-CONTROL	337	7736	22.96	46
FUZZY-LOGIC	255	4204	16.49	31
FUZZY-NUMBERS	243	3758	15.47	31
FUZZY-MEASURE	157	2785	17.74	24
LINEAR-MATRIX-INEQUALITY	132	4424	33.52	34
FUZZY-TOPOLOGY	124	456	3.68	9
T-NORM	124	1733	13.98	22
FUZZY-RELATION	96	1068	11.13	18
UNCERTAINTY	91	1235	13.57	20
FUZZY-CLUSTERING	87	1792	20.60	23
OWA-OPERATORS	55	1721	31.29	23
LEAST-SQUARES	43	644	14.98	16
FUZZY-MAPPING	36	195	5.42	9
FUZZY-SUBGROUP	32	219	6.84	9
INTERACTIVE-METHODS	27	938	34.74	12
INTUITIONISTIC-FUZZY-SET	20	647	32.35	13
NECESSITY-MEASURE	13	313	24.08	6
L-FUZZY-TOPOLOGY	11	80	7.27	5

Tabla 3.5: Medidas de rendimiento para los temas del periodo 2000-2004.

En el último periodo de tiempo analizado, 2005-2009, tal y como se muestra en la Figura 3.10 y en la Tabla 3.6, los temas básicos y los temas motores alcanzaron unos mayores niveles de impacto y de citación, repitiéndose el patrón de los últimos periodos analizados.



(a) Diagrama estratégico basado en el número de documentos publicados.

(b) Diagrama estratégico basado en el número de citas.

Figura 3.10: Diagramas estratégicos del periodo 2005-2009.

En este periodo el 84 % de los documentos (1325) pudo asociarse con algún tema.

Como con conclusión general, cabe destacar que en todos los periodos de tiempo analizados, los temas básicos lograron las mayores cotas de citación e impacto. Este hecho nos indica que los temas identificados a través de la metodología presentada en la Sección 3.1 son consistentes, ya que es lógico pensar que los temas considerados

como básicos y transversales tienen una mayor probabilidad de atraer la atención de la comunidad y por tanto de ser altamente citados.

Nombre del tema	Número de documentos	Número de citas	Media de citas	Índice h
FUZZY-CONTROL	346	2436	7.04	23
FUZZY-LOGIC	267	1103	4.13	16
CLASSIFICATION	242	1415	5.85	18
UNCERTAINTY	239	1372	5.74	18
H-INFINITY-CONTROL	221	1936	8.76	24
FUZZY-NUMBERS	162	629	3.88	12
T-NORM	152	547	3.60	11
SYSTEM-IDENTIFICATION	136	744	5.47	14
GROUP-DECISION-MAKING	124	630	5.08	13
FUZZY-MEASURE	79	263	3.33	8
FUZZY-RELATIONAL-EQUATIONS	64	339	5.30	9
FUZZY-TOPOLOGY	62	185	2.98	6
FUZZY-CLUSTERING	60	333	5.55	11
FUZZY-REGRESSION	55	240	4.36	9
FUZZY-ROUGH-SETS	43	394	9.16	10
SIMILARITY-RELATIONS	35	148	4.23	8
L-TOPOLOGY	31	62	2.00	4
UNIVERSAL-APPROXIMATORS	28	137	4.89	7
CAUCHY-PROBLEM	22	86	3.91	5
PROBABILISTIC-METRIC-SPACE	21	46	2.19	4

Tabla 3.6: Medidas de rendimiento para los temas del periodo 2005-2009.

3.2.2. Evolución de los Temas del Área de Teoría y Conjuntos Difusos

En esta sección, estudiamos y analizamos la evolución temática y conceptual del campo científico TCD a través de las áreas temáticas. En primer lugar, estudiamos el solapamiento, continuidad y discontinuidad de las palabras clave a lo largo de los distintos periodos de tiempo estudiados. Posteriormente, mostramos la evolución de los temas detectados en la sección anterior.

En cada periodo de tiempo, el conjunto de palabras clave que describe a los documentos no ha sido el mismo, tanto en un sentido lexicográfico como en un sentido numérico. Es decir, la terminología de TCD ha evolucionado a lo largo de los perio-

dos, utilizando diferentes palabras clave para describir el contendido de los documentos científicos publicados, surgiendo nuevos términos, y desapareciendo otros. Por otro lado, ha habido un subconjunto de palabras clave que se ha mantenido constante durante algunos periodos de tiempo consecutivos, en cambio otros subconjuntos sólo se han utilizado en algún periodo. Por ejemplo, las palabras clave *fuzzy-control*, *fuzzy-topology* y *neuro-fuzzy-systems* aparecen en todos los periodos de tiempo estudiados. Por el contrario, la palabra clave *multi-valued-logic* exclusivamente apareció en el primer periodo estudiado (1978-1989).

Siguiendo la filosofía de Price [81], en la Figura 3.11 mostramos la evolución de las palabras clave del área. Los círculos representan cada periodo de tiempo, además, el número dentro de ellos representa el total de palabras clave del periodo. Las flechas entre dos periodos consecutivos representan el número de palabras clave compartidas por ambos periodos, y entre paréntesis se muestra el índice de estabilidad (fracción de solapamiento). Las flechas entrantes representan el número de palabras clave nuevas del periodo. Finalmente, las flechas salientes representan las palabras clave que no han sido utilizadas por el siguiente periodo de tiempo, es decir, el número de palabras clave que no tienen continuación temporal en el periodo de tiempo inmediatamente posterior.

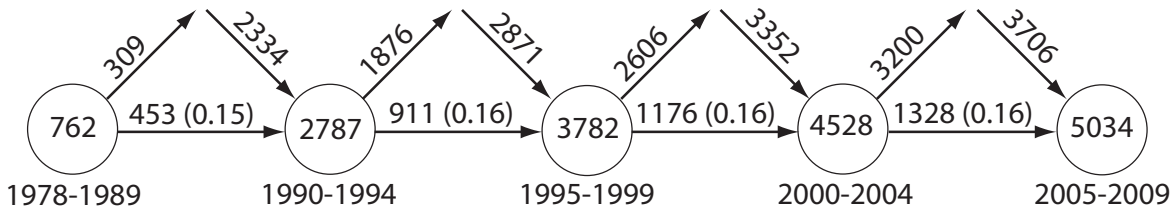


Figura 3.11: Continuidad de palabras clave entre periodos contiguos.

Como ejemplo, el tercer periodo analizado (1995-1999) contiene 3782 palabras clave, de ellas, 1176 palabras clave permanecieron en el periodo siguiente (2000-2004). Por el contrario, las restantes 2606 palabras clave no se utilizaron el siguiente periodo. El

índice de similaridad entre el tercer y cuarto periodo es 0.16.

El número de palabras clave se ha incrementado drásticamente a lo largo de los años; de hecho, el último periodo contenía seis veces más palabras clave que el primer periodo. Asimismo, el número de palabras clave compartidas entre periodos consecutivos también creció, pasando de 453 entre el primer y segundo periodo, a 1328 entre el cuarto y quinto periodo. Por otro lado, el índice de estabilidad se ha mantenido constante a lo largo del tiempo, lo que nos indica que la comunidad ha ido afianzando su vocabulario para describir a los documentos publicados. Finalmente, el número de palabras clave que no han sido utilizadas en el periodo contiguo es también elevado, lo que implica que hay un gran número de palabras clave transversales que sólo son usadas en un periodo y nunca más se vuelven a utilizar. Por ejemplo, en el cuarto periodo (2000-2004), existen un total de 3200 palabras clave efímeras, o discontinuadas, de un total de 4528. De hecho, el número de palabras clave efímeras es casi similar al número de palabras clave nuevas del periodo.

Una vez que hemos descrito y analizado la evolución de las palabras clave a lo largo de los años, estudiaremos la evolución temática del área científica TCD, mediante la utilización de la técnica de visualización descrita en la Sección 3.1.3: las áreas temáticas.

En la Figura 3.12 mostramos la evolución temática del área científica TCD. Al igual que se describió anteriormente (Sección 3.1.3), las líneas sólidas representan, o bien que ambos temas comparten el nombre, o bien que el nombre de uno de ellos es parte del otro. Es decir, que el nombre de alguno de los temas enlazados pertenece al nexo conceptual o temático. Por otro lado, una línea discontinua nos indica que entre los elementos de los temas enlazados no se encuentra los nombres de ambos temas. Finalmente, el grosor de cada enlace es proporcional al índice de inclusión de los temas enlazados; el volumen de las esferas es proporcional al número de documentos asociados

a cada tema.

Aunque el grafo de la Figura 3.12 es muy denso, pudimos detectar correctamente un conjunto significativo de áreas temáticas. En la Figura 3.12 los diferentes sombreados agrupan a los temas que pertenecen a la misma área temática. En este caso, existen temas que pertenecen a más de una zona sombreada, lo que implica que existen temas que pertenecen a más de un área temática. Por otro lado, hay temas que no están bajo ninguna zona sombreada, lo que implica que no pertenecen a ningún área temática.

Una vez identificadas las diversas áreas temáticas y los temas asociados a cada una de ellas, pudimos calcular sus medidas de rendimiento y de impacto. En concreto, en la Tabla 3.7 se muestran los nombres de cada área temática (aunque el nombre en este caso se corresponde con algún tema del área, se ha elegido manualmente), así como el número de documentos (calculado como la unión de los documentos principales y documentos secundarios de cada tema del área temática) y medidas de calidad e impacto.

Analizando detalladamente los resultados mostrados en la Figura 3.12 y en la Tabla 3.7, cabe resaltar los siguientes aspectos:

Área temática	Número de documentos	Número de citas	Media de citas	Índice h
FUZZY-CONTROL	2461	49726	20.21	92
FUZZY-LOGIC	1217	24477	20.11	69
FUZZY-NUMBER	1008	13896	13.79	50
T-NORM	604	8999	14.90	44
FUZZY-TOPOLOGY	581	3678	6.33	28
FUZZY-RELATION	447	4679	10.47	33
UNCERTAINTY	386	4094	10.61	29
GROUP-DECISION-MAKING	219	3164	14.45	29
FUZZY-SUBGROUP	166	1062	6.40	16
FUZZY-MAPPING	128	991	7.74	16

Tabla 3.7: Medidas cuantitativas y de impacto de las áreas temáticas detectadas (1978-2009).

- Si observamos el desarrollo del campo científico TCD de acuerdo a sus áreas temáticas y temas, podemos concluir que éste presenta una gran cohesión. Esto es debido a que la mayor parte de los temas identificados pudieron asociarse con algún área temática y estos a su vez provenían de algún tema (en algunos casos

hasta de varios) del periodo anterior. Por otro lado, encontramos temas que no estaban asociados con ningún área temática lo que es debido a: i) el tema es muy reciente y podría considerarse como el comienzo de un área temática nueva, por ejemplo, como ocurre con los temas *INTUITIONISTIC-FUZZY-SET* o *FUZZY-ROUGH-SET* en el cuarto y quinto periodo; ii) el tema está conectado con muchas áreas temáticas (es un tema básico) y es difícil clasificarlo en algún área, por ejemplo, como ocurre con el tema *DECISION-MAKING* del primer periodo; o iii) el tema no contiene unas palabras clave los suficientemente descriptivas como para poder detectar sus conexiones con otros temas, tal y como ocurre con el tema *RELATIONS* del primer periodo.

- La mayor parte de las áreas temáticas evolucionan de forma continua desde sus comienzos hasta el último periodo analizado. Es decir, no hay huecos en su evolución, lo que nos indica que atrajeron el interés de los miembros de la comunidad difusa a lo largo de todos los periodos estudiados. Aunque esto es cierto para la mayoría de los casos, no se cumple para el área *FUZZY-SUBGROUP* (Figura 3.18), la cual desaparece tras el cuarto periodo.

- Con respecto a la evolución del número de documentos, fijándonos en el volumen de las esferas, detectamos que la mayor parte de las áreas temáticas evolucionaron incrementalmente, es decir, aumentaron su número de documentos. De hecho, detectamos un interés creciente por parte de la comunidad difusa en dichas áreas temáticas debido al incremento de trabajo y por lo tanto, documentos publicados. De nuevo, la evolución del área temática *FUZZY-SUBGROUP* no presenta este comportamiento, al igual que ocurre con el área temática *FUZZY-MAPPING* (el detalle de su evolución se muestra en la Figura 3.15).

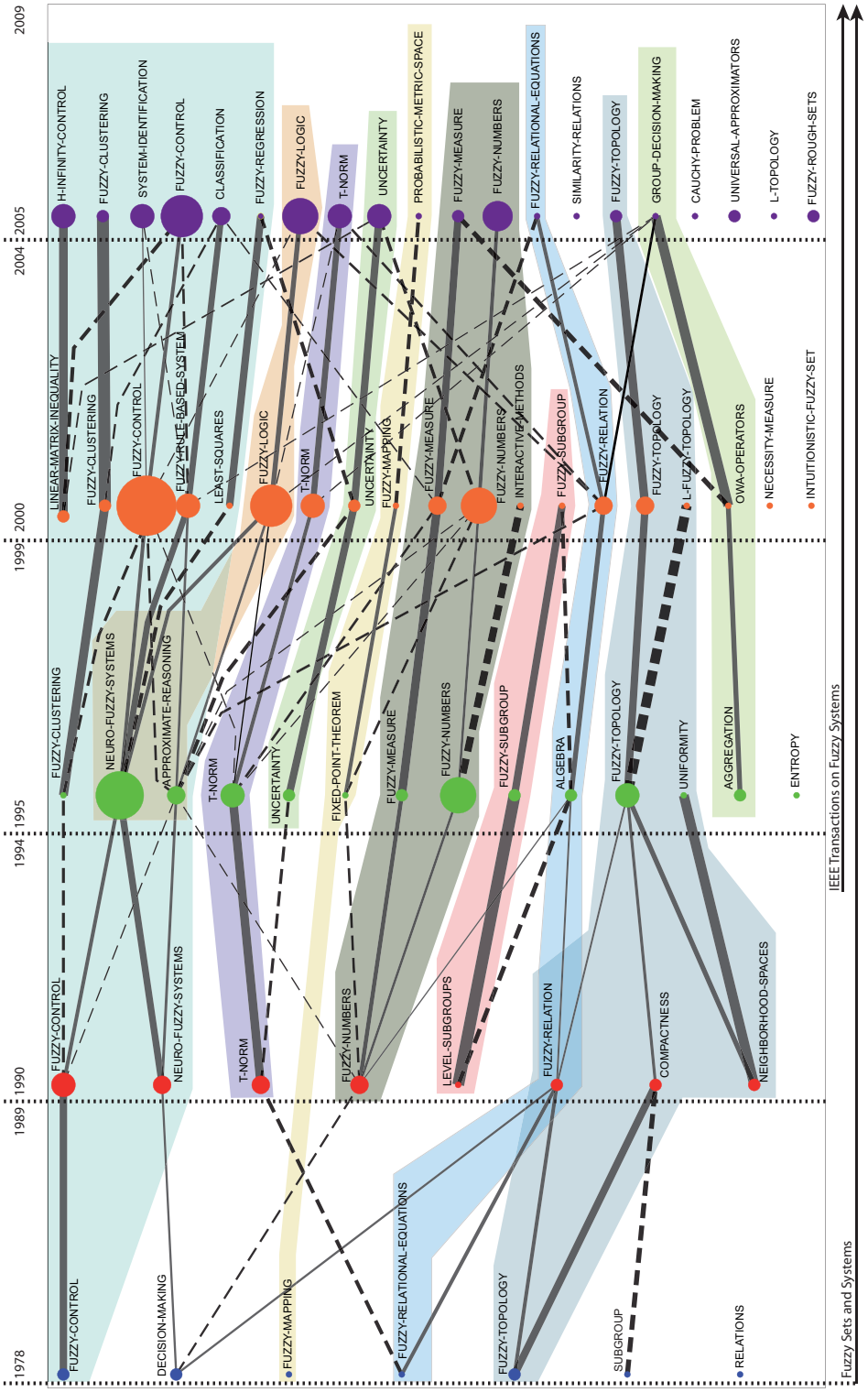


Figura 3.12: Evolución temática del área de Teoría y Conjuntos Difusos (1978-2009).

- Atendiendo a la evolución del numero de temas, encontramos que sólo el área temática *FUZZY-CONTROL* (el detalle de su evolución se muestra en la Figura 3.13) evoluciona de forma incremental, es decir, aumenta el número de temas a los largo de los años. Esta área temática es el origen de otra área temática importante: *FUZZY-LOGIC*. El resto de las áreas temáticas evolucionan de modo constante, como *T-NORM* (el detalle de su evolución se muestra en la Figura 3.21) o *GROUP-DECISION-MAKING* (el detalle de su evolución se muestra en la Figura 3.20), o de modo decreciente, como *FUZZY-LOGIC* (el detalle de su evolución se muestra en la Figura 3.14).

- Además, debemos resaltar que existen sólo tres áreas temáticas presentes a lo largo de todos los periodos, es decir, con temas en cada uno de ellos: *FUZZY-CONTROL*, *FUZZY-TOPOLOGY* (Figura 3.19) y *FUZZY-RELATION* (Figura 3.17). De hecho, podemos afirmar que estas tres áreas temáticas han mantenido el interés de la comunidad difusa en todos los periodos de tiempo estudiados, pero claramente, atendiendo a los resultados de la Tabla 3.7, el área temática con mayor número de temas, *FUZZY-CONTROL*, es la que presenta mejores indicadores de calidad (citas e índice h).

- Con respecto a lo composición temática de cada área temática encontramos que existen:

  - Dos áreas temáticas sólidas, las cuales están compuestas por temas motores y temas básicos en cada uno de los periodos de tiempo: *FUZZY-CONTROL*, *T-NORM*.

  - Dos áreas temáticas importantes que muestran signos de agotamiento, debido a que están compuestas por temas básicos y temas motores en los primeros

subperiodos, pero en el último, están compuestas por temas en extinción: *FUZZY-LOGIC* y *FUZZY-NUMBERS* (Figura 3.16).

- Dos áreas temáticas específicas o periféricas: *FUZZY-SUBGROUP* y *FUZZY-MAPPING*.
- Tres áreas temáticas en auge. Es decir, compuestas inicialmente por temas específicos y que en el último periodo empezaron a consolidarse como temas motores o temas básicos: *FUZZY-RELATIONS*, *UNCERTAINTY* (Figura 3.22) y *GROUP-DECISION-MAKING*.
- Existe sólo un área temática en descenso: *FUZZY-TOPOLOGY*.

- En resumen, cabe resaltar los siguientes aspectos:

1. *FUZZY-CONTROL* es el área temática más importante en el campo científico TCD, presentando el mejor comportamiento evolutivo, y los mejores indicadores bibliométricos de acuerdo a los valores de la tabla 3.7.
2. *T-NORM* es también un área temática importante y a la vez básica en el campo TCD, la cual presenta una sólida evolución y un buen valor de impacto (índice h=44), de acuerdo a los valores de la Tabla 3.7.
3. *FUZZY-LOGIC* y *FUZZY-NUMBERS*, fueron también áreas temáticas importantes, las cuales presentan buenos indicadores de impacto (índice h=69 y índice h=50, respectivamente), pero que actualmente presentan síntomas de agotamiento.
4. *FUZZY-RELATIONS*, *UNCERTAINTY* y *GROUP-DECISION-MAKING* son tres áreas temáticas en ascenso, y además presentan buenos indicadores de calidad: índice h=33, índice h=29, índice h=29, respectivamente.

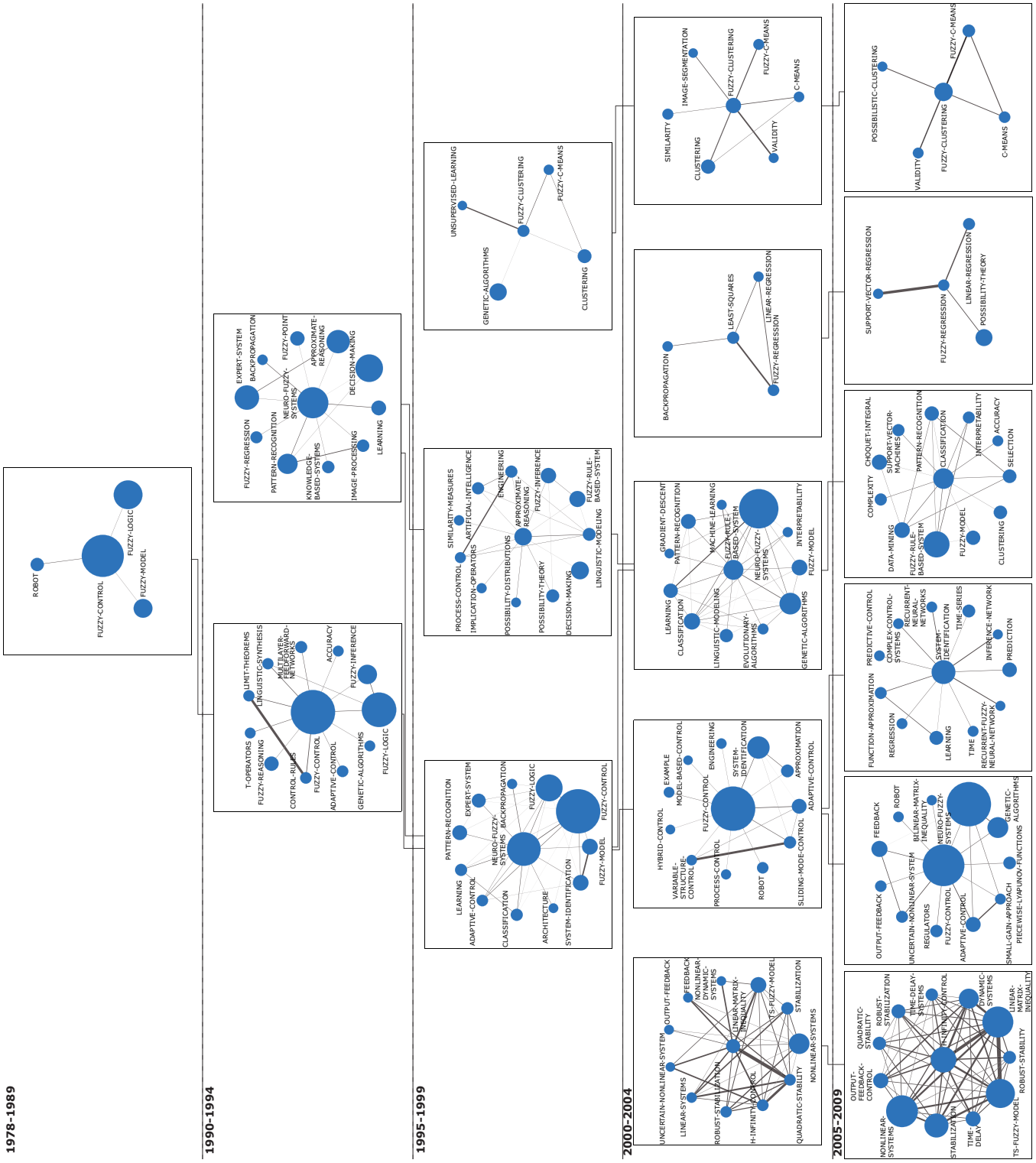


Figura 3.13: El área temática FUZZY-CONTROL (1978-2009).

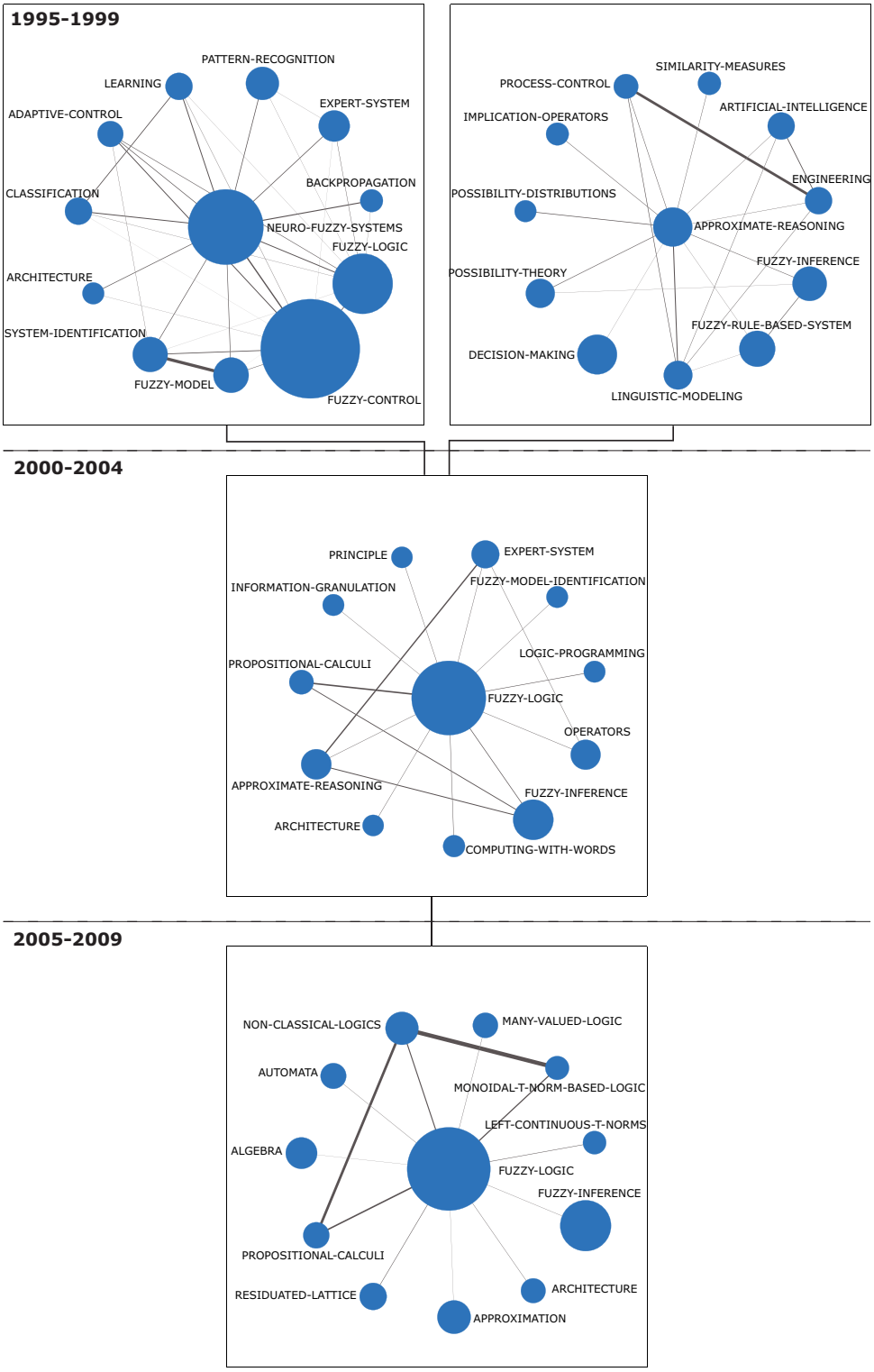


Figura 3.14: El área temática FUZZY-LOGIC (1978-2009).

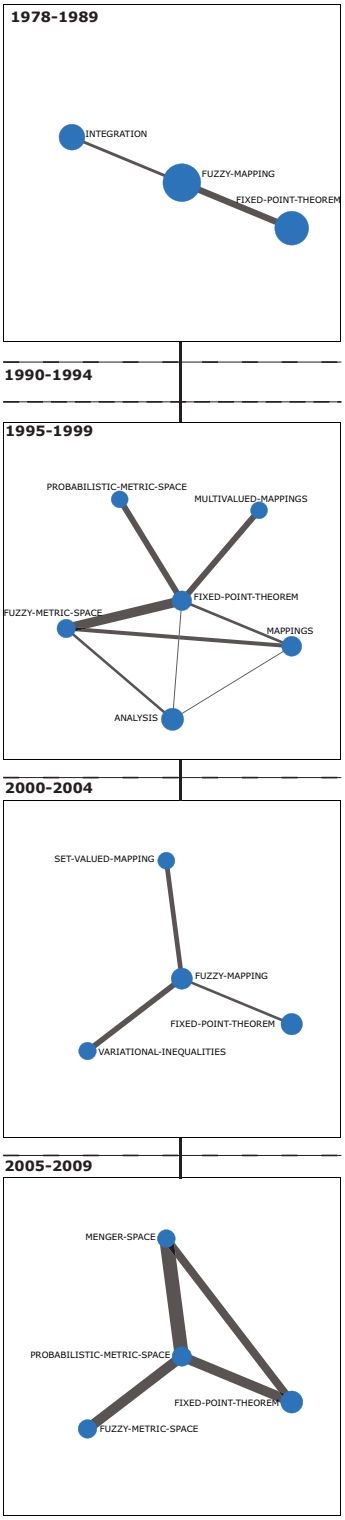


Figura 3.15: El área temática FUZZY-MAPPING (1978-2009).

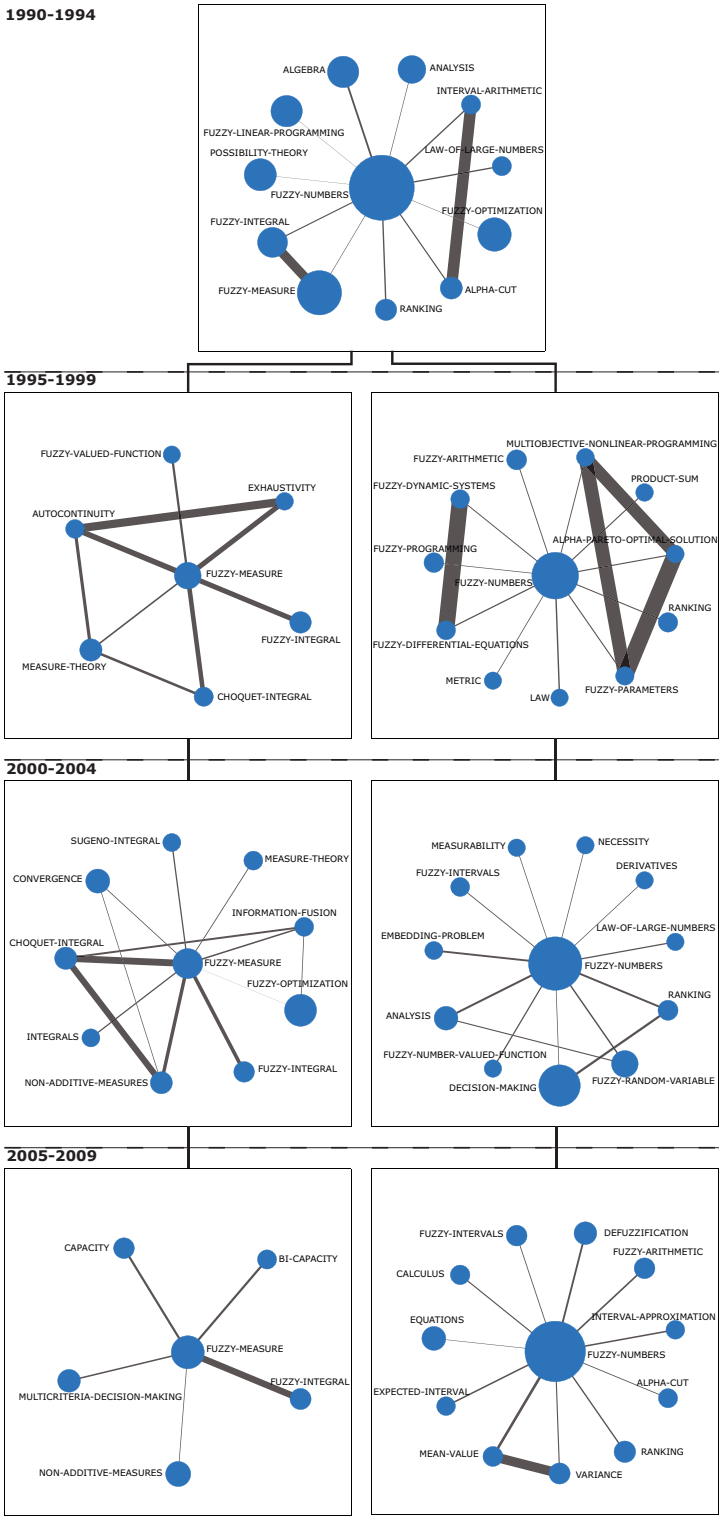


Figura 3.16: El área temática FUZZY-NUMBER (1978-2009).

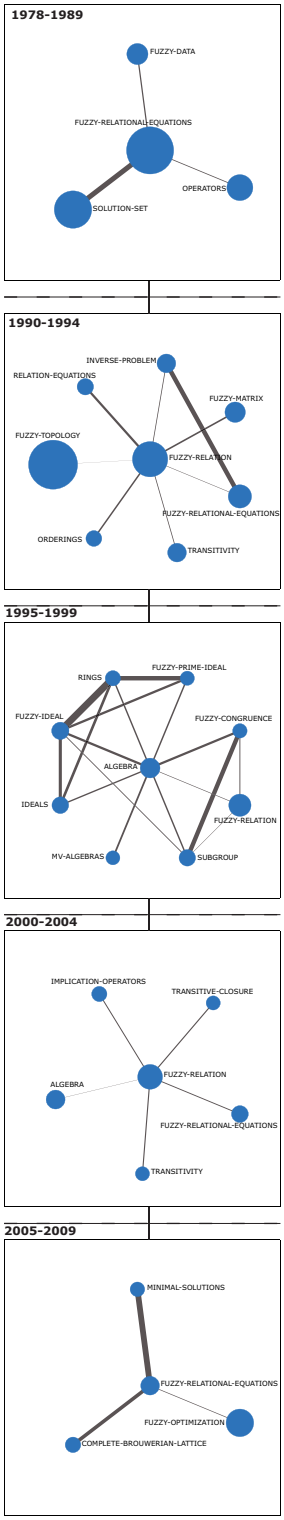


Figura 3.17: El área temática FUZZY-RELATION (1978-2009).

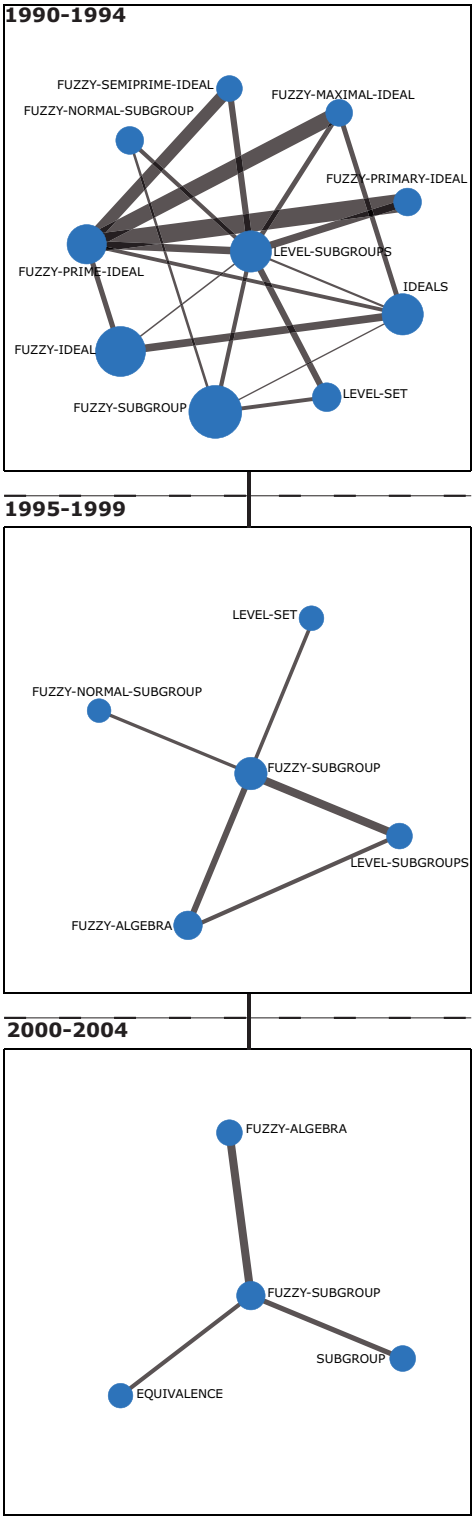


Figura 3.18: El área temática FUZZY-SUBGROUP (1978-2009).

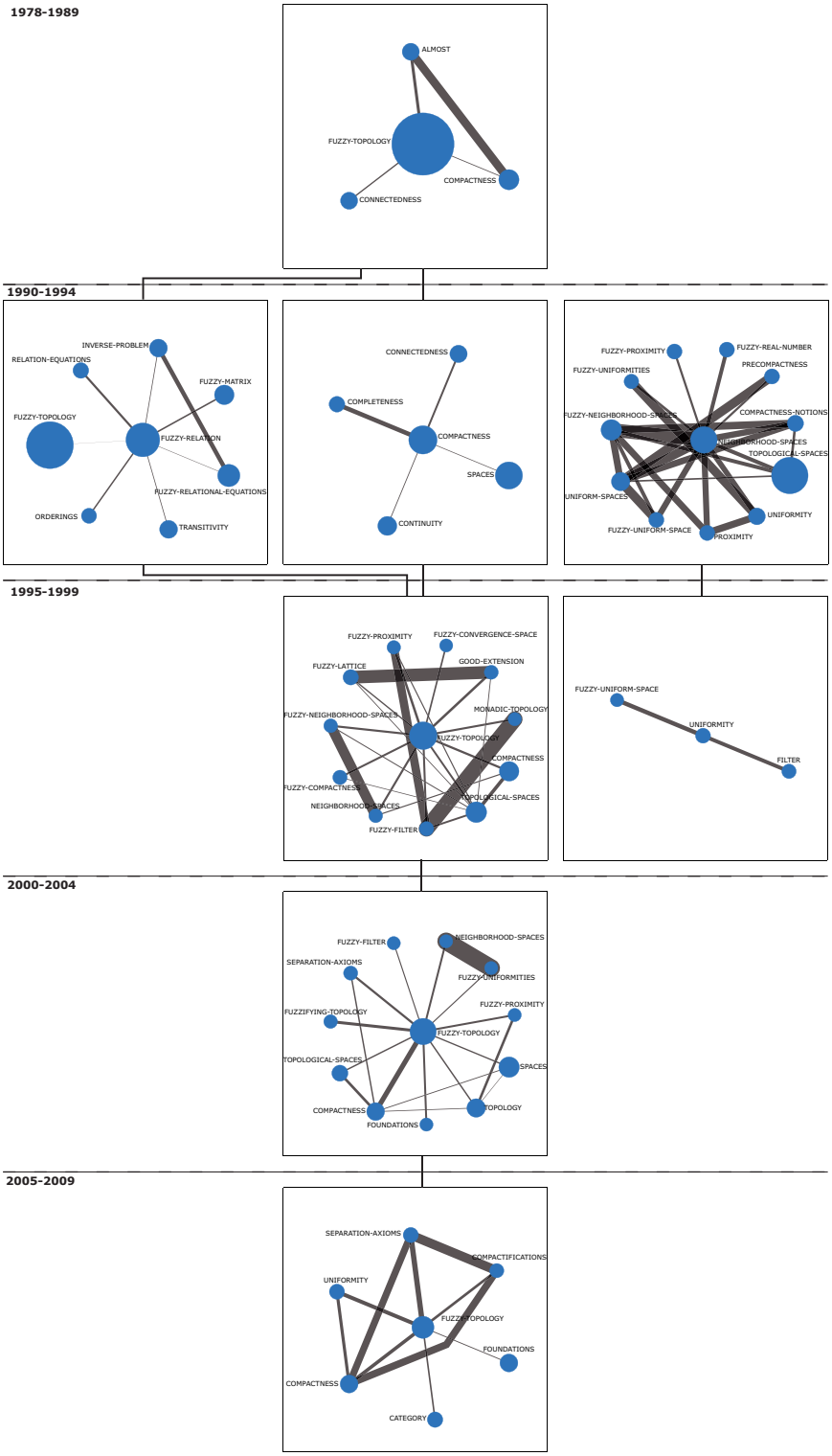


Figura 3.19: El área temática FUZZY-TOPOLOGY (1978-2009).

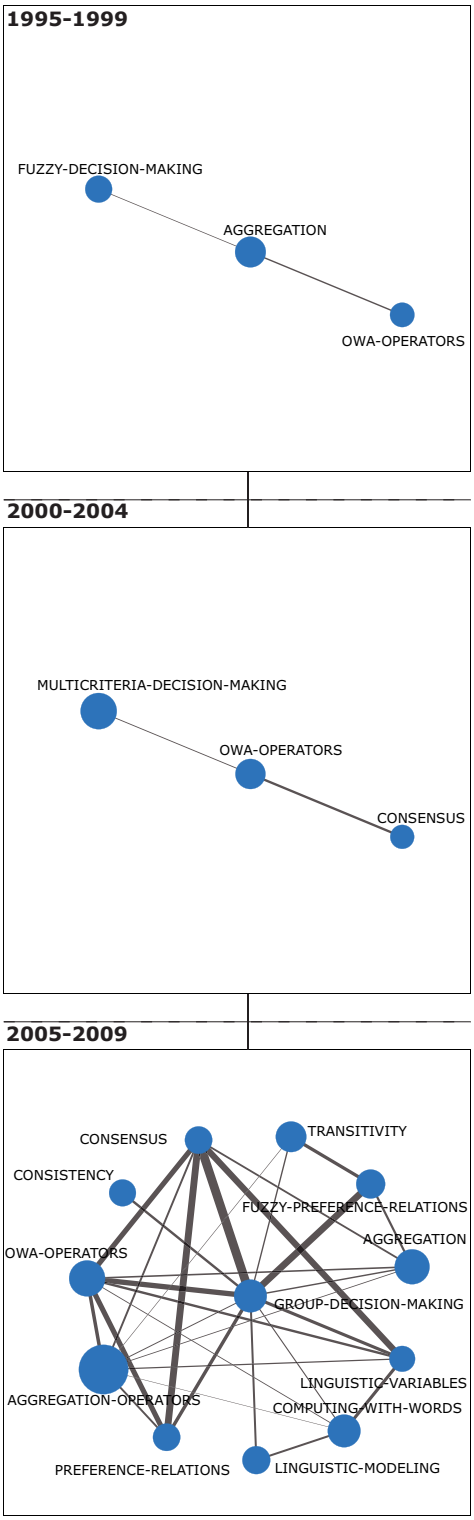


Figura 3.20: El área temática GROUP-DECISION-MAKING (1978-2009).

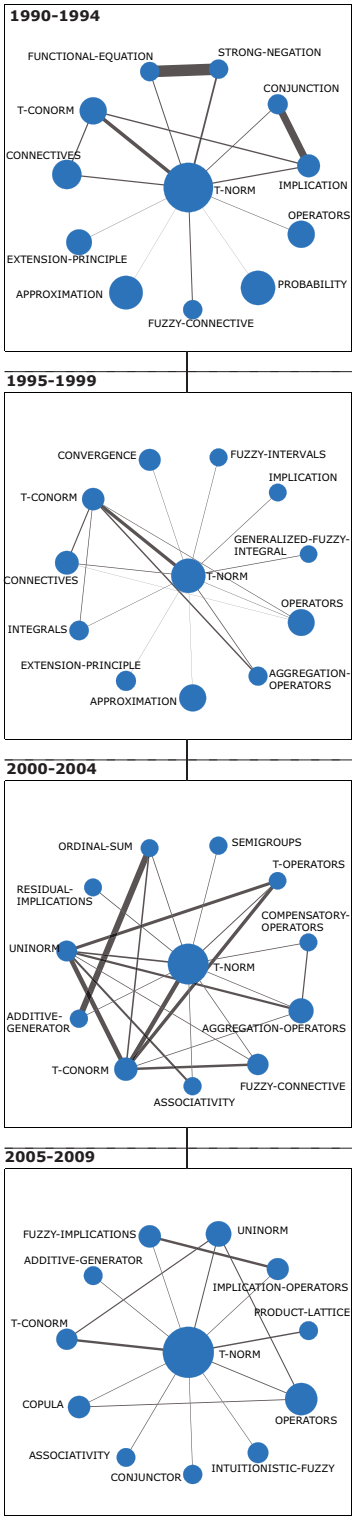


Figura 3.21: El área temática T-NORM (1978-2009).

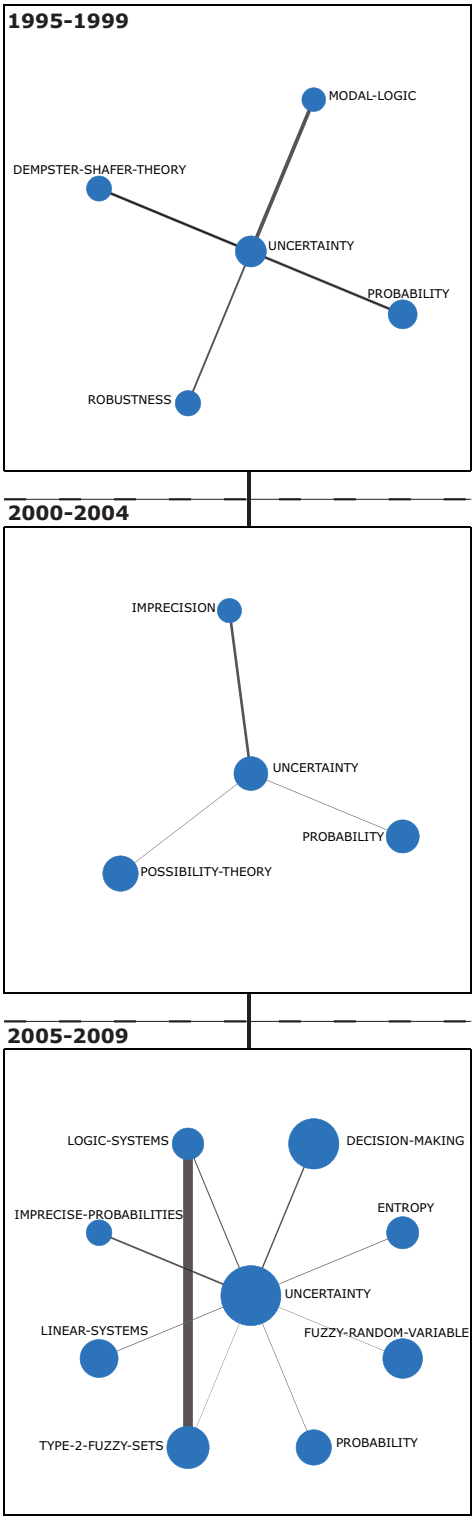


Figura 3.22: El área temática UNCERTAINTY (1978-2009).

5. *FUZZY-TOPOLOGY* no es un área temática muy importante, tiene un comportamiento decreciente y además presenta unos valores de calidad bajos (índice h=28). De todos modos, ha estado presente en todos los periodos de tiempo analizados.

6. *FUZZY-SUBGROUP* y *FUZZY-MAPPING* son áreas temáticas periféricas, las cuales presentan los indicadores bibliométricos más bajos de acuerdo con los valores de la Tabla 3.7.

7. Por último, identificamos dos temas que podrían ser el origen de nuevas áreas temáticas en el futuro: *INTUITIONISTIC-FUZZY-SETS* y *FUZZY-ROUGH-SETS*.

Entre los años comprendidos en el estudio, 1978-2009, el total del corpus analizado logró un índice h de 115. Como cabe esperar, no todas las áreas temáticas contribuyen equitativamente a este indicador de calidad, ya que algunas de ellas contienen un gran número de documentos altamente citados, y otras contiene documentos pobremente citados. Concretamente, 60 de los documentos del área temática *FUZZY-CONTROL* pertenecen al núcleo de los documentos del índice h global, por lo que casi la mitad de los documentos altamente citados de la comunidad TCD pertenecen a esta área temática. Otras áreas temáticas importantes fueron *FUZZY-LOGIC* y *FUZZY-NUMBER*, las cuales contribuyeron con 31 y 11 documentos altamente citados, respectivamente. Las áreas temáticas restantes contribuyeron con menos de 10 documentos altamente citados, por ejemplo, el área temática *GROUP-DECISION-MAKING* contribuyó con 3 documentos muy citados. Por otro lado, las áreas temáticas *FUZZY-MAPPING* y *FUZZY-SUBGROUP* no contribuyeron con ningún documento, lo que confirma que nuestra metodología identifica de forma adecuada a los temas periféricos y especializados.

Es necesario comentar que los conjuntos de documentos pertenecientes a cada área temática no son disjuntos, debido a que un mismo documento puede pertenecer a más de un tema (motivado por el uso conjunto de los documentos importantes + documentos secundarios), y un mismo tema puede pertenecer a diferentes áreas temáticas. Por este motivo, los documentos pertenecientes al núcleo del índice h global pertenezcan a su vez a varias áreas temáticas. Por otro lado, es también posible que un documento altamente citado no pertenezca a ningún área temática.

3.2.3. ¿Qué nos Indica el Análisis Realizado?

Como consecuencia directa de la aplicación de la metodología presentada en este capítulo al campo científico TCD, debemos remarcar varios aspectos:

1. *Aspectos técnicos de nuestra metodología:*

   - Esta metodología combina diferentes técnicas bibliométricas para analizar la evolución de la estructura cognitiva y conceptual de un campo científico, permitiéndonos descubrir conocimiento importante relacionado con sus temas y áreas temáticas. De este modo, tal y como se señaló en la Sección 3.2.1, nuestra metodología fue capaz de identificar adecuadamente los temas básicos del campo científico TCD, para cada uno de los periodos de tiempo analizados, que se correspondían con aquellos que lograron unas mayores cotas de citación e impacto. Adicionalmente, como se mostró en la Sección 3.2.2, nuestra metodología es capaz de identificar las áreas temáticas (ver la Tabla 3.7), mostrando su evolución a lo largo del tiempo, como es el caso de *FUZZY-CONTROL*, la cual tuvo un comportamiento incremental, o *FUZZY-LOGIC*, que drecrementó con el paso de los años.

- Nuestra metodología está soportada por diferentes técnicas de visualización que permiten identificar fácilmente los temas y áreas temáticas y comprender su evolución, importancia y futuras tendencias, lo que ayudará en la toma de decisiones. Por ejemplo, hemos mostrado la evolución del campo TCD en la Figura 3.12, identificando que *FUZZY-CONTROL* es el área temática más importante y con el mayor impacto de acuerdo a los valores de la Tabla 3.7. De un modo similar, hemos concluido que *FUZZY-ROUGH-SETS* parece ser el origen de una nueva área temática.
- La metodología se completa mediante la incorporación de indicadores bibliométricos complejos, que nos ayudaran a analizar de un mejor modo la calidad e impacto de los temas y áreas temáticas. En concreto, en nuestro análisis sobre el campo TCD, utilizamos el índice h para evaluar el impacto tanto de áreas temáticas, como de temas.

2. *Aspectos relacionados con la aplicación de la metodología a un campo científico concreto*: la aplicación de nuestra metodología para analizar la evolución del campo TCD ha sido muy efectiva, permitiéndonos extraer, de un modo sencillo, información en cada uno de los periodos analizados. Además, muchos de los resultados obtenidos pueden ser entendidos por una gran cantidad de usuarios gracias a las técnicas de visualización empleadas.

3.3. Extensión y Generalización de la Metodología

La metodología presentada en este Capítulo puede ser fácilmente extendida para, además de permitir realizar un análisis conceptual de un campo científico (a través de un análisis de co-palabras), permitir analizar otros aspectos de éste como aquellos

sociales (utilizando los autores como unidad de análisis) e intelectuales (utilizando las referencias como unidad de análisis). En concreto, la metodología permite generalizar y adaptar cada paso, en la forma en la que se comenta a continuación:

- La metodología permite utilizar cualquiera de las unidades de análisis descritas en la Sección 1.2.2 para construir la red bibliométrica.
- Del mismo modo, permite crear cualquier tipo de red bibliométrica a partir de los distintos tipos de relaciones entre unidades de análisis descritos en la Sección 1.2.2.
- Las redes bibliométricas pueden normalizarse empleándose cualquier tipo de medida de similitud, como las descritas en la Sección 1.2.4.
- El mapa científico puede crearse utilizando cualquier algoritmo de clustering que sea capaz de dividir una red o grafo (matriz de adyacencia) en diferentes subredes. Además, El algoritmo tiene que ser capaz de identificar el nodo más importante de cada subred. En el caso de que esto último no sea posible, el nodo más importante de cada subred se tendrá que establecer manualmente.
- Como medidas de rendimiento para medir la actividad, calidad e impacto de grupos[3] , redes y áreas de evolución, se puede utilizar cualquiera de los indicadores bibliométricos propuestos en la Sección 1.1.

Por ejemplo, eligiendo como unidad de análisis los autores, y como relación entre ellos, la co-ocurrencia, podemos realizar un análisis de coautorías para analizar los aspectos sociales de un campo científico, así como su evolución temporal. Mediante

[3] Dado que la generalización de la metodología permite analizar aspectos sociales, intelectuales y conceptuales, no tiene sentido hablar de temas, redes y áreas temáticas. En este caso, dicha terminología será reemplazada por grupos, redes y áreas de evolución, respectivamente.

esta configuración, los mapas científicos mostrarían los distintos grupos de autores que han trabajado conjuntamente en los diversos periodos de tiempo analizados.

Capítulo 4

SciMAT: una Herramienta para el Análisis de Mapas Científicos Enriquecidos con Medidas Bibliométricas

En los capítulos anteriores, presentamos un análisis de nueve de las herramientas más representativas para realizar un análisis de mapas científicos (Capítulo 2), así como una propuesta metodológica para la creación y visualización de mapas científicos en un contexto longitudinal (Capítulo 3). Tal y como se detalló en la Sección 3.3, la metodología puede extenderse de un modo fácil e intuitivo para utilizar cualquier tipo de unidad de análisis y red bibliométrica, para de este modo, analizar cualquier aspecto de un campo científico.

En este capítulo presentamos una nueva herramienta para el análisis de mapas científicos llamada SciMAT. Esta herramienta implementa la metodología desarrollada en el Capítulo 3, por lo que nos permitirá analizar la evolución de los aspectos sociales, intelectuales y conceptuales de un campo científico.

El presente capítulo se estructura de la siguiente forma: en la Sección 4.1 situamos en contexto a SciMAT, para posteriormente describir sus características fundamentales,

en la Sección 4.2 describimos el diseño conceptual y físico de la base de conocimiento de SciMAT, en la Sección 4.3 describimos los módulos, métodos y algoritmos que incorpora SciMAT. Por último, en la Sección 4.4 realizamos un estudio guiado con SciMAT para demostrar su potencial.

4.1. SciMAT: Contexto y Descripción

Gracias al análisis comparativo de las nueve herramientas analizadas en el Capítulo 2, pudimos detectar ciertas carencias en ellas, así como descubrir sus principales ventajas, desventajas y diferencias. En concreto podemos resaltar los siguientes aspectos:

- Las herramientas disponibles para el análisis de mapas científicos, aunque tienen diferentes características, no implementan todos los aspectos necesarios para poder realizar un análisis completo. Por ejemplo, VOSViewer se centra principalmente en la fase de creación del mapa y en su visualización, por otro lado, Loet Leydesdorff's Software no implementa ninguna técnica de visualización.

- Atendiendo al preprocesamiento, concretamente a la unificación de elementos similares, CoPalRed y VantagePoint son las únicas herramientas que disponen de un buen módulo para realizar esta tarea. Otras herramientas como NWB y Sci^2, aunque disponen de un módulo para la unificación de elementos, este necesita la ayuda de una herramienta externa para poder realizarse correctamente. Por otro lado, herramientas como BibExcel o VOSViewer, no incluyen ningún tipo de preprocesamiento.

- Otra característica importante es la capacidad de la herramienta de extraer diferentes tipos de redes bibliométicas. Aunque, la mayor parte de herramientas permiten extraer un gran número de redes, no existe una única herramienta que

permita construir todos los tipos de redes. De hecho, algunas herramientas, como CoPalRed, se centran sólo en un tipo de red, mientas que otras herramientass, como VOSViewer, no son capaces de extraer ningún tipo.

- Cada herramienta emplea diferentes técnicas de visualización. Por ejemplo, VOSViewer utiliza mapas basados en proximidad, CoPalRed utiliza diagramas estratégicos para categorizar los temas detectados, y VantagePoint realiza la visualización utilizando diferentes técnicas entre las que se encuentra los mapas de factores.

- Además, estas herramientas no permiten establecer diferentes configuraciones para realizar el análisis de mapas científicos. De hecho, la mayoría de ellas sólo permiten seleccionar un tipo de análisis, un tipo de medida de similitud, un tipo de algoritmo de clustering, etc.

- Por último, tal y como señalamos en el Capítulo 3, sería adecuado que una herramienta para el análisis de mapas científicos fuera capaz de enriquecer los resultados utilizando medidas de calidad e impacto basadas en indicadores bibliométricos. Esto ayudaría al usuario a analizar el mapa y determinar el impacto y calidad de los diferentes resultados, para de este modo, indicar los aspectos más importantes del campo científico analizado. Por ejemplo, si hubiésemos realizado un análisis de mapas científicos basado en una red de co-palabras para analizar conceptualmente un campo científico, mediante la utilización de indicadores bibliométricos el analista podría medir cuales de los temas detectados habrían tenido una mayor repercusión en la comunidad científica, y por lo tanto, cuáles habrían sido más importantes en el campo científico estudiado. En la comparativa de herramientas realizada en el Capítulo 2, detectamos que CiteSpace es la única herramienta que

que permite utilizar las citas en su análisis, pero en cualquier caso, la utilización que hace de ellas es muy básica.

A la vista de esto, pensamos que sería muy adecuado definir una herramienta para el análisis de mapas científicos que satisfaga los siguientes resultados:

1. Debería incorporar módulos para realizar todos los pasos del flujo de trabajo de un análisis de mapas científicos.
2. Debería contener un buen módulo para la unificación de unidades similares.
3. Debería ser capaz de extraer el mayor número posible de redes bibliométricas.
4. Debería poseer buenas técnicas de visualización que ayudaran al analista a comprender e interpretar los resultados, para que de este modo, pudiera extraer conocimiento de los resultados.
5. Debería enriquecer los resultados utilizando indicadores bibliométricos de impacto y calidad, tales como: índice h [1, 46], índice g [34], índice hg [2], índice q^2 [17], etc.

En este sentido, surge SciMAT (*Science Mapping Analysis software Tool*), como una nueva herramienta gratuita y de código libre (*open source*) para la realización de análisis de mapas científicos. Integra los puntos fuertes de otras herramientas, mientras que reduce la dependencia de herramientas externas. Es decir, SciMAT integra todo lo necesario para realizar un análisis de mapas científicos completo, en un marco longitudinal y utilizando medidas bibliométricas de impacto. Además, la herramienta permite analizar la evolución social, intelectual y conceptual de un campo científico.

SciMAT puede utilizarse de forma gratuita, de hecho, puede modificarse y distribuirse de acuerdo a los términos de la licencia GPLv3[1] . SciMAT se encuentra disponible

[1] http://www.gnu.org/licenses/gpl-3.0.html

en Internet, a través de su portal web[2] , el cual nos permitirá descargarnos el archivo ejecutable, o ejecutar la aplicación directamente desde Internet, así como acceder a su manual de usuario.

Las principales características que describen a SciMAT y lo diferencian del resto de herramientas son:

- SciMAT incorpora todos los módulos necesarios para realizar todos los pasos del flujo de trabajo del análisis de mapas científicos, desde la carga de los datos, hasta la visualización e interpretación de estos. Además, la mayor parte de los pasos son configurables, de modo que permiten seleccionar diversos algoritmos y medidas.

- SciMAT incorpora métodos para extraer un gran número de redes bibliométricas, múltiples medidas para normalizar las redes, diferentes algoritmos de clustering, y diversas técnicas de visualización de gran utilidad para la interpretación de los resultados.

- SciMAT implementa diversas técnicas de preprocesamiento que nos permitirá detectar elementos similares que deben ser unificados, dividir los datos en diferentes periodos de tiempo, filtrar los datos para realizar el análisis con los datos más significativos, y filtrar las redes para quedarnos con las relaciones entre las unidades de análisis más importantes.

- SciMAT permite al analista realizar el análisis de mapas científicos bajo un marco longitudinal, para de este modo poder estudiar y detectar la evolución social, conceptual o intelectual de un campo científico a lo largo de periodos de tiempo consecutivos.

[2] http://sci2s.ugr.es/scimat

- Por último, de acuerdo con la metodología presentada en el Capítulo 3, SciMAT enriquece los mapas con medidas bibliométricas basadas en citas, como: índice h [1, 46], índice g [34], índice hg [2], índice q^2 [17], etc.

4.2. La Base de Conocimiento

SciMAT construye una base de conocimiento a partir de un conjunto de publicaciones científicas y almacena en ella las relaciones de cada publicación (documento) con las diferentes entidades (autores, palabras clave, revista, referencias, etc.) comprendidas en él. Esta base de datos ayudará al usuario a editar y preprocesar la información para de este modo, mejorar la calidad de los datos y consecuentemente, obtener unos mejores resultados en el análisis de mapas científicos.

En esta sección describimos detalladamente la base de conocimiento de SciMAT, profundizando primeramente en su diseño conceptual (Sección 4.2.1), y finalmente mostrando el diseño Entidad/Relación (E/R) de la base de datos que la soporta (Sección 4.2.2).

4.2.1. Definición Conceptual: Entidades

La base de conocimiento se compone de dieciséis entidades, siendo el *Documento* la principal entidad. Un Documento representa a una publicación científica, que normalmente suele ser un artículo, carta, revisión o un artículo procedente de un congreso. El Documento contiene una serie de información, como el título, resumen, doi, número de citas recibidas, etc. Además, los Documentos tienen una gran cantidad de piezas de información asociadas o relacionadas con ellos, como los autores y sus afiliaciones, las palabras clave, las referencias citadas en el documento, la revista o conferencia en

donde fue publicado el documento, así como, el año de publicación, siendo cada una de estas una entidad en la base de conocimiento.

El *Autor* es la entidad que representa a la persona envuelta en el desarrollo de un Documento. En este sentido, un Autor se encuentra relacionado con un conjunto de Documentos, y similarmente, un documento está asociado con un conjunto de autores. Además, la relación entre un Autor y un Documento tiene asociada una *posición* que representará el lugar que le corresponde al Autor en un Documento concreto.

La entidad *Afiliación* representa la afiliación de un autor, es decir, el lugar donde trabajaba cuando realizó el documento. Debido a que los autores pueden trabajar en lugares diferentes (universidades, institutos, centros de investigación, etc.) a lo largo de su vida profesional, un Autor puede tener asociadas un conjunto de Afiliaciones.

Usualmente, los documentos científicos tienen un conjunto de palabras clave asociadas con ellos. De hecho, dependiendo de la base de datos bibliográfica que hayamos usado para recopilar los datos, los documentos pueden contener palabras clave o descriptores añadidos por la base de datos. Por ejemplo, la base de datos ISIWoS añade un conjunto de palabras clave llamadas *ISI Keywords Plus* a cada documento. En este sentido, la entidad *Término* representa a un término descriptivo o palabra clave del documento. Los Términos pueden aparecer en múltiples Documentos, y de forma similar, un Documento puede tener asociados un conjunto de Términos. Además, cada Término tiene un rol específico en cada documento donde aparece, de modo que, un documento puede tener términos procedentes de los autores (palabras clave de autor), provenientes de la base de datos bibliográfica (palabras clave de la fuente de información), o añadidos por el usuario durante el preprocesamiento (términos extraídos).

La entidad *Referencia* representa la base intelectual de un documento científico, es decir, el conjunto de documentos citados por él. Al igual que los Términos, un Documen-

to tiene asociado un conjunto de Referencias y viceversa. Las referencias, normalmente, pueden dividirse en piezas de información más pequeñas, las cuales varían en función de la base de datos bibliográfica usada en la recopilación de los documentos. De todos modos, cierta información, como los autores, revista y año, suele aparecer siempre. Por esta razón, las Referencias tienen dos entidades asociadas: *Autor-Referencia* (*Autor de la Referencia*) y *Revista-Referencia* (*Revista de la Referencia*).

Otras entidades asociadas con un Documento son la *Revista* y la *Fecha de publicación*. Lógicamente, un Documento se encuentra asociado exclusivamente con una Revista (o conferencia) y una Fecha de publicación, mientras que ambas entidades pueden tener asociado un conjunto de Documentos. Además, ambas entidades tienen asociadas una *Categoría*, la cual representa una categoría general de conocimiento, dada usualmente por la base de datos bibliográfica, y que clasifica a la revista en categorías de conocimiento. Una Revista puede estar asociada con varias Categorías, y además esta relación puede variar a lo largo del tiempo.

La entidad *Periodo* representa un conjunto de años (Fechas de publicación), los cuales no tienen que ser consecutivos. De hecho, los periodos no tienen que ser disjuntos, por lo que varios años pueden estar asociados simultáneamente a varios periodos. Los periodos se utilizarán en el proceso del análisis de mapas científicos para analizar la evolución estructural del campo estudiado.

Finalmente, SciMAT permite utilizar cinco de las entidades descritas en el proceso de creación del mapa científico: Autor, Término, Referencia, Autor-Referencia y Revista-Referencia. Estas entidades deben de ser preprocesadas cuidadosamente, prestando especial atención a los errores ortográficos y a los elementos que representan a la misma entidad (proceso de unificación). Normalmente, el proceso de unificación de elementos similares une aquellos elementos parecidos en uno solo. Por ejemplo, supon-

gamos que tenemos dos elementos almacenados en la base de conocimiento: *Garfield, E.* y *Eugene Garfield*. Ambos elementos representan al mismo autor, y por lo tanto, deberían unificarse, uniendo también sus asociaciones con otras entidades.

En este sentido, cuando unimos dos elementos, sólo mantendremos uno de ellos en la base de conocimiento, y éste contendrá las asociaciones con otras entidades de ambos elementos. Esto hace imposible saber qué elementos fueron unidos ya que del grupo de elementos unificados exclusivamente mantendremos uno. Por esta razón, en la base de conocimiento de SciMAT incorporamos el concepto de *grupo* para cada unidad de análisis. Un grupo, es un conjunto de elementos que representan a la misma entidad. En este sentido, en la base de conocimiento existen cinco clase de grupos distintos: *Grupo de Autores*, *Grupo de Autor-Referencia*, *Grupo de Revista-Referencia*, *Grupo de Referencias* y, por último, *Grupo de Términos*. Además un grupo puede marcarse como *útil* o *no útil*, no utilizándose en el proceso en el segundo caso.

En la Figura 4.1 se muestra un ejemplo de grupos y cómo estos pueden ayudar en el proceso de unificación de elementos similares. En la parte izquierda, podemos observar los elementos antes de ser unificados, por otro lado, en la parte derecha podemos observar el grupo de elementos después de la unificación. La elipse sombreada representa al grupo (en este caso un Grupo de Autores), y las restantes elipses representan las entidades (Autores) asociadas con el grupo.

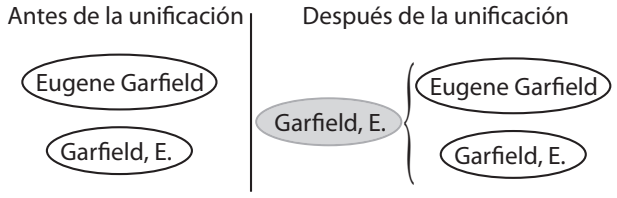


Figura 4.1: Grupo de autores después del proceso de unificación.

4.2.2. Diseño Entidad/Relación de la Base de Conocimiento

Una vez descrito el modelo conceptual de la base de conocimiento, así como las entidades que lo forman, podemos describir en detalle el modelo E/R que soporta la base de conocimiento de SciMAT. Particularmente, cada una de las entidades se representa mediante una tabla.

A continuación mostramos en detalle la estructura de la base de conocimiento, los atributos de cada tabla, así como las relaciones entre ellas. Además, en la Figura 4.2 se detalla el modelo E/R, mostrando cada una de las tablas, sus atributos y relaciones.

En la Figura 4.2, el símbolo de una llave representa a la clave principal de la tabla. El resto de atributos de la tabla vienen marcados por un rombo. Los rombos azules son atributos de la entidad, mientras que los rombos rosas representan a una clave foránea, es decir, la relación con otra entidad. Además, si un rombo se encuentra completamente relleno de color significa que los atributo no puede ser *null*. Por contra, un rombo exclusivamente con los bordes de color implica que el atributo puede ser *null*, es decir, es un atributo opcional. Por último, atendiendo a las relaciones, si la línea que une dos tablas es continua implica que la relación tiene que existir. Por el contrario, una línea discontinua implica que la relación puede ser opcional.

La tabla *Document*, quizás la más importante, representa a la entidad Documento. Como se mencionó en la Sección 4.2.1, es la entidad central de la base de conocimiento y la que articula la mayor parte de las relaciones con las entidades restantes. En este sentido, establece relaciones de muchos a muchos con las entidades Afiliación, Autor, Referencia y Término. Además establece una relación de uno a muchos con las entidades Revista (un documento se publica en una revista) y Fecha de publicación (un documento se publica en una fecha concreta). Los atributos que contiene son:

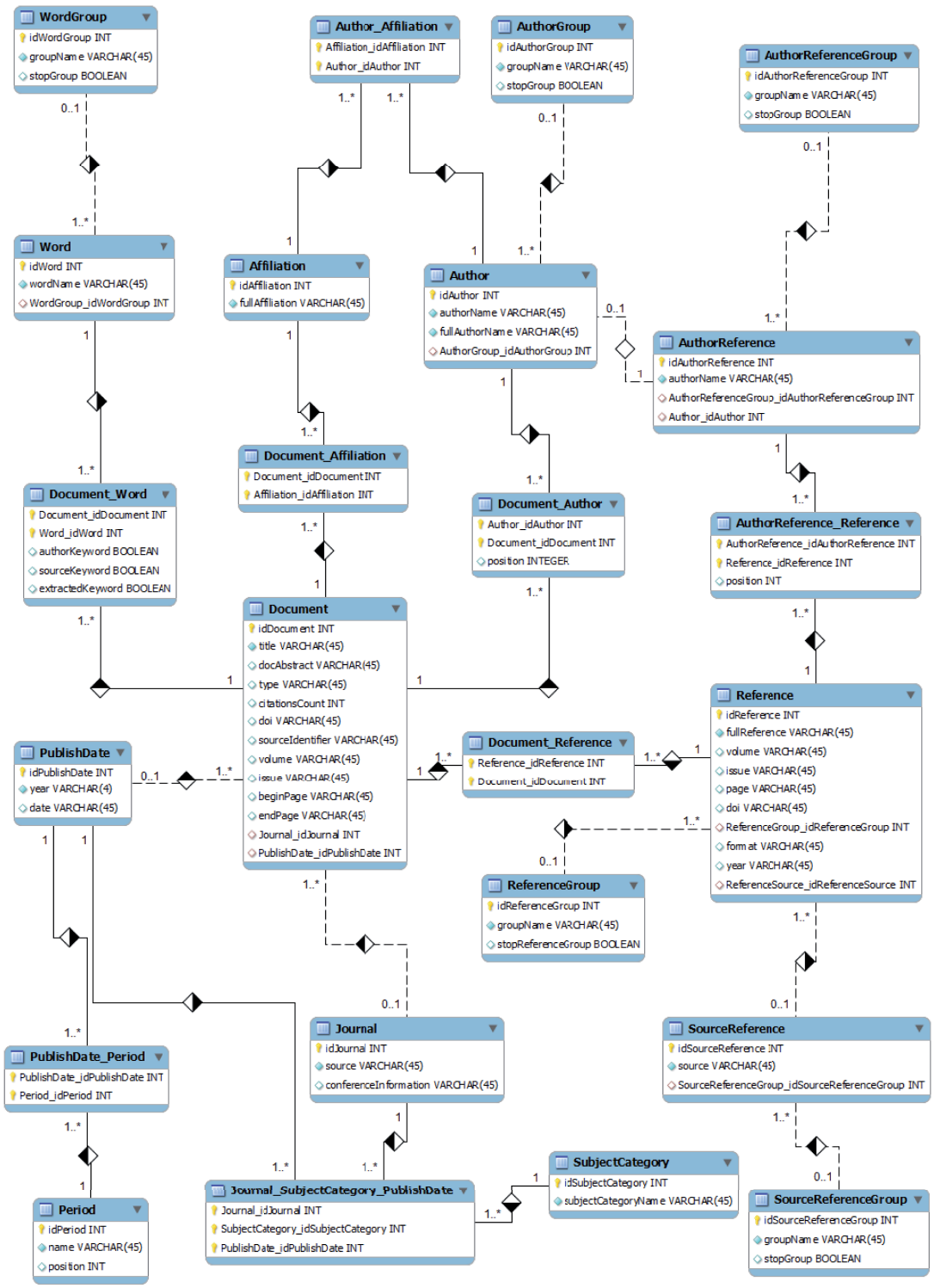


Figura 4.2: Modelo Entidad/Relación de la base de conocimiento.

- **idDocument**: clave principal del documento.
- **title**: título del documento.
- **docAbstract**: resumen del documento.
- **type**: tipo de documento, el cual no está restringido a un conjunto cerrado de tipos. De esta forma, el usuario puede establecer los tipos que más les convengan, como artículo, congreso, revisión, etc.
- **citationsCount**: número de citas recibidas por el documento, el cual por defecto será 0.
- **doi**: *Digital Object Identifier* del documento.
- **sourceIdentifier**: identificador interno del documento dentro de la base de datos bibliográfica de donde se descargó la información del documento.
- **volume**: volumen de la revista en donde se publicó el documento.
- **issue**: número de la revista en donde se publicó el documento.
- **beginPage**: número de la página de comienzo del documento.
- **endPage**: número de la página de fin del documento.
- **Journal_idJournal**: clave foránea que representa la revista relacionada con el documento (relación de la entidad Documento con la entidad Revista). Es decir, representa la revista en dónde el documento fue publicado.
- **PublishDate_idPublishDate**: clave foránea que representa la fecha cuando el documento fue publicado. Es decir, representa la relación entre la entidad Documento y Fecha de publicación.

La tabla *Affiliation* representa a la entidad Afiliación. Dado que los diversos documentos que un autor publica a lo largo de su carrera puede haberlos realizado en diversos centros de trabajo, una Afiliación mantiene una relación de muchos a muchos con la entidad Autor. Además, dado que la información descargada de las bases de datos bibliográficas, normalmente especifica las afiliaciones, pero no especifica a qué autores pertenecen, una Afiliación mantiene también una relación de muchos a muchos con la entidad Documento. Particularmente, los atributos que contiene son:

- **idAffiliation**: clave principal de la afiliación.
- **fullAffiliation**: contiene una afiliación completa, es decir, la dirección postal que el autor puso en el Documento. Este atributo no puede ser *null* y además tiene que ser único.

La tabla *Author* representa a la entidad Autor, la cual está relacionada, además de las relaciones vistas anteriormente, con la entidad Grupo de Autores (un autor puede tener asociado uno o ningún grupo de autores). Particularmente, los atributos que contiene son:

- **idAuthor**: clave principal del autor.
- **authorName**: nombre abreviado del autor (nombre de firma).
- **fullAuthorName**: nombre completo del autor. Además el conjunto formado por este atributo y por *authorName* tiene que ser único.
- **AuthorGroup_idAuthorGroup**: clave foránea que representa el Grupo de Autores relacionado con este Autor.

La tabla *Journal* representa a la entidad Revista. Además de la relación de muchos a uno con la entidad Documento vista anteriormente, establece una relación ternaria

de muchos a muchos con las entidades Categoría y Fecha de publicación, la cual se detallará posteriormente en esta sección. En particular, los atributos que contiene son:

- **idJournal**: clave principal de la revista.
- **source**: nombre de la revista, el cual tiene que ser único.
- **conferenceInformation**: campo que representa el nombre de la conferencia. Este campo es útil para los artículos de revista procedentes de congresos, como los números especiales.

La tabla *SubjectCategory* representa a la entidad Categoría. Los atributos que contiene son:

- **idSubjectCategory**: clave principal de la categoría.
- **subjectCategoryName**: nombre de la categoría, el cual tiene que ser único.

La tabla *PublishDate* representa a la entidad Fecha de publicación, la cual suele estar representada por un año concreto, y ocasionalmente, por una fecha completa (día, mes y año). Además de las relaciones vistas anteriormente con la entidad Documento y la ternaria con las entidades Revista y Categoría, una Fecha de publicación se relaciona de muchos a muchos con la entidad Periodo. En particular, los atributos que contiene son:

- **idPublishDate**: clave principal de la fecha de publicación.
- **year**: año de la fecha de publicación, el cual tiene que ser único.
- **date**: campo que representa a la fecha completa de publicación.

La tabla *Period* periodo representa a un conjunto de años (Fechas de publicación). Dado que los diversos periodos de la base de conocimiento pueden contener conjuntos de fechas de publicación solapados, como mencionamos anteriormente, la relación entre la entidad Periodo y Fecha de publicación es de muchos a muchos. Los atributos de esta tabla son:

- **idPeriod**: clave principal del periodo.
- **name**: nombre del periodo, el cual tiene que ser único.
- **position**: este atributo representa la posición que ocupará el periodo dentro del análisis longitudinal realizado en el análisis de mapas científicos. Por defecto contendrá el valor de 1.

La tabla *Word* representa a la entidad Término. Además de la relación de muchos a muchos con la entidad Documento (un término puede pertenecer a varios documentos y un documento contiene un conjunto de términos), un Término se encuentra relacionado con un Grupo de Términos. En particular, los atributos que contiene son:

- **idWord**: clave principal del término.
- **wordName**: nombre del término, es decir, la cadena de caracteres que representa al término. Este atributo tiene que ser único.
- **WordGroup_idWordGroup**: clave foránea que representa el Grupo de Términos relacionado con este Término.

La tabla *Reference* representa a la entidad Referencia. Además de la relación muchos a muchos con la entidad Documento, una Referencia tiene una relación muchos a muchos con la entidad Autor-Referencia, y una relación uno a muchos con las entidades Grupo de Referencias y Revista-Referencia. Los atributos de esta tabla son:

- **idReference**: clave principal de la referencia.
- **fullReference**: referencia completa, tal y como se encuentra en el registro descargado de la base de datos bibliográfica. Este atributo tiene que ser único.
- **volume**: si la referencia se puede dividir en diversos campos, este atributo representa el volumen de la revista donde se publicó el documento al que hace mención esta referencia.
- **issue**: número de la revista donde se publicó el documento al que hace mención esta referencia.
- **page**: páginas de la revista donde se publicó el documento al que hace mención esta referencia.
- **doi**: *Digital Object Identifier* del documento al que hace mención esta referencia.
- **format**: dado que dependiendo de la base de datos bibliográfica que se haya utilizado para la descarga de datos, el formato de la referencia puede ser distinto, este campo identifica el formato en el que se encuentra la referencia, para de este modo facilitar su preprocesamiento y así extraer las distintas piezas de información contenida en la referencia.
- **year**: fecha (año) cuando se publicó el documento al que hace mención la referencia.
- **ReferenceGroup_idReferenceGroup**: clave foránea que representa el Grupo de Referencias relacionado con esta Referencia. Este atributo puede ser *null*, por lo que una Referencia puede o no estar asociado con un Grupo de Referencias.

- **ReferenceSource_idReferenceSource**: clave foránea que representa la Revista-Referencia relacionada con esta Referencia. Es decir, la revista donde se publicó el documento al que hace mención esta referencia. Este atributo es opcional, lo que implica que una Referencia puede o no estar relacionada con una Revista-Referencia.

La tabla *AuthorReference* representa a la entidad Autor-Referencia. Además de la relación muchos a muchos con la entidad Referencia, un Autor-Referencia esta relacionado con un Grupo de Autores. Un caso particular es la relación de uno a uno que se mantiene con la entidad Autor, lo que nos permite identificar un Autor-Referencia con un Autor de la base de conocimiento. En concreto, los atributos que contiene son:

- **idAuthorReference**: clave principal del autor-referencia.
- **authorName**: nombre del Autor-Referencia, el cual tiene que ser único.
- **AuthorReferenceGroup_idAuthorReferenceGroup**: clave foránea que representa el Grupo de Autor-Referencia relacionado con este Autor-Referencia.
- **Author_idAuthor**: clave foránea que representa al Autor relacionado con este Autor-Referencia.

Tabla *SourceReference*

- **idSourceReference**: clave principal de la Revista-Referencia.
- **source**: nombre de la Revista-Referencia, el cual tiene que ser único.
- **ReferenceSourceGroup_idReferenceSourceGroup**: clave foránea que representa el Grupo de Revista-Referencia relacionado con esta Revista-Referencia.

Las tablas *AuthorGroup*, *ReferenceGroup*, *AuthorReferenceGroup*, *SourceReference-Group* y *WordGroup* representan a los grupos: Grupo de Autores, Grupo de Referencias, Grupo de Autor-Referencia, Grupo de Revisa-Referencia y Grupo de Términos, respectivamente. Dado que la estructura de estas cinco entidades es similar, la describiremos de forma conjunta. En particular, los atributos que contiene un Grupo son:

- **idAuthorGroup, idReferenceGroup, idSourceReferenceGroup, idWordGroup**: clave principal de los grupos.
- **groupName**: nombre del grupo, el cual tiene que ser único.
- **stopGroup**: indica si el grupo debe ser utilizado (*true*) o no (*false*) en el análisis de mapas científicos.

Por último, como hemos venido describiendo a lo largo de la sección, existen un gran número de relaciones de muchos a muchos entre varias de las entidades de la base de conocimiento. Estas relaciones se representan mediante tablas. A continuación se detallan cada una de estas relaciones.

La tabla *Document_Affiliation* representa la relación entre un Documento y una Afiliación.

- **Document_idDocument**: clave foránea que representa al Documento.
- **Affiliation_idAffiliation**: clave foránea que representa a la Afiliación.

La tabla *Document_Author* representa la relación entre un Documento y un Autor. Además, esta relación permite identificar la posición que ocupa el autor en el documento.

- **Author_idAuthor**: clave foránea que representa al Autor.
- **Document_idDocument**: clave foránea que representa al Documento.

- **position**: posición que ocupa el Autor en el Documento.

La tabla *Document_Reference* representa la relación entre un Documento y una Referencia.

- **Reference_idReference**: clave foránea que representa a la Referencia.
- **Document_idDocument**: clave foránea que representa al Documento.

La tabla *Document_Word* representa la relación entre un Documento y un Término. Además, esta relación permite identificar el rol que mantiene el Término en el Documento.

- **Document_idDocument**: clave foránea que representa al Documento.
- **Word_idWord**: clave foránea que representa al Término.
- **authorKeyword**: indica si el Termino mantiene el rol "palabras clave de autor", lo que se indicara con un valor *true*.
- **sourceKeyword**: indica si el Termino mantiene el rol "palabras clave de la fuente de información", lo que se indicara con un valor *true*.
- **extractedKeyword**: indica si el Termino mantiene el rol "términos extraídos", lo que se indicara con un valor *true*.

La tabla *Author_Affiliation* representa la relación entre un Documento y una Afiliación.

- **Affiliation_idAffiliation**: clave foránea que representa a la Afiliación.
- **Author_idAuthor**: clave foránea que representa al Autor.

La tabla *PublishDate_Period* representa la relación entre una Fecha de publicación y un Periodo.

- **PublishDate_idPublishDate**: clave foránea que representa a la Fecha de publicación.
- **Period_idPeriod**: clave foránea que representa al Periodo.

La tabla *Journal_SubjectCategory_PublishDate* representa la relación ternaria entre una Revista, Categoría y Fecha de publicación.

- **Journal_idJournal**: clave foránea que representa a la Revista.
- **SubjectCategory_idSubjectCategory**: clave foránea que representa a la Categoría.
- **PublishDate_idPublishDate**: clave foránea que representa a la Fecha de publicación.

Finalmente, la tabla *AuthorReference_Reference* representa la relación entre una Referencia y un Autor-Referencia. Además, al igual que la relación *Document_Author*, permite definir la posición que ocupa el Autor-Referencia dentro de la Referencia.

- **AuthorReference_idAuthorReference**: clave foránea que representa al Autor-Referencia.
- **Reference_idReference**: clave foránea que representa a la Referencia.
- **position**: posición que ocupa el Autor-Referencia en la Referencia.

4.3. Módulos, Funcionalidades y Algoritmos

En esta descripción analizamos los diversos métodos y algoritmos que incorpora SciMAT para realizar el análisis de mapas científicos.

En este sentido, primeramente estudiaremos los tres módulos en los que se divide la interfaz gráfica de SciMAT (Sección 4.3.1), para después describir la arquitectura de SciMAT (Sección 4.3.2).

4.3.1. Descripción de los Módulos

La interfaz gráfica de SciMAT se divide en tres módulos diferentes: i) un módulo dedicado a la gestión de la base de conocimiento y sus entidades, ii) un módulo encargado de realizar el análisis de mapas científicos, y iii) un módulo para visualizar los mapas y resultados generados. Estos módulos permiten al analista desarrollar los diferentes pasos del flujo de trabajo de un análisis de mapas científicos.

El *módulo para gestionar la base de conocimiento* es el responsable de construirla, importando la información a partir de datos recopilados de varias fuentes de información, así como limpiar dicha información y eliminar los posibles errores en las diversas entidades. Este módulo puede considerarse como una primera etapa en el preprocesamiento.

Este módulo incorpora diferentes métodos para importar información de diferentes fuentes de información bibliográfica. En particular, permite importar datos procedentes de ISIWoS y Scopus (en formato RIS). Además, el módulo de gestión permite añadir nuevos datos (importar información de fuentes bibliográficas) a una base de conocimiento existente.

Teniendo en cuenta las capacidades de edición del módulo de gestión, cada entidad puede editarse y sus relaciones con otras entidades también pueden ser modificadas.

Además, a través de los Grupos, se puede realizar el proceso de unificación de elementos similares, en donde el usuario podrá unir los elementos que él considere que representan a la misma entidad bajo un mismo grupo. Asimismo, este módulo incorpora métodos que ayudan al analista en el proceso de unificación, como búsqueda de elementos similares por sus formas en plural o singular, por la distancia de edición (Distancia de Levenshtein) entre dos elementos, o incluso importando los grupos y sus elementos asociados desde un fichero en formato XML (previamente generado). Finalmente, comentar que la división temporal (preprocesamiento) se realiza a través de la entidad Periodo.

El *módulo para realizar el análisis de mapas científicos* se ha implementado en forma de asistente, en donde el usuario puede seleccionar los métodos y algoritmos que se usarán para realizar cada paso. Finalmente, el análisis de mapas científicos se realizará con lo configuración dada por el usuario. Aunque el asistente ha sido implementado en concordancia con el flujo de trabajo descrito en la Sección 1.2, algunos de los pasos se realizarán en distinto orden. Por ejemplo, el proceso de unificación y división temporal debe de realizarse antes, utilizando el módulo de gestión de la base de conocimiento.

De este modo, la secuencia de pasos a seguir con el asistente, puede dividirse en cuatro etapas principales: i) construcción del conjunto de datos a analizar (dataset), ii) creación y normalización de la red bibliométrica, iii) aplicación de un algoritmo de clustering para realizar el mapa, y iv) realizar un conjunto de análisis sobre el mapa para extraer conocimiento de él. A continuación, se describe en detalle cada una de estas etapas:

1. Construcción del conjunto de datos: en esta etapa, el usuario puede configurar los periodos a usar en el análisis, los aspectos (sociales, intelectuales o conceptuales) que quiere estudiar, así como la porción de los datos que desea utilizar.

a) Seleccionar los periodos: se debe de seleccionar un conjunto de periodos de los existentes en la base de conocimiento, los cuales han de haber sido previamente definidos. Estos periodos se usarán en el análisis longitudinal, para analizar la evolución del aspecto seleccionado.

b) Seleccionar la unidad de análisis: como unidad de análisis el usuario puede seleccionar cualquiera de los cinco grupos existentes en la base de conocimiento: Grupo de Autores, Grupo de Autor-Referencia, Grupo de Revista-Referencia, Grupo de Referencias, o Grupo de Términos. Únicamente se puede seleccionar uno de los grupos para realizar el análisis. En el caso de que el Grupo de Términos haya sido seleccionado, el usuario tendrá que seleccionar el rol del término que quiere analizar. En este caso, el usuario puede seleccionar las palabras clave de autor, las palabras clave procedentes la fuente de información, los términos extraídos, o cualquier combinación de estos.

c) Reducción de datos: SciMAT permite filtrar los datos a usar mediante un umbral de frecuencia mínima. Para cada periodo seleccionado en los pasos previos, se debe seleccionar un umbral, por lo que en cada periodo, sólo se tendrán en cuenta los elementos que aparezcan en al menos n documentos.

2. Creación y normalización de la red bibliométrica: en esta etapa se construye la red utilizando datos de co-ocurrencia o de emparejamiento, incluso emparejamiento agregado. Tras la creación de la red, ésta es filtrada para dejar las relaciones más importantes, es decir, los enlaces entre elementos más importantes. Finalmente, se realiza un proceso de normalización utilizando una medida de similitud.

a) Seleccionar el modo en que se construirá la red bibliométrica: la red bibliométrica puede construirse basándose en co-ocurrencias o en empareja-

miento. Mediante el uso de co-ocurrencias, se pueden construir las siguientes redes: co-autor, co-palabras, co-citación (usando las referencias), co-citación de autores (usando los grupos de la entidad autores-referencia) y co-citación de revista (usando los grupos de la entidad revista-referencia). Por otro lado, el emparejamiento puede usarse de forma básica o agregada. En la primera, la red bibliométrica se construye emparejando documentos mediante la unidad de análisis seleccionada. Es decir, si por ejemplo se seleccionó las Referencias como unidad de análisis, se construirá una red de emparejamiento bibliográfico. En la segunda forma, la red de emparejamiento se construye agregando una red de emparejamiento básico utilizando los autores o revistas, en donde los elementos a comparar en el emparejamiento serán la unidad de análisis seleccionada. En este caso, si elegimos las referencias como unidad a analizar y los autores como elementos de agregación construiremos una red de emparejamiento bibliográfico de autores.

b) Reducción de la red: cada enlace en la red tiene un valor, es decir, tiene un peso asociado, el cual puede ser el valor de co-ocurrencia o de emparejamiento de los elementos enlazados. En este sentido, SciMAT permite filtrar la red utilizando un umbral de peso mínimo del enlace. Al igual que en la reducción de datos, para cada periodo se tiene que seleccionar un valor para el umbral, por lo que en cada periodo sólo aparecerán los enlaces con un peso mayor o igual al umbral elegido.

c) Seleccionar una medida de similitud para normalizar la red: SciMAT permite al usuario elegir entre las medidas de similitud comúnmente utilizadas en la literatura para normalizar una red bibliométrica. En concreto, las medidas disponibles son: fuerza de asociación, índice de equivalencia, índice de

inclusión, índice de Jaccard, y por último el coseno de Salton.

3. Aplicación de un algoritmo de clustering para crear el mapa científico y sus grupos o subredes asociadas: en esta etapa, se utiliza un algoritmo de clustering o agrupamiento para construir el mapa científico, pudiendo elegir entre diferentes algoritmos: Centros Simples [29], o Enlace simple, [100] y sus variantes, como Enlace completo, Enlace medio, o Suma de enlaces.

4. Aplicar un conjunto de análisis: la etapa final del asistente consite en seleccionar los análisis que se tienen que realizar sobre los mapas generados.

   a) Análisis de red: por defecto, SciMAT añade la densidad y centralidad de Callon como medida de red a cada uno de los subgrupos encontrados en cada periodo. Ambas medias serán útiles para categorizar los subgrupos detectados en cada periodo en un diagrama estratégico (ver Sección 3.1.2).

   b) Análisis del rendimiento: SciMAT es capaz de medir y evaluar las salidas mediante diversas medidas de rendimiento y calidad basadas en indicadores bibliométricos. Para realizar este análisis, SciMAT añade a cada subgrupo detectado un conjunto de documentos, para posteriormente calcular su rendimiento, y calidad basándose en medias cuantitativas y cualitativas (usando las citas recibidas, contabilizando el número de documentos de cada subgrupo, etc.).

   SciMAT incorpora métodos diferentes para asignar documentos a los subgrupos detectados dependiendo de si son redes basadas en co-ocurrencia, o en emparejamiento.

   Para las redes basadas en co-ocurrencia, SciMAT incorpora cinco métodos diferentes: i) documentos principales (ver Sección 3.1.2), ii) intersección, el

cual añade los documentos que contienen todos los elementos del subgrupo, iii) k-documentos importantes, el cual añade a cada grupo los documentos que tienen al menos k elementos en común con él, iv) documentos secundarios (ver Sección 3.1.2), y v) unión, el cual añade a cada subgrupo los documentos que tienen al menos un elemento en común con él (es similar a la unión de los documentos secundarios y principales).

Por otro lado, para las redes basadas en emparejamiento, SciMAT dispone dos tipos de métodos diferentes para asignar documentos, dependiendo del tipo de emparejamiento utilizado. En este sentido, si se realizó un emparejamiento básico (cada elemento del subgrupo será un documento), exclusivamente podremos utilizar un método de asignación de documentos básico, el cual añade como documentos, sus mismos elementos. Finalmente, si se realizó un emparejamiento agregado sólo se puede elegir el método agregado de asignación de documentos, el cual añade al subgrupo los documentos asociados con cada uno de sus elementos.

Además, debemos resaltar que cada elemento de los subgrupos (nodos), tiene también asociado un conjunto de documentos, los cuales se corresponden con el conjunto de documentos asociados a cada elemento del conjunto de datos extraído en la primera etapa.

Una vez que los documentos han sido asignados a cada uno de los subgrupos detectados en los diferentes periodos de tiempo seleccionados, se pueden calcular un conjunto de medidas de rendimiento y calidad a cada subgrupo. En este sentido, SciMAT permite utilizar las citas recibidas por cada uno de los documentos asignados a los subgrupos para calcular medidas básicas, como la suma, media, máximo y mínimo de citas, e indicadores bibliométricos

complejos, como el índice h [1, 46], índice g [34], índice hg [2] o índice q^2 [17]. Además, SciMAT añade por defecto, como medida de rendimiento, el número de documentos de cada subgrupo.

c) Análisis temporal o longitudinal: permite al usuario descubrir la evolución social, intelectual o conceptual del campo científico analizado. SciMAT construye un mapa de evolución para detectar las áreas de evolución tal y como fue descrito en la Sección 3.1.3. Además, permite analizar el solapamiento entre los elementos de periodos de tiempo consecutivos a través de un mapa de solapamiento (descrito en la Sección 3.1.3). Tanto para el cálculo de los pesos de los nexos entre temas de diferentes periodos en el mapa de evolución, como para calcular el índice de estabilidad entre periodos en el mapa de solapamiento, SciMAT permite elegir entre las siguientes medidas: fuerza de asociación [29, 105], índice de equivalencia [20], índice de inclusión [84], índice de Jaccard [75], y por último el coseno de Salton [86].

Finalmente, el *módulo de visualización* muestra los resultados obtenidos por el análisis, ayudando al usuario a entenderlos, interpretarlos y por lo tanto, obtener conocimiento de ellos. El módulo de visualización permite al usuario navegar e interactuar con los resultados, pudiendo así, centrarse en las zonas deseadas y analizarlas en profundidad. Dicho módulo incorpora las técnicas de visualización descritas en la metodología presentada en el Capítulo 3: diagramas estratégicos, redes de grupos, mapas de evolución y mapas de solapamiento.

El módulo de visualización es capaz de exportar los resultados a un informe en formato LaTeX o HTML. Las diversas figuras (diagramas estratégicos, mapas de evolución, etc.) se exportan también en formato PNG y SVG, para que de este modo, el usuario puede editarlas utilizando una herramienta de diseño gráfico. Además, las redes

de subgrupos y los mapas de evolución se exportan en formato Pajek.

Para resumir, en la Figura 4.3 se muestra el flujo de trabajo que implementa SciMAT para realizar un análisis de mapas científicos.

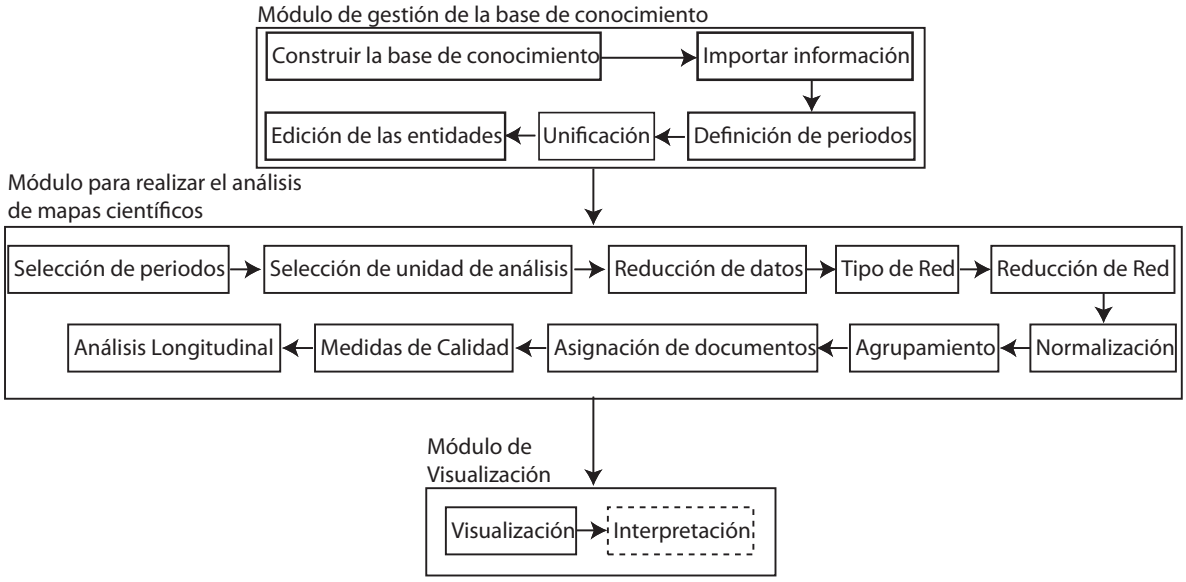


Figura 4.3: Flujo de trabajo de SciMAT.

4.3.2. Arquitectura de SciMAT

Una vez descritos los módulos de los que se compone SciMAT, podemos mostrar su arquitectura, para de este modo detallar el funcionamiento interno de SciMAT.

Internamente, SciMAT se compone de diversos bloques independientes que interactúan entre sí para desarrollar los diversos pasos del flujo de trabajo mostrado en la Figura 4.3. Una parte de dichos bloques se ocupan de la gestión de la interfaz gráfica así como de la interacción con el usuario. Por otro lado, ciertos módulos realizan tareas opacas para el usuario, siendo éstos el núcleo central del funcionamiento de SciMAT. En particular, en la Figura 4.4 se puede apreciar en detalle la arquitectura de SciMAT. En ella, la zona sombreada representa los bloques de gestión de la interfaz.

El núcleo principal de SciMAT lo forman:

- El bloque para la comunicación y gestión de la base de datos (llamado "*modelo*" en este contexto).
- Y aquel que contiene los métodos y algoritmos necesarios para realizar el análisis de mapas científicos completo, desde la carga de datos hasta su visualización (*SciMAT API* -SciMAT Application Programming Interface-).

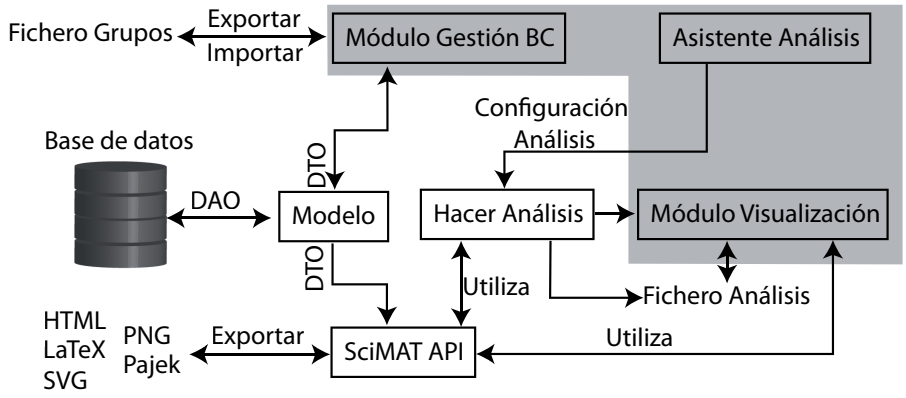


Figura 4.4: Arquitectura de SciMAT.

El *modelo* se encarga de hacer de nexo de unión entre la interfaz gráfica y el fichero físico donde se aloja la base de conocimiento. Para ello, implementa clases que desarrollan los patrones de diseño software *Data Access Object* (DAO) y *Data Transfer Object* (DTO). Los *DAO* se encargan de la comunicación con la base de datos, realizando todas las operaciones de inserción, modificación y borrado de las diversas entidades. Por otro lado, los *DTO* representan a las diversas entidades de la base de conocimiento, y sirven para transportar la información desde la base de datos, hasta los diferentes métodos que hagan uso de ella.

Por otro lado, el API (Application Programming Interface) de SciMAT contiene todos los métodos necesarios para realizar un análisis de mapas científicos. En particular,

dispone de métodos para importar datos desde las bases de datos bibliográficas ISIWoS y Scopus a la base de conocimiento, métodos para exportar e importar los grupos de entidades (en formato XML), diversos métodos de filtro, una gran variedad de técnicas de construcción de la red bibliométrca, las medidas más comunes para normalizar, diferentes algoritmos de clustering para crear el mapa, varias técnicas de análisis, etc. Es decir, el API de SciMAT contiene todos los métodos, técnicas, medidas y algoritmos necesarios para realizar las diversas configuraciones que se pueden realizar a través del asistente para realizar el análisis de mapas científicos (Sección 4.3.1). Gracias a las técnicas de programación orientada a objetos en las que se basa SciMAT, se puede extender el API de un modo sencillo para incorporar nuevos métodos, algoritmos y medidas.

El API de SciMAT implementa las técnicas de visualización necesarias para visualizar los resultados de acuerdo a la metodología presentada en el Capítulo 3. Además, incorpora diversos métodos para exportar los resultados a formatos, como SVG, PNG o Pajek. Además, dispone de métodos para realizar informes en HTML y LaTeX con todos los resultados del análisis de mapas científico realizado.

Además, el usuario avanzado podrá utilizar dicha API para desarrollar herramientas ad-hoc, que le permitan configurar y desarrollar un análisis de mapas científicos a su medida. Un ejemplo de este uso avanzado puede verse en el Anexo A, en donde se muestra el código usando el API de SciMAT que permite desarrollar el análisis de mapas científicos realizado en las Secciones 3.2 y 4.4.

Por otro lado, el bloque de gestión de la base de conocimiento es el encargado de desarrollar todas las tareas del módulo que tienen que ver con éste (Sección 4.3.1). Como hemos comentado, se comunica con la base de datos (base de conocimiento) a través del *modelo*, el cual responde a las peticiones sobre las entidades devolviendo información

encapsulada en objetos *DTO*. Además, este bloque se encarga de exportar los grupos de entidades a un fichero XML, así como su importación.

Uno de los bloques más importantes de la interfaz gráfica de SciMAT, es aquel encargado de configurar el análisis de mapas científicos, es decir, el asistente de configuración. Este bloque se encarga de generar un configuración determinada para la realización del análisis de mapas científicos. Una vez que el usuario a especificado la configuración deseada, ésta es enviada al bloque para realizar el análisis, el cual, utilizando el API, se encargará de realizar el análisis de mapas científicos. Al finalizar los resultados se almacenarán en un fichero y a su vez, se enviarán al bloque de visualización.

El bloque de visualización es el encargado de desarrollar todas las tareas del módulo de visualización descritos en la Sección 4.3.1. Para ello, emplea el API de SciMAT.

4.3.3. Tecnologías

SciMAT ha sido desarrollado utilizando el lenguaje de programación Java, lo que posibilita que la herramienta pueda ser ejecutada en cualquier plataforma o sistema operativo, como Microsoft Windows, Linux, MacOS, etc.

Atendiendo a las técnicas de programación empleadas, SciMAT se ha desarrollado bajo la metodología de programación orientada a objetos. Además, se han empleado diversos patrones de diseño software, como *Observable*, *Observer*, *Command*, *Edit*, *Singleton*, *Factory*, *Data Access Object*, *Data Transfer Object* etc. Gracias a los patrones empleados, y a las técnicas abstractas desarrolladas, SciMAT puede extenderse con gran facilidad, para de este modo incorporar nuevos métodos y algoritmos.

Sobre el desarrollo de la base de datos, comentar que ésta sigue un modelo relacional, tal y como se mostró en la Sección 4.2.2 y en la Figura 4.2. Ademas, se ha utilizado *SQLite* [53] como motor de base de datos para almacenar la base de conocimiento. *SQ-*

Lite es un paquete de dominio público que nos provee de un sistema de administración de bases de datos relacionales. Elegimos este motor de base de datos debido a sus características: no hace falta configurar un servidor, es multi-plataforma, auto-contenido, necesita muy pocos recursos de la máquina donde se ejecuta, y alta disponibilidad [53]. Gracias a que la base de datos se encuentra en formato *SQL*, cualquier usuario puede abrir la base de conocimiento utilizando cualquier lector y modificarla[3] .

Por último, se han utilizado tecnologías como SVG, HTML y XML, para exportar diversos resultados generados por SciMAT.

4.4. Usando SciMAT: Un Análisis Guiado

En las secciones previas hemos descrito SciMAT, mostrando sus principales características, así como sus módulos, métodos y algoritmos. En esta sección, ilustramos el modo de utilizar SciMAT, así como su interfaz y sus módulos a través de un ejemplo práctico (un completo manual de usuario de SciMAT puede encontrarse en el Anexo B). Esto nos permitirá mostrar el flujo particular de trabajo de SciMAT.

Como ejemplo, mostramos como realizar con SciMAT el análisis de mapas científicos realizado en la Sección 3.2. En dicha sección, el análisis se realizó utilizando diversas herramientas: i) CoPalRed para el proceso de unificación y construcción del mapa, y ii) herramientas desarroladas ad-hoc para realizar los análisis de rendimiento, temporal, así como para la visualización. Como veremos más adelante, SciMAT es capaz de realizar todo el proceso sin utilizar herramientas externas, y lógicamente, generando los mismos resultados.

El principal objetivo del análisis realizado en la Sección 3.2 era detectar los temas más importantes, productivos e impactantes tratados por las dos revistas más impor-

[3] Como por ejemplo el complemento *SQLite Manager* del navegador *Firefox*.

tantes del área TCD a lo largo de cinco periodos de tiempo consecutivos. Para realizar esto se utilizaron medidas de rendimiento, como el número de documentos (documentos principales + documentos secundarios), e indicadores bibliométricos basados en citas, como el número total de citas recibidas y el índice h.

A continuación, describimos como realizar dicho análisis utilizando SciMAT, desde la carga de los datos en bruto obtenidos de la fuente bibliográfica, hasta la visualización de los resultados. Para ello, mostraremos los diferentes pasos envueltos en el proceso.

Una vez que los datos bibliográficos han sido descargados de la fuente de información (en este caso se utilizaron ISIWoS, Scopus y Science Direct[4]), el *primer paso* en SciMAT es construir una base de conocimiento e importar los datos recopilados utilizando el módulo de gestión de la base de conocimiento.

El *segundo paso* es la edición de la base de conocimiento, para de ste modo, eliminar los posibles errores (en títulos, autores, referencias, etc.) y mejorar así la calidad de los datos. Para realizar esto, SciMAT incorpora un submódulo de gestión para cada entidad, por lo que el usuario puede editar fácilmente la información asociada con cada entidad y sus relaciones con otras entidades.

Debemos señalar que todos los módulos de gestión de entidades tienen la misma estructura: en la parte izquierda se encuentra la lista de entidades, y en la parte derecha se muestran los campos asociados con la entidad seleccionada.

En la Figura 4.5, se muestra el módulo para gestionar los documentos. En la lista de documentos (parte izquierda), se encuentra seleccionado uno de los artículos más citados de la base de conocimiento. En la parte izquierda (la zona de detalle), se muestra la información relativa a este documento.

Dado que necesitamos estudiar la evolución conceptual, se seleccionaron como unidad de análisis las palabras clave. Por esta razón, necesitamos realizar un proceso de

[4] Science Direct utiliza el formato RIS de Scopus para exportar su información.

unificación sobre la entidad Término, mediante la utilización de los Grupos de Términos. Para ello uniremos aquellas palabras que representen el mismo concepto bajo un mismo grupo. Esta tarea puede realizarse utilizando el módulo para crear manualmente grupos de términos.

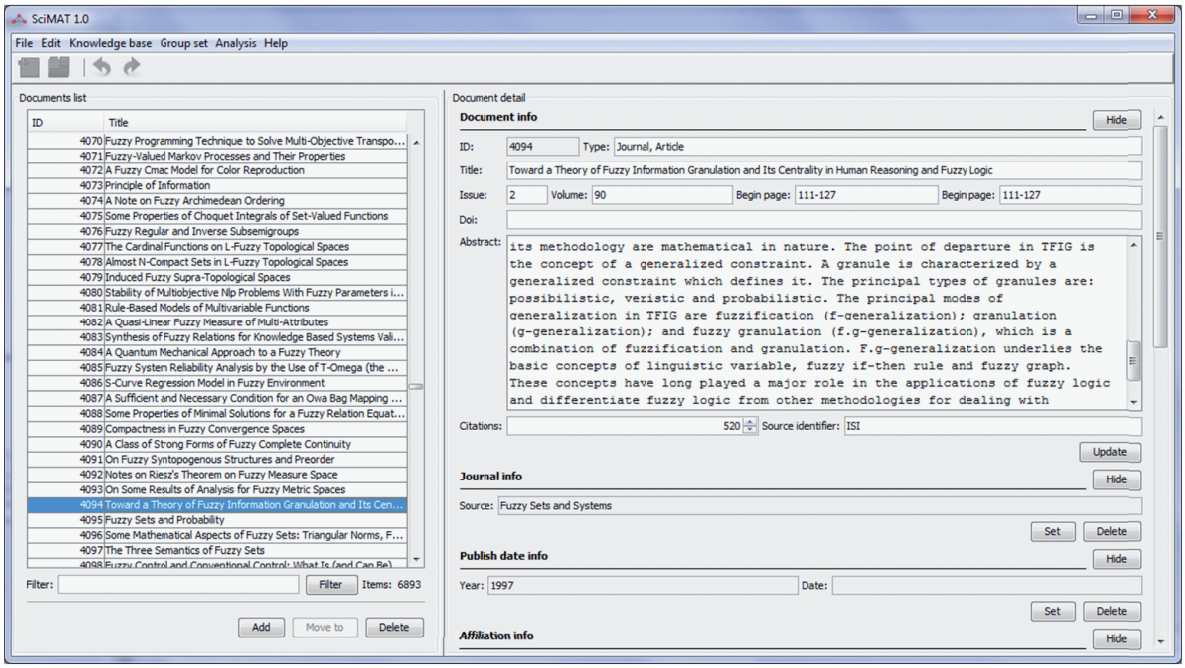


Figura 4.5: Módulo de gestión de documentos.

Debida a la naturaleza especial de los grupos (se emplean como unidad de análisis en la creación del mapa científico), estos tienen dos módulos de gestión: uno similar al del resto de entidades, y otro específico para realizar el proceso de unificación (manualmente). Al igual que los módulos de gestión anteriores, los módulos para crear grupos de elementos tienen una estructura similar: en la parte izquierda se encuentra la lista de grupos definidos, y en la parte derecha se muestran las entidades asociadas con el grupo seleccionado (tabla superior) y las entidades que todavía no han sido asignadas a ningún grupo (tabla inferior).

La Figura 4.6 muestra el módulo para crear grupos de términos. Por ejemplo, se

puede observar que se ha definido un grupo específico con nombre *GROUP-DECISION-MAKING*. Además, podemos ver cómo este grupo aglutina diferentes variantes del concepto (tabla superior). La tabla inferior permite al usuario añadir más variantes al concepto. Además, los grupos se pueden unir fusionando sus elementos.

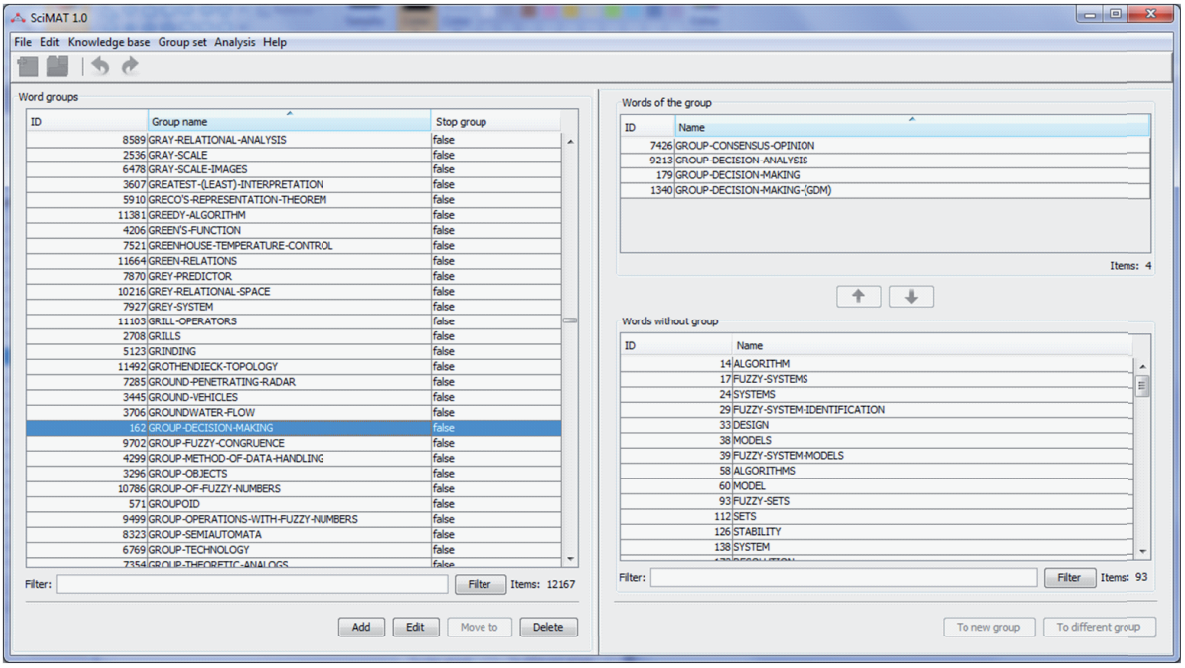


Figura 4.6: Módulo de creación de Grupo de Términos.

El *tercer paso* es definir los periodos, es decir, establecer los grupos de años que se usaran en el análisis longitudinal. Los periodos se tienen que definir mediante el módulo para gestionar periodos. En particular, en este ejemplo se definieron cinco periodos de tiempo: 1978-1989, 1990-1994, 1995-1999, 2000-2004 y 2005-2009.

Una vez que la base de conocimiento ha sido limpiada, y hemos creado los grupos y periodos, el *cuarto paso* es configurar el análisis de mapas científicos utilizando el asistente para tal efecto. Como se mostró en la Sección 4.3, este módulo nos permite seleccionar los periodos a utilizar (Figura 4.7), la unidad de análisis a utilizar[5] (Figura

[5] Exclusivamente se pueden seleccionar las unidades de análisis que contengan algún grupo definido.

4.8), los métodos de reducción de datos a emplear (Figura 4.9), la forma en la que crearemos la red bibliométrica (Figura 4.10), los parámetros de la reducción de redes (Figura 4.11), la medida de similitud empleada para normalizar la red (Figura 4.12), el algoritmo de clustering empleado para detectar los subgrupos (Figura 4.13), la forma de asignar los documentos a los subgrupos (Figura 4.14), los indicadores bibliométricos a calcular (Figura 4.15), así como medidas a emplear en el análisis longitudinal (Figura 4.16).

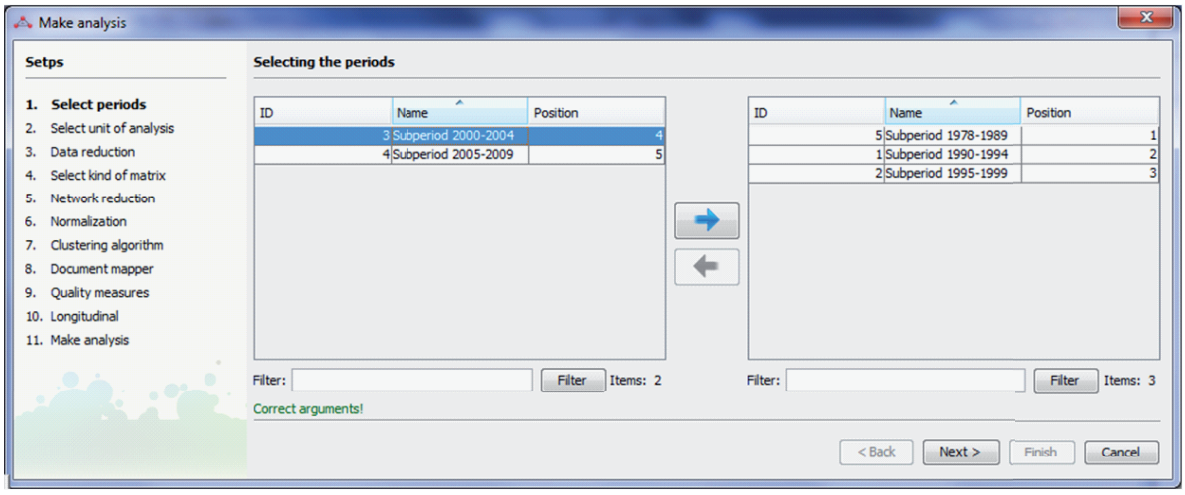


Figura 4.7: Seleccionando los periodos (Paso 1).

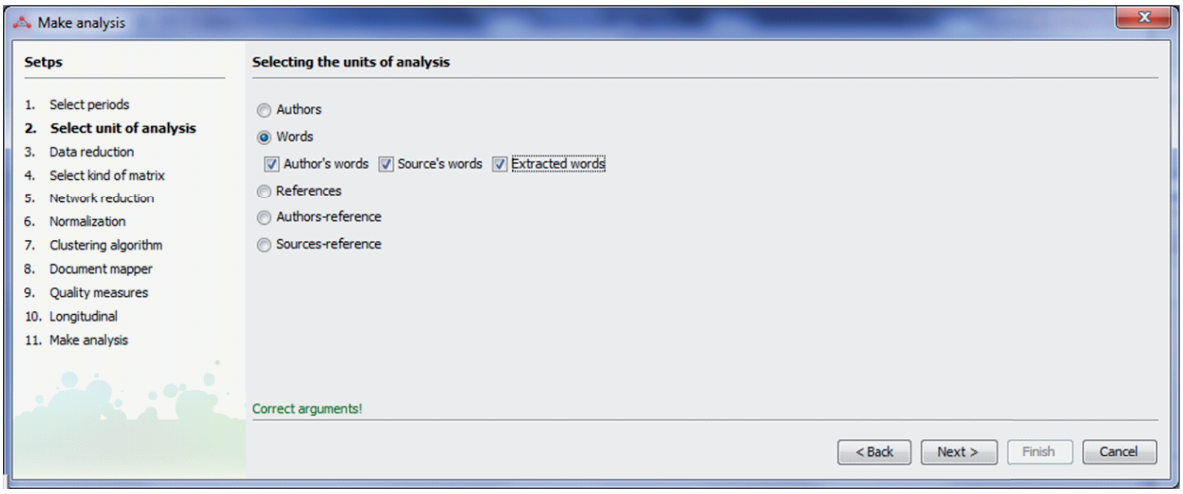


Figura 4.8: Seleccionando la unidad de análisis (Paso 2).

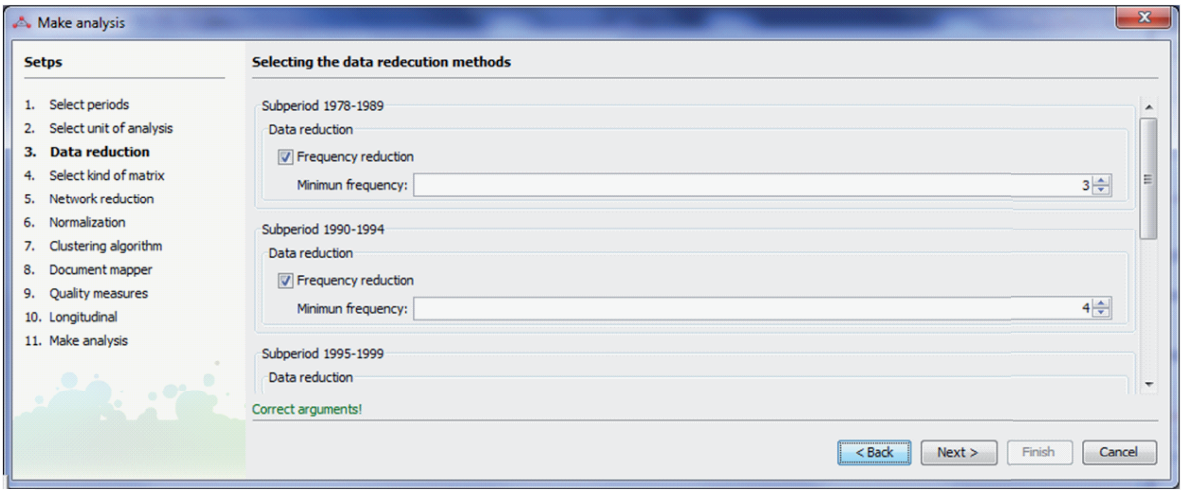


Figura 4.9: Seleccionando los umbrales para la reducción de datos (Paso 3).

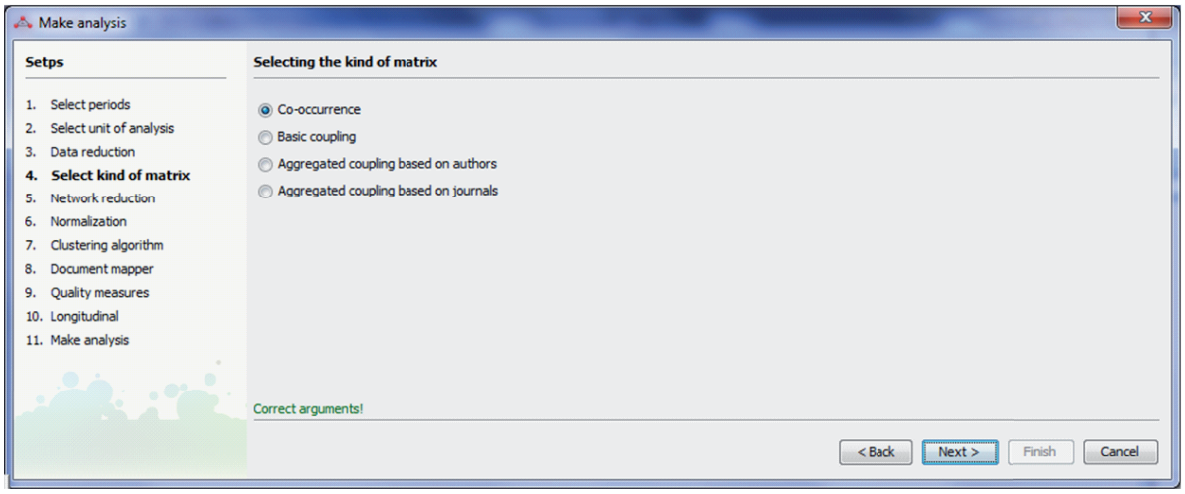


Figura 4.10: Seleccionando el tipo de red bibliométrica (Paso 4).

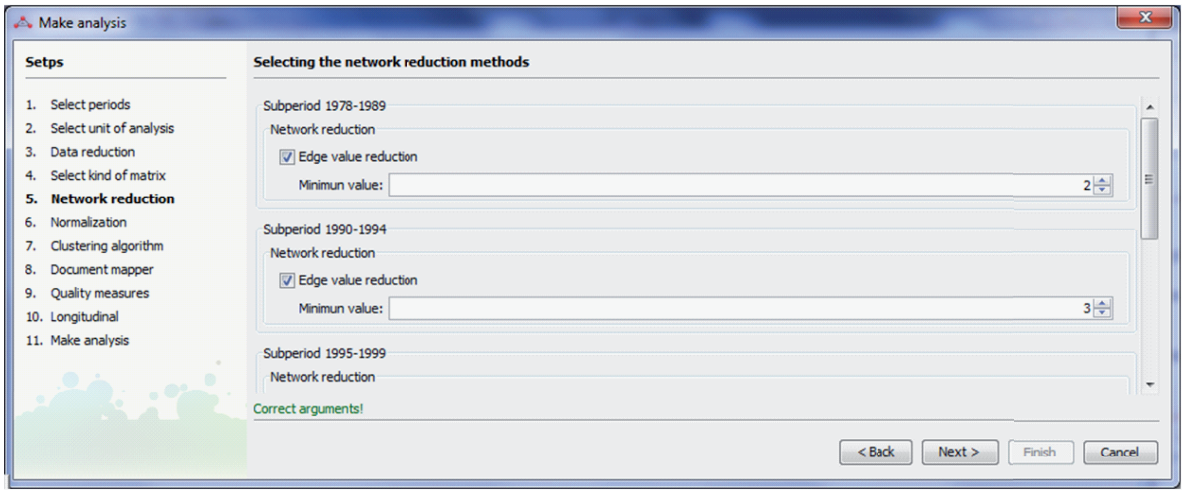


Figura 4.11: Seleccionando los umbrales para la reducción de red (Paso 5).

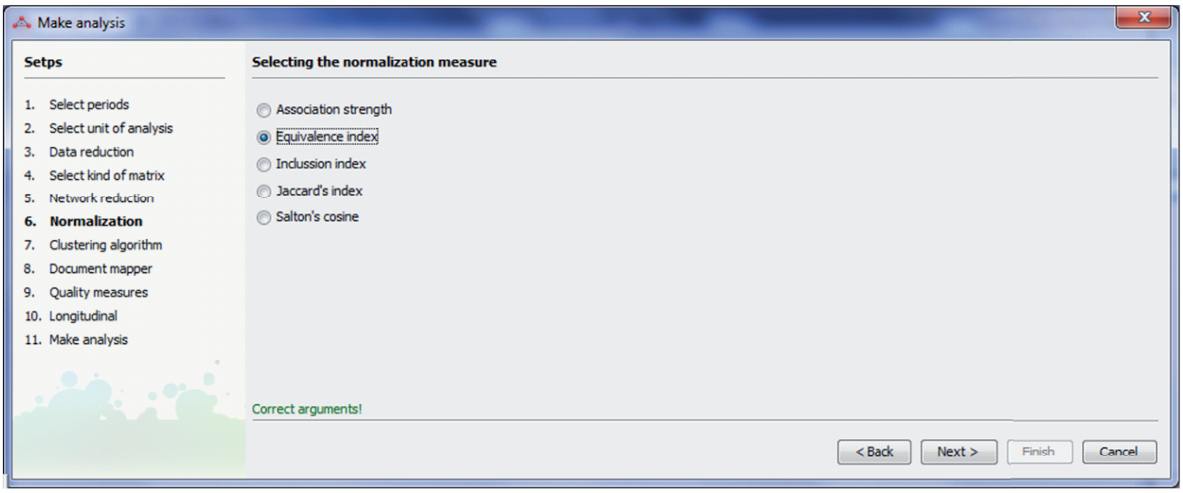


Figura 4.12: Seleccionando la medida de similitud para normalizar la red (Paso 6).

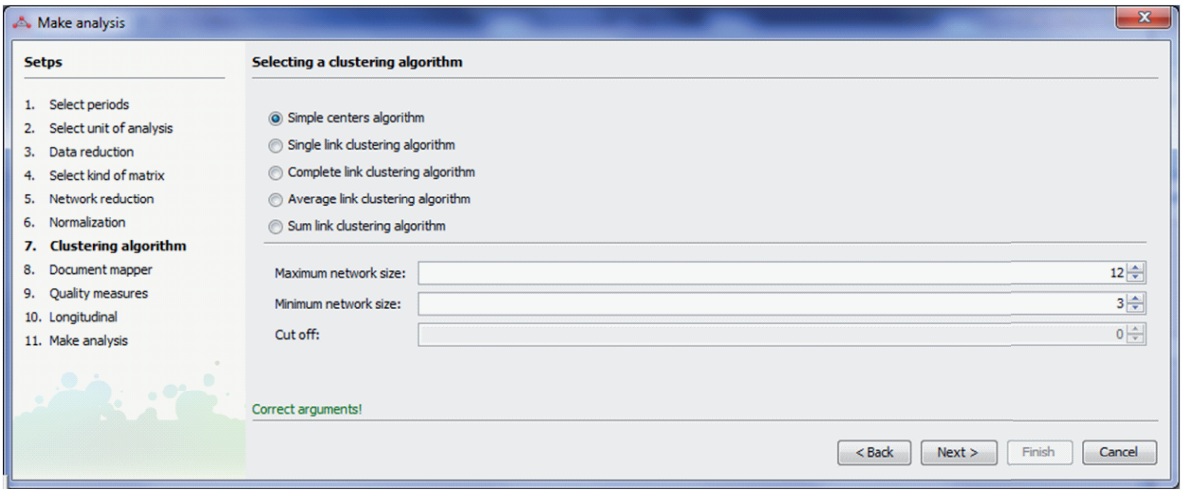


Figura 4.13: Seleccionando el algoritmo de clustering (Paso 7).

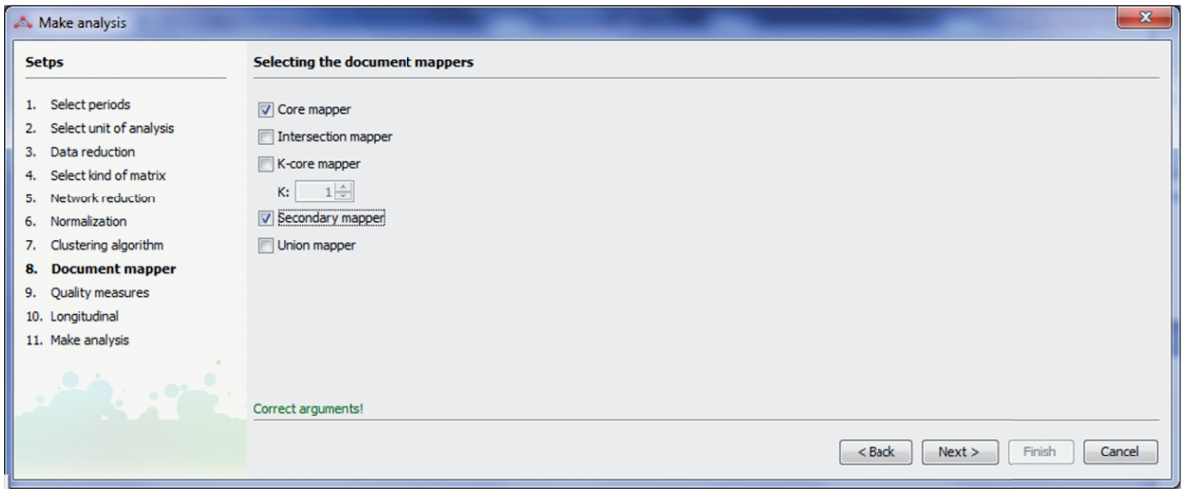


Figura 4.14: Seleccionar el modo en el que los documentos se asignarán a los grupos detectados (Paso 8).

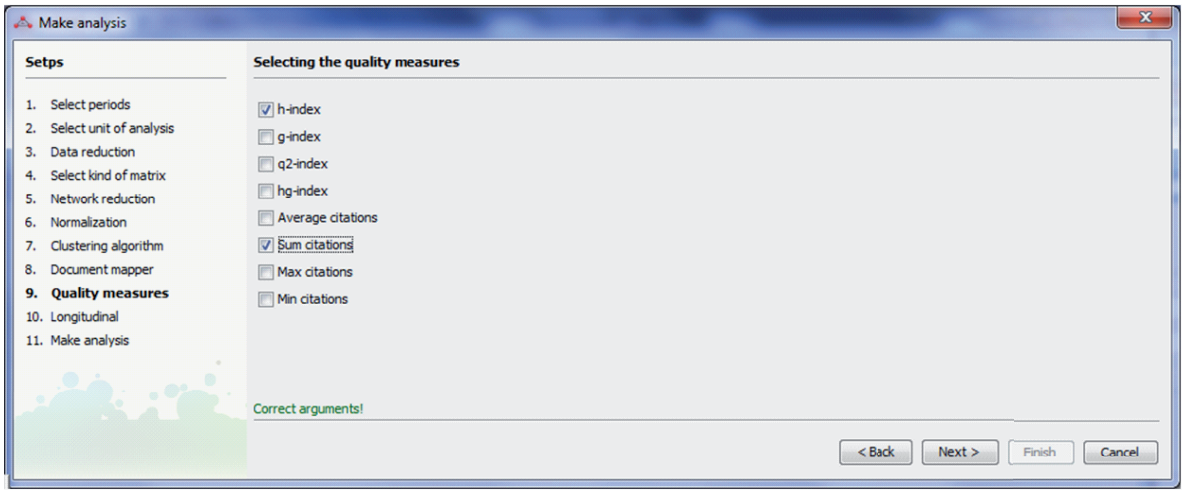


Figura 4.15: Seleccionando los indicadores bibliométricos (Paso 9).

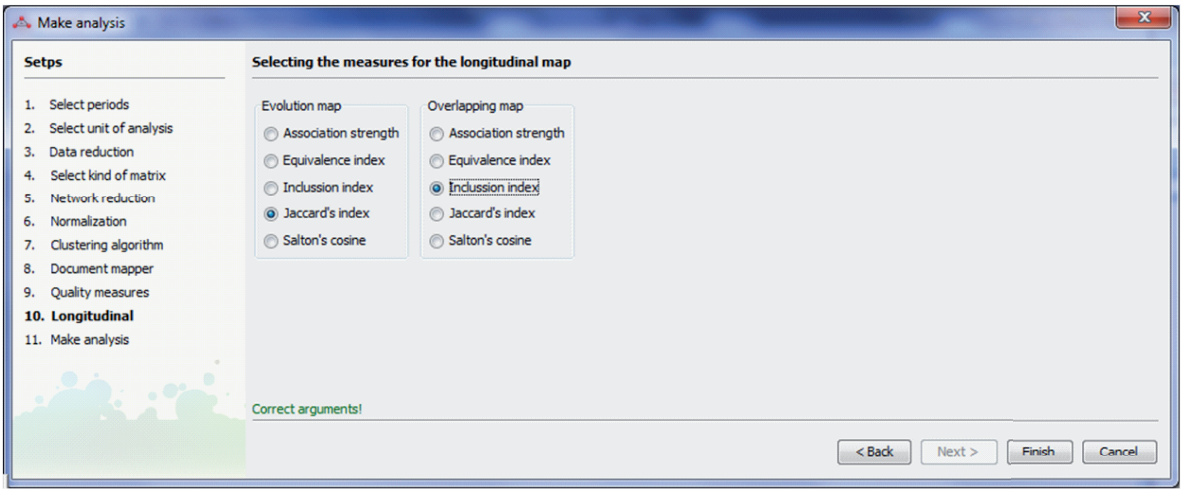


Figura 4.16: Seleccionar la medidas de similitud para los mapas longitudinales (Paso 10).

Particularmente, para realizar el análisis desarrollado en la Sección 3.2, la configuración utilizada para construir los mapas científicos fue: todos los periodos, los términos como unidad de análisis con sus tres roles posibles, co-ocurrencia como método para extraer la red bibliométrica, el índice de equivalencia para normalizar la red, y el algoritmo de los centros simples para detectar las subredes. Como indicadores bibliométricos se seleccionaron la suma de las citas recibidas por los documentos y el índice h. Estos indicadores se calcularon para los documentos secundarios y principales. Esta configuración se puede observar en las Figuras 4.7-4.16.

Al final de todos los pasos, el asistente realiza el análisis de mapas científicos utilizando la configuración seleccionada. Es entonces cuando los resultados son guardados, y el módulo de visualización es cargado. Dicho módulo tiene dos vistas: vista longitudinal y vista de periodos. En la *vista longitudinal*, se muestran los mapas de evolución y de solapamiento, ayudándonos a detectar la evolución de los subgrupos detectados a lo largo de los periodos de tiempo seleccionados, así como la continuidad y fugacidad de

los elementos en periodos consecutivos. En la Figura 4.17 se muestra la vista longitudinal, pero por claridad sólo se muestra el mapa de evolución. Finalmente, la *vista de periodos* (ver Figura 4.18) muestra información detallada de los resultados obtenidos para cada periodo, como su diagrama estratégico, subgrupos asociados, etc. Además, para cada subred detectada se muestra sus medidas de calidad e impacto.

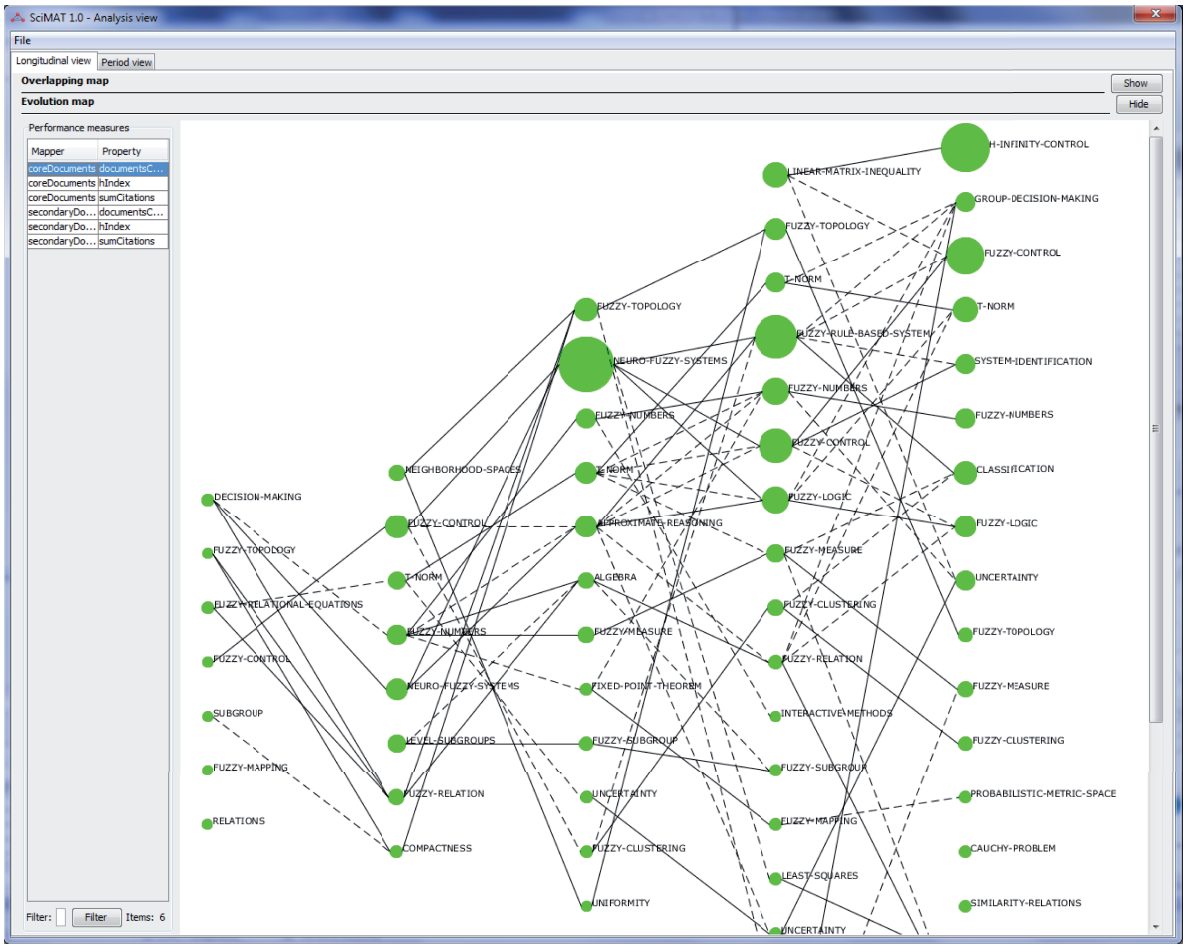


Figura 4.17: Vista longitudinal.

Finalmente, comentar que aunque en este estudio nos hemos centrado en los aspectos conceptuales de un área científica, utilizando para ello una red bibliométrica de co-palabras para detectar los temas más importantes y de más calidad e impacto, SciMAT

permite analizar aspectos diferentes, simplemente escogiendo otra unidad de análisis. Por ejemplo, para analizar la estructura social del campo analizado bastaría con replicar este ejemplo, pero seleccionando los autores como unidad de análisis.

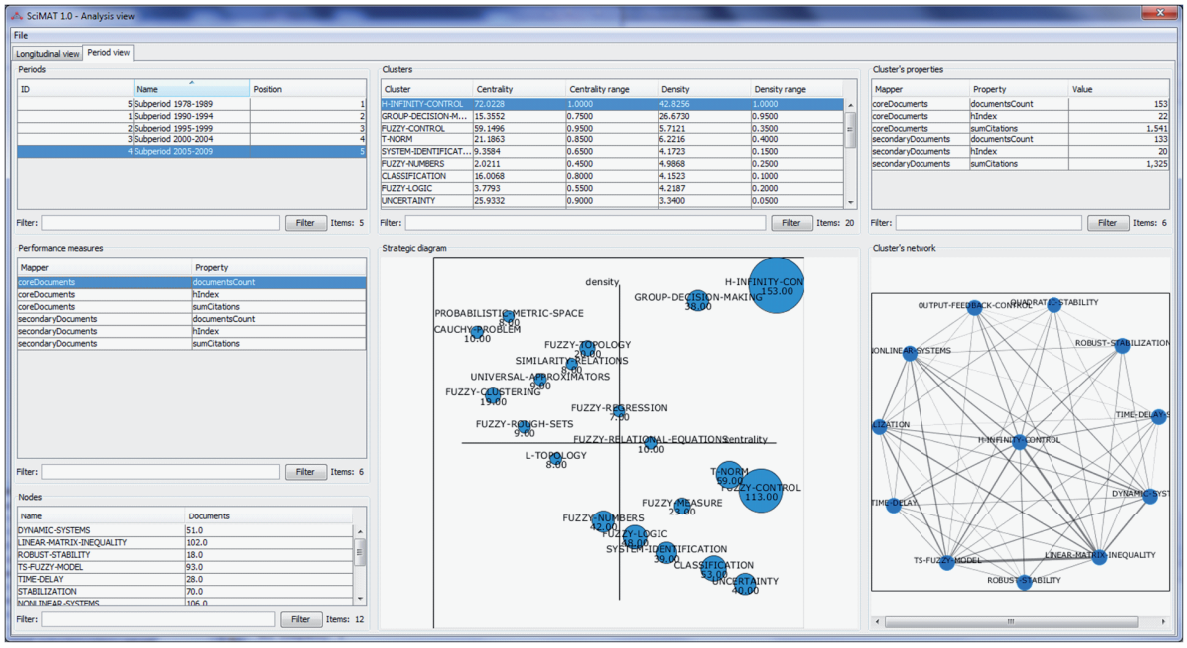


Figura 4.18: Vista de periodos.

Conclusiones y Trabajos Futuros

A continuación resaltaremos las conclusiones del trabajo realizado en la presente memoria de tesis y concluiremos la misma presentando las líneas de trabajo futuro, y los trabajos publicados en revistas científicas y congresos asociados a la tesis.

Resumen y Conclusiones

El análisis de mapas científicos es una técnica bibliométrica que nos permite representar gráficamente las relaciones entre los documentos que las disciplinas o campos científicos concretos publican. Este nos muestra las subáreas de investigación en la que la disciplina ha estado centrada a lo largo de los años, para de este modo, identificar, analizar y visualizar su estructura intelectual, social y conceptual, así como su evolución temporal.

En la presente memoria de tesis nos hemos centrado en el análisis de mapas científicos. Para ello, la memoria de tesis se ha articulado en tres ejes fundamentales:

- Un análisis comparativo de las herramientas software específicamente diseñadas para el análisis de mapas científicos existentes en el mercado.
- Definición de una metodología que nos permita analizar cualquier aspecto de un campo científico, así como su evolución temporal, además de permitir cuantificar los resultados mediante indicadores bibliométricos de actividad, calidad e impacto.

- Desarrollo de una herramienta de código abierto que implementa la metodología anterior, capaz de ayudar al analista a realizar un análisis de mapas científicos, desde la carga de los datos hasta la visualización de los resultados.

Gracias al análisis comparativo realizado entre un grupo de herramientas software (Bibexcel, CiteSpace II, CoPalRed, IN-SPIRE, Loet Leydesdorff's Software, Network Workbench Tool, Science of Science Tool, VantagePoint y VOSViewer -Capítulo 2-), pudimos observar que cada una de ellas contienen características fundamentalmente diferentes. Por ejemplo, algunas de ellas se centran principalmente en la visualización de los resultados, mientras que otras poseen potentes módulos de preprocesamiento. Como conclusión general destacamos que ninguna de las herramientas analizadas puede considerarse como la mejor, por lo que, consecuentemente, observamos que un análisis completo y profundo usando mapas bibliométricos de una disciplina científica, debería ser realizado con la asistencia de diferentes herramientas software, para de este modo, extraer la mayor cantidad de conocimiento posible, así como diferentes perspectivas del mismo. Por ejemplo, el preprocesamiento de los datos objeto de análisis es una tarea muy importante y que influye en la calidad de los resultados obtenidos, por lo que el analista debería emplear herramientas software que le faciliten esta tarea. Asimismo, no todas las herramientas que fueron analizadas permiten trabajar con todas las unidades de análisis (revistas, documentos, referencias, autores y palabras descriptivas), ni permiten construir todos los tipos de redes bibliométricas (co-palabras, co-citation, co-autor, emparejamiento bibliográfico y enlace directo), por lo que el análisis de los diferentes aspectos (social, intelectual o conceptual) de un campo científico deberá realizarse usando varias herramientas.

Por otro lado, cada herramienta software posee su propio conjunto de métodos de análisis (temporal, geoespacial, de redes y detección de estallidos), y ninguna de

las herramientas analizadas, con la excepción de CiteSpace, Science of Science Tool y VantagePoint, implementa todas las técnicas disponibles, por lo que usualmente, es necesario el uso simultáneo de varias de estas herramientas para analizar los diferentes aspectos de la evolución de un campo científico.

Otro aspecto fundamental a tener en cuenta es la capacidad de visualización de los resultados de cada herramienta. Usualmente, cada una de ellas se preocupa de mostrar aspectos diferentes de los análisis realizados, y por tanto, las metáforas de visualización que cada una de ellas implementa suelen ser diferentes. Así, cada herramienta suele mostrar su propia visión e interpretación de los datos analizados. Para evitar sesgos que el uso de una sola metáfora pudiera introducir en la visualización e interpretación de los resultados, sería deseable hacer uso simultáneo de varias de estas metáforas, lo que necesariamente implica la utilización de varias de estas herramientas software.

De este estudio, como principal conclusión, mostramos que el uso cooperativo de diferentes herramientas software proporciona una sinergia positiva, que permitiría al analista extraer, analizar y visualizar desde todos los puntos de vista posible todo el conocimiento escondido en los datos objeto de estudio.

Un segundo resultado de esta memoria de tesis ha sido la definición de una metodología (Capítulo 3) para analizar, cuantificar (mediante indicadores bibliométricos) y visualizar los aspectos sociales, intelectuales y conceptuales de un campo científico, y cómo dichos aspectos evolucionan a lo largo de diversos periodos de tiempo. La metodología propuesta fue fundamentalmente concebida para detectar, analizar y visualizar la evolución conceptual de campos científicos mediante el uso de palabras como unidad de análisis. Sin embargo, gracias a su concepción paramétrica, es fácil la generalización de cada una de sus fases y pasos. En este sentido, se puede utilizar cualquier tipo de unidad de análisis (autores, términos, referencias, etc.), establecer diferentes tipos de

relaciones para crear la red bibliométrica, emplear cualquier tipo de medida de similitud para normalizar la red, crear el mapa mediante diversos algoritmos de clustering, y cuantificar los resultados mediante indicadores bibliométricos de actividad, calidad e impacto, como los vistos en la Sección 1.1. Como técnicas de visualización se utilizan diagramas estratégicos para categorizar los grupos detectados y redes de grupos para ver las interconexiones de los elementos de cada grupo. Además, se emplean mapas de evolución y solapamiento para visualizar la evolución temporal del campo científico analizado.

Como ejemplo, y para validar la metodología, ésta fue aplicada al análisis del área de la Teoría de los Conjuntos Difusos (TCD), empleando para ello los documentos publicados por las dos revistas más importantes (de acuerdo a su Factor de Impacto) y longevas del área: *Fuzzy Sets and Systems* y *IEEE Transactions on Fuzzy Systems*. Tal y como se señaló en la Sección 3.2.1, nuestra metodología fue capaz de identificar adecuadamente los temas básicos del campo científico TCD, para cada uno de los periodos de tiempo analizados, que se correspondían con aquellos que lograron unas mayores cotas de citación e impacto. Adicionalmente, como se mostró en la Sección 3.2.2, nuestra metodología fue capaz de identificar las áreas temáticas (ver la Tabla 3.7), mostrando su evolución a lo largo del tiempo, como es el caso de *FUZZY-CONTROL*, la cual tuvo un comportamiento incremental, o *FUZZY-LOGIC*, que drecrementó con el paso de los años.

Gracias a este estudio pudimos concluir que existe una fuerte correlación entre los temas con mayor centralidad (cuadrantes de la zona derecha de los diagramas estratégicos) y el número de citas recibidas por sus documentos asociados, como se pudo observar en las Figuras 3.6.b, 3.7.b, 3.8.b, 3.9.b y 3.10.b. Esta correlación nos mostró que la metodología propuesta es muy adecuada para su utilización como herramienta de análisis

de cualquier campo científico.

Además, la propuesta metodológica presentada en esta memoria de tesis, ha sido aplicada para analizar el campo de los sistemas inteligentes de transporte [27], las hibridaciones entre la TCD y otras técnicas de inteligencia computacional [59], así como la TCD a nivel internacional y nacional [60].

Por último, como consecuencia directa de la revisión de herramientas para el análisis de mapas científicos presentada en el Capítulo 2 y de la metodología propuesta en el Capítulo 3, se ha desarrollado una herramienta para analizar mapas científicos y que se ha denominado SciMAT (Science Mapping Analysis Tool).

SciMAT es una herramienta de código libre (bajo licencia GPLv3) que permite realizar análisis de mapas científicos en un marco longitudinal, y cuantificar los resultados mediante indicadores bibliométricos, ya que implementa y extiende la metodología propuesta en el Capítulo 3. Además, SciMAT integra en una única herramienta la mayor parte de las ventajas de las herramientas analizadas en el Capítulo 2, mientras que reduce la dependencia de herramientas externas. De este modo, SciMAT proporciona diferentes módulos para ayudar al analista a realizar los distintos pasos de un análisis de mapas científicos, desde la carga de datos, hasta la visualización de los resultados.

SciMAT presenta dos características fundamentales que otras herramientas no poseen (o están muy limitadas):

- Un potente módulo de preprocesamiento de los datos objeto de estudio, y
- El uso de indicadores bibliométricos de impacto y calidad.

Atiendo al módulo de preprocesamiento, éste puede realizar la detección y unificación de elementos similares, división temporal, reducción de datos y de red. Por otro lado, atendiendo a los indicadores bibliométricos de calidad (basados en citas), SciMAT incorpora: índice h [1, 46], índice g [34], índice hg [2] o índice q^2 [17].

Trabajos Futuros

A la vista de los resultados obtenidos, nos planteamos algunas líneas de trabajo futuro que nos permitirán ahondar en cada uno de los aspectos del análisis de mapas científicos.

- Los algoritmos utilizados tanto en nuestra propuesta metodológica como en el desarrollo de la herramienta SciMAT, son algoritmos clásicos de clustering. Actualmente, existen una gran cantidad de técnicas de clustering que se podrían emplear en la construcción del mapa científico, además, recientemente se han desarrollado algunos algoritmos específicos para este problema. Por ello, sería útil, analizar dichas técnicas para determinar cual de ellas funcionan mejor para extraer los subgrupos de una red bibliométrica, para incorporarlas en SciMAT. Además, observando las deficiencias o puntos débiles de dichos algoritmos, se podría desarrollar nuevos algoritmos de clustering que nos permitieran obtener mejores resultados y de forma más rápida.

- Definir medidas de similitud para la normalización de redes bibliométricas que integren medidas de calidad, de modo que aquellas relaciones formadas por unidades de análisis procedentes de documentos muy citados, tengan un mayor peso en la red. Es decir, potenciar las relaciones que asocien elementos altamente citados.

- Las técnicas de visualización son de vital importancia para el correcto análisis y comprensión e interpretación de los resultados. Las técnicas definidas en la metodología e incorporadas en SciMAT son muy útiles y permiten extraer fácilmente conclusiones. No obstante, sería adecuado diseñar nuevas técnicas de visualización que ayuden a la mejor comprensión de los mapas científicos, así como su evolución.

- Ampliar SciMAT con nuevos filtros para leer información de más fuentes de datos bibliográficas, con técnicas que faciliten aún más el preprocesamiento, y con nuevos algoritmos que flexibilicen el análisis de mapas científicos.

- Como vimos en la Sección 1.2.6, se pueden aplicar diversos análisis sobre los mapas científicos. Particularmente, el análisis geoespacial es de gran utilidad para situar geográficamente los elementos de un mapa científico. Por ejemplo, se podría determinar en qué temáticas están centradas diferentes zonas geográficas, y cómo éstas han evolucionado a lo largo del tiempo. Incluso, se podría estudiar las relaciones de colaboración internacional. Por este motivo, creemos necesario que SciMAT, en un futuro, incorpore el análisis geoespacial.

- Por último, aunque la presente memoria de tesis se ha centrado en el análisis mapas científicos basados en publicaciones científicas, todas las técnicas, metodologías y herramientas diseñadas pueden aplicarse para realizar vigilancia tecnológica. Particularmente, los mapas científicos pueden realizarse con patentes, por lo que de este modo, pueden emplearse para estudiar los aspectos estructurales de un campo tecnológico particular. Por este motivo, sería adecuado que en un futuro SciMAT incorpore los métodos necesarios para desarrollar mapas científicos basados en patentes.

Publicaciones Asociadas a la Memoria de Tesis Doctoral

En esta Sección listamos las publicaciones tanto en revistas científicas como en congresos nacionales e internacionales asociados a la presente memoria de tesis doctoral.

Las publicaciones en revistas científicas asociados a la memoria de tesis han sido los siguientes:

- M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma, F. Herrera, Science Mapping Software Tools: Review, Analysis and Cooperative Study among Tools. *Journal of the American Society for Information Science and Technology*, 62:7, pp. 1382-1402 (2011) doi: 10.1002/asi.21525.

- M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma, F. Herrera, An Approach for Detecting, Quantifying, and Visualizing the Evolution of a Research Field: A Practical Application to the Fuzzy Sets Theory Field. *Journal of Informetrics* 5:1, pp. 146-166 (2011). doi: 10.1016/j.joi.2010.10.002.

- M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma, F. Herrera, SciMAT: A new Science Mapping Analysis Software Tool. **Submitted to** *Journal of the American Society for Information Science and Technology.*

- M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma, F. Herrera, A Note on the ITS Topic Evolution in the Period 2000-2009 at T-ITS. *IEEE Transactions on Intelligent Transportation Systems* **in press**.

- A.G. López-Herrera, M.J. Cobo, E. Herrera-Viedma, F. Herrera, A Bibliometric Study about the Research Based on Hybridating the Fuzzy Logic Field and the Other Computational Intelligent Techniques: A Visual Approach. *Internacional Journal of Hybrid Intelligent Systems* 17:7 (2010) 17-32. doi:10.3233/HIS-2010-0102.

- A.G. López-Herrera, M.J. Cobo, E. Herrera-Viedma, F. Herrera, R. Bailón, E. Jiménez-Contreras, Visualization and Evolution of the Scientific Structure of

Fuzzy Sets Research in Spain. *Information Research* 14:4, paper 421 (2009), Available at http://InformationR.net/ir/14-4/paper421.html.

Las trabajos publicados en congresos asociados a la memoria de tesis han sido los siguientes:

- A.G. López-Herrera, M.J. Cobo, E. Herrera-Viedma, F. Herrera, Visualizing the Hybridizations Between the Fuzzy Logic Field and the Other Soft-Computing Techniques. Proceedings of the 8th International Conference on Hybrid Intelligent Systems (HIS08), Barcelona (Spain), 252-257, 10-12 September 2008.

- A.G. López-Herrera, M.J. Cobo, E. Herrera-Viedma, F. Herrera, A Study on the Evolution of Fuzzy Sets Based Research in Spain Based on Bibliometric Maps. Proceedings of the XIV Congreso Español Sobre Tecnologías Y Lógicas Fuzzy (ESTYLF 2008), pp. 265-270. ISBN: 978-84-691-5807-4, Mieres-Langreo (Spain), 17-19 September 2008.

Apéndice A

Un Ejemplo Práctico de Utilización del API de SciMAT

SciMAT dispone de una potente API que permite al usuario avanzado realizar un análisis de mapas científicos personalizado, mediante la utilización de los diversos métodos implementados en ella.

En particular, en este anexo mostramos cómo realizar el estudio llevado a cabo en las Secciones 3.2 y 4.4 utilizando exclusivamente la API. El ejemplo parte de una base de conocimiento existente, en donde los grupos de términos ya han sido creados.

```
1  package scimat.apps;
2
3  import java.util.ArrayList;
4  import scimat.api.analysis.network.labeller.*;
5  import scimat.api.analysis.network.statistics.*;
6  import scimat.api.analysis.performance.*;
7  import scimat.api.analysis.performance.docmapper.*;
8  import scimat.api.analysis.performance.quality.*;
9  import scimat.api.analysis.performance.quantity.*;
10 import scimat.api.analysis.temporal.*;
11 import scimat.api.dataset.Dataset;
12 import scimat.api.dataset.UndirectNetworkMatrix;
13 import scimat.api.dataset.datasetbuilder.*;
```

```
14 import scimat.api.dataset.networkbuilder.*;
15 import scimat.api.mapping.clustering.*;
16 import scimat.api.mapping.clustering.result.ClusterSet;
17 import scimat.api.preprocessing.reduction.data.FilterItemByFrequency;
18 import scimat.api.preprocessing.reduction.network.FilterByEdgeValue;
19 import scimat.api.similaritymeasure.*;
20 import scimat.api.similaritymeasure.direct.*;
21 import scimat.model.knowledgebase.entity.Period;
22 import scimat.project.CurrentProject;
23
24 /**
25  *
26  * @author mjcobo
27  */
28 public class AnaylsisApp {
29
30   /**
31    * @param args the command line arguments
32    */
33   public static void main(String[] args) {
34
35     int i;
36     ArrayList<Period> periods;
37     Period period;
38     ArrayList<Dataset> datasets;
39     DatasetBuilder datasetBuilder;
40     FilterItemByFrequency frequencyFilter;
41     NetworkBuilder networkBuilder;
42     UndirectNetworkMatrix[] networks;
43     FilterByEdgeValue coOccurrenceFilter;
44     Normalizer normalizer;
45     ClusteringAlgorithm clusterAlgorithm;
46     ArrayList<ClusterSet> clusterSets;
47     NodeLabelSetter nodeLabelAssigner;
48     ClusterSetLabelSetter clusterSetLabelAssigner;
49     ClusterSetDocumentsSetter clusterSetDocumentsSetter;
50     ClusterSetAggregationDocumentsMeasureSetter
           clusterSetAggregationDocumentsMeasureSetter;
51     ClusterSetNetworkMeasureSetter clusterSetNetworkMeasureSetter;
```

```
52      CalculateNormalizedRange calculateNormalizedRange;
53      EvolutionMapBuilder evolutionMapBuilder;
54      EvolutionMap evolutionMap;
55      OverlappingMapBuilder overlappingMapBuilder;
56      OverlappingMap overlappingMap;
57
58      try {
59
60        CurrentProject.getInstance().loadProyect("FSSandIEEETFS.sqlite");
61
62        periods = CurrentProject.getInstance().getFactoryDAO().getPeriodDAO().
              getPeriodsOrderedByPosition();
63
64        datasetBuilder = new DatasetBasedOnWordsBuilder(CurrentProject.getInstance().
              getKnowledgeBase(), true,false,false);
65
66        datasets = new ArrayList<Dataset>();
67
68        // Build the data sets base on words
69        for (i = 0; i < periods.size(); i++) {
70
71          period = periods.get(i);
72
73          datasets.add(datasetBuilder.execute(CurrentProject.getInstance().getFactoryDAO
                ().getPeriodDAO().getPublishDates(period.getPeriodID())));
74        }
75
76        // Filter the datasets
77        frequencyFilter = new FilterItemByFrequency(3);
78        frequencyFilter.execute(datasets.get(0));
79
80        frequencyFilter = new FilterItemByFrequency(4);
81        frequencyFilter.execute(datasets.get(1));
82
83        frequencyFilter = new FilterItemByFrequency(4);
84        frequencyFilter.execute(datasets.get(2));
85
86        frequencyFilter = new FilterItemByFrequency(4);
87        frequencyFilter.execute(datasets.get(3));
```

```
88
89      frequencyFilter = new FilterItemByFrequency(5);
90      frequencyFilter.execute(datasets.get(4));
91
92      // Build the network base on co-occurence
93      networks = new UndirectNetworkMatrix[datasets.size()];
94
95      networkBuilder = new NetworkCoOccurrenceBuilder(datasets.get(0));
96      networks[0] = networkBuilder.execute();
97
98      networkBuilder = new NetworkCoOccurrenceBuilder(datasets.get(1));
99      networks[1] = networkBuilder.execute();
100
101     networkBuilder = new NetworkCoOccurrenceBuilder(datasets.get(2));
102     networks[2] = networkBuilder.execute();
103
104     networkBuilder = new NetworkCoOccurrenceBuilder(datasets.get(3));
105     networks[3] = networkBuilder.execute();
106
107     networkBuilder = new NetworkCoOccurrenceBuilder(datasets.get(4));
108     networks[4] = networkBuilder.execute();
109
110     // Filter the network
111     coOccurrenceFilter = new FilterByEdgeValue(2);
112     coOccurrenceFilter.execute(networks[0]);
113
114     coOccurrenceFilter = new FilterByEdgeValue(3);
115     coOccurrenceFilter.execute(networks[1]);
116
117     coOccurrenceFilter = new FilterByEdgeValue(3);
118     coOccurrenceFilter.execute(networks[2]);
119
120     coOccurrenceFilter = new FilterByEdgeValue(3);
121     coOccurrenceFilter.execute(networks[3]);
122
123     coOccurrenceFilter = new FilterByEdgeValue(4);
124     coOccurrenceFilter.execute(networks[4]);
125
126     // Normalize the network
```

```
127         normalizer = new CoOccurrenceNormalizer(new EquivalenceIndexMeasure());
128
129         for (i = 0; i < datasets.size(); i++) {
130
131           normalizer.execute(datasets.get(i), networks[i]);
132         }
133
134         // Apply a cluster algorithm
135         clusterAlgorithm = new CentersSimpleGroupingAlgorithm(3, 12);
136
137         clusterSets = new ArrayList<ClusterSet>();
138
139         for (i = 0; i < networks.length; i++) {
140
141           clusterSets.add(clusterAlgorithm.execute(networks[i]));
142         }
143
144         // Set a label to each node
145         for (i = 0; i < clusterSets.size(); i++) {
146
147           nodeLabelAssigner = new NodeLabelSetter(new BasicNodeLabeller(datasets.get(i))
                  );
148           nodeLabelAssigner.execute(clusterSets.get(i).getWholeNetwork().getNodes(), "
                  name");
149         }
150
151         // Set a label to each cluster
152         for (i = 0; i < clusterSets.size(); i++) {
153
154           clusterSetLabelAssigner = new ClusterSetLabelSetter(new
                  BasicClusterLabellerBasedOnMainNode(datasets.get(i)));
155           clusterSetLabelAssigner.execute(clusterSets.get(i), "name");
156         }
157
158         // Calculate networks measures for each cluster
159         clusterSetNetworkMeasureSetter = new ClusterSetNetworkMeasureSetter(new
              CallonCentrality());
160         clusterSetNetworkMeasureSetter = new ClusterSetNetworkMeasureSetter(new
              CallonDensity());
```

```
161         calculateNormalizedRange = new CalculateNormalizedRange();
162
163         for (i = 0; i < clusterSets.size(); i++) {
164
165           clusterSetNetworkMeasureSetter.execute(clusterSets.get(i), "callonCentrality")
                  ;
166
167           clusterSetNetworkMeasureSetter.execute(clusterSets.get(i), "callonDensity");
168
169           calculateNormalizedRange.calculateMeasures(clusterSets.get(i), "
                  callonCentrality", "callonCentralityRange");
170           calculateNormalizedRange.calculateMeasures(clusterSets.get(i), "callonDensity"
                  , "callonDensityRange");
171         }
172
173         // Map documents to cluster
174         for (i = 0; i < clusterSets.size(); i++) {
175
176           clusterSetDocumentsSetter = new ClusterSetDocumentsSetter(new
                  CoreDocumentMapper(clusterSets.get(i).getWholeNetwork(), datasets.get(i)));
177           clusterSetDocumentsSetter.execute(clusterSets.get(i), "coreDocuments");
178
179           clusterSetDocumentsSetter = new ClusterSetDocumentsSetter(new
                  SecondaryDocumentMapper(datasets.get(i)));
180           clusterSetDocumentsSetter.execute(clusterSets.get(i), "secondaryDocuments");
181         }
182
183         // Calculate bibliometric indicator
184         for (i = 0; i < clusterSets.size(); i++) {
185
186           clusterSetAggregationDocumentsMeasureSetter = new
                  ClusterSetAggregationDocumentsMeasureSetter(new
                  DocumentCountAggregationMeasure());
187           clusterSetAggregationDocumentsMeasureSetter.execute(clusterSets.get(i), "
                  coreDocuments", "coreDocumentsCount");
188           clusterSetAggregationDocumentsMeasureSetter.execute(clusterSets.get(i), "
                  secondaryDocuments", "secondaryDocumentsCount");
189
190           clusterSetAggregationDocumentsMeasureSetter = new
```

```
                ClusterSetAggregationDocumentsMeasureSetter(new
                SumCitationAggregationMeasure(datasets.get(i)));
191             clusterSetAggregationDocumentsMeasureSetter.execute(clusterSets.get(i), "
                coreDocuments", "coreDocumentsCitations");
192             clusterSetAggregationDocumentsMeasureSetter.execute(clusterSets.get(i), "
                secondaryDocuments", "secondaryDocumentsCitations");
193
194             clusterSetAggregationDocumentsMeasureSetter = new
                ClusterSetAggregationDocumentsMeasureSetter(new HIndex(datasets.get(i)));
195             clusterSetAggregationDocumentsMeasureSetter.execute(clusterSets.get(i), "
                coreDocuments", "coreDocumentsH-Index");
196             clusterSetAggregationDocumentsMeasureSetter.execute(clusterSets.get(i), "
                secondaryDocuments", "secondaryDocumentsH-Index");
197         }
198
199         evolutionMapBuilder = new EvolutionMapBuilder(new OverlappingMeasure(new
                InclusionIndexMeasure()));
200         evolutionMap = evolutionMapBuilder.buildEvolutionMap(clusterSets);
201
202         overlappingMapBuilder = new OverlappingMapBuilder(new OverlappingMeasure(new
                JaccardIndexMeasure()));
203         overlappingMap = overlappingMapBuilder.buildOverlappingMap(datasets);
204
205     } catch (Exception e) {
206
207         e.printStackTrace(System.err);
208     }
209   }
210 }
```

Apéndice B

Guía de Usuario de SciMAT

En este anexo se muestra la guía o manual de usuario de SciMAT. En él se explican detalles concretos del uso de la herramienta, dejando de un lado los detalles técnicos o metodológicos, los cuales han sido descritos profundamente en los Capítulos 3 y 4.

Como comentamos en la Sección 4.3.1, la interfaz gráfica de SciMAT se divide en tres módulos diferentes: i) un módulo dedicado a la gestión de la base de conocimiento y sus entidades, ii) un módulo encargado de realizar el análisis de mapas científicos, y iii) un módulo para visualizar los mapas y resultados generados. Estos módulos permiten al analista desarrollar los diferentes pasos del flujo de trabajo de un análisis de mapas científicos. A continuación describimos el uso de cada uno de estos módulos.

Tenemos que señalar que el manual de usuario se ha realizado en inglés para facilitar la lectura y comprensión de éste por parte de la comunidad científica internacional.

B.1. Knowledge Base Manager

The module to manage the knowledge base is responsible for building it, importing the data from different bibliographical sources, and cleaning and fixing the possible errors in the entities. It can be considered as a first stage in the preprocessing step.

The first step in this module is to build a new project or load an existing one. It can be done through the menu *File* (Figure B.1) or using the buttons of the toolbar.

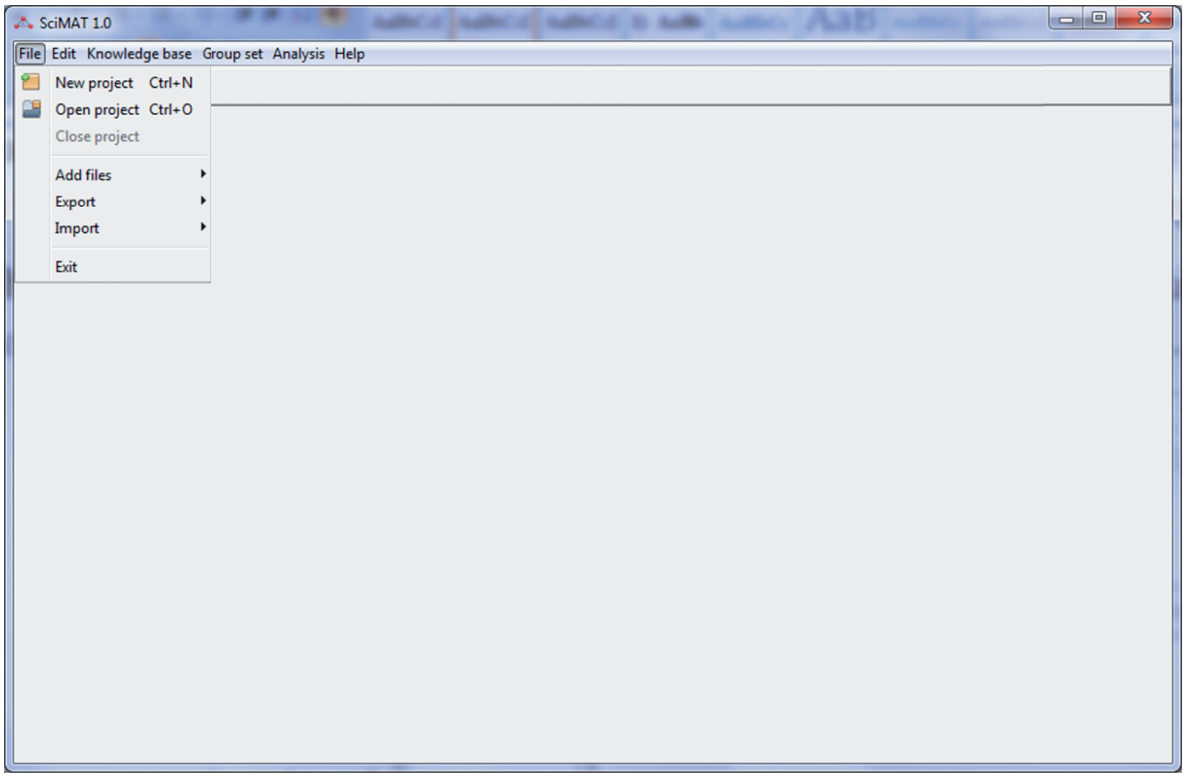


Figura B.1: Menu *File*.

If a new project is selected, a new window will appear asking for the path where the knowledge base file will be stored and the name of the file. We can give any extension for the file.

SciMAT uses the SQLite database engine in order to store the knowledge base built in the previous step. Thanks to these capabilities, the knowledge base can be opened with any database browser that reads SQLite files.

Once we have a project loaded (new or existing), the options under the menus *Knowledge base* and *Group set* will be enabled. Furthermore, the import, export and add options under the menu *File* (Figure B.2) will be enabled too.

The add files option allows the user to add bibliographical information, exported from bibliographical databases, to the knowledge base. Particularly, SciMAT is able to read bibliographical information exported in ISI Web of Knowledge format (ISI-CE) or RIS (Scopus) format.

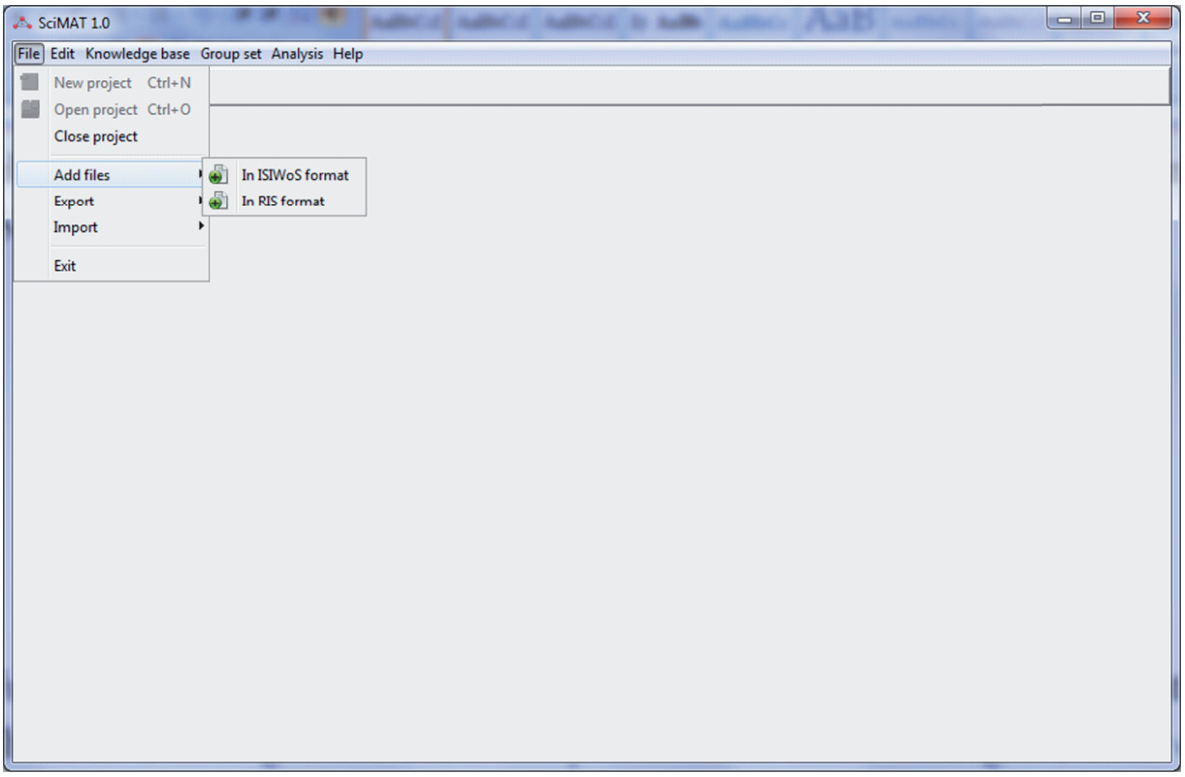


Figura B.2: Menu *Add Files*.

Under the menu *Knowledge base* (Figure B.3) the manager for the sixteen entities can be found. There is a manager for each entity. Thank to these managers, each entity can be edited and its attributes and associations can be modified.

We should point out that all the managers have the same structure, on the left-side a list of entities is shown, and on the right-side the fields of the selected entity and its relations with other entities are shown.

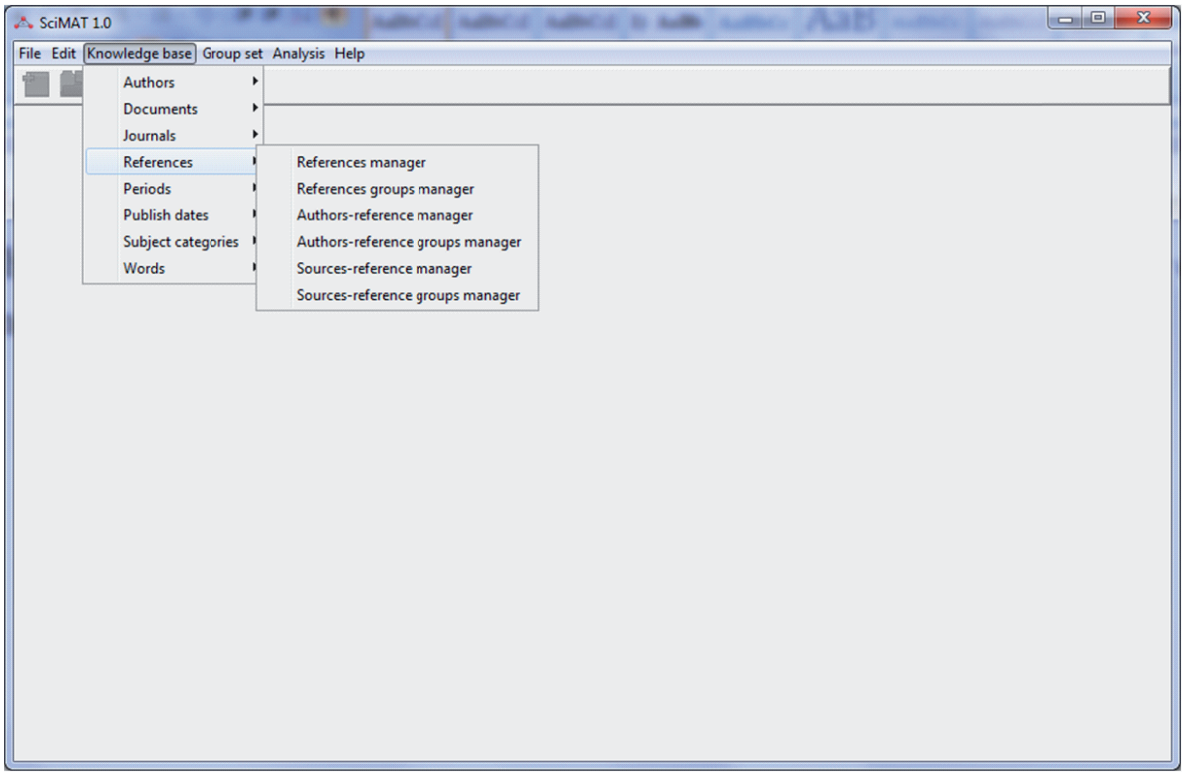


Figura B.3: Menu *Knowledge base*.

As an example the *Document's manager* (Figure B.4) is shown below. In the list of documents (left-side), one of the most cited articles in the knowledge base is selected, and on the right-side its associated information (title, abstract, publication data, citations, etc.) and associations are shown.

The manager allows us to add a new Document (filling manually each attribute, see Figure B.5), delete a set of Documents and join (*move to* button) a set of Document. Furthermore, the user can use the filter box to introduce a regular expression and find the wanted entities.

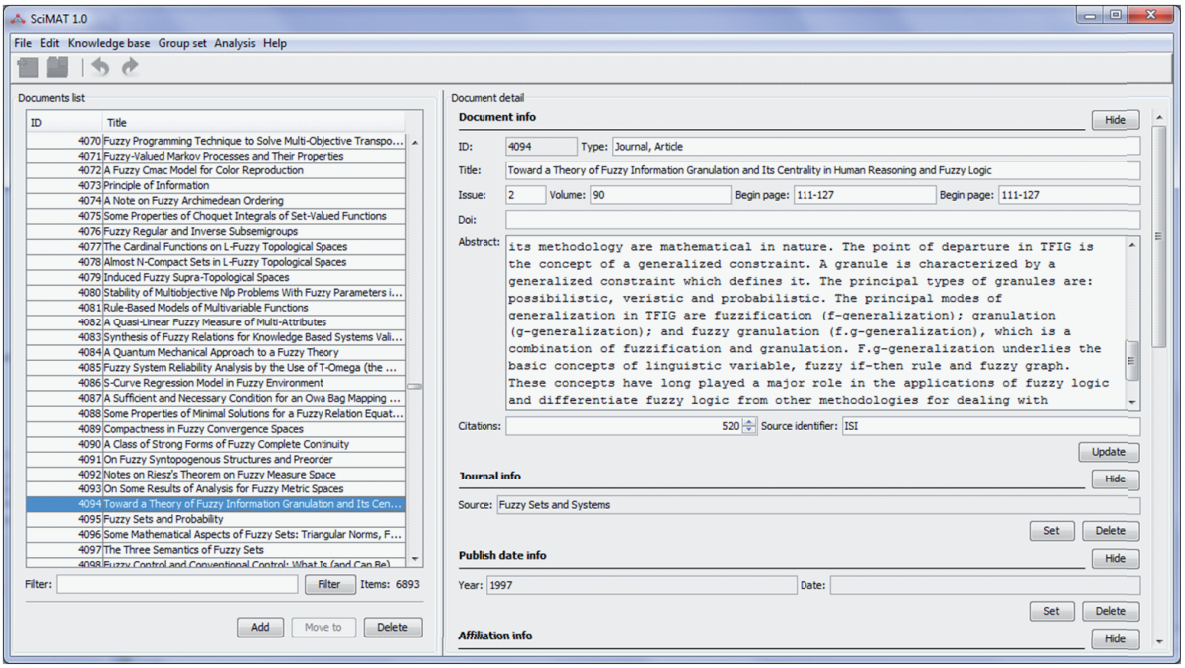


Figura B.4: Document's manager.

The *move to* or *join* capability allows us to join a set of entities under other. It is especially useful when we are working with groups. Once we have selected the set of entities that we want to join, a new dialog will appears (Figure B.6). In this dialog, the user has to select the entity under the remaining entities will be joined. The main or target entity will maintain its associations with others entities and the associations of the joined entities. As an example, the six selected documents below will be joined under the document with ID 4070. So, the target document will be associated with the words, references, affiliations, etc of the remaining five documents.

The right-side of the manager allows us to edit the field of the selected entity or its associations with other entities.

SciMAT incorporate powerful capabilities to perform a de-duplicating step over the unit of analysis by means of groups. To do that, there are five especial managers to perform this task (they can be found under the menu *Group set*). Similarly to the

entity manager, the manual groups set manager have a common structure: the left-side shows a list of defined groups, and the right-side shows the entities associated with the selected entity (header-table) and the entities without groups (foot-table).

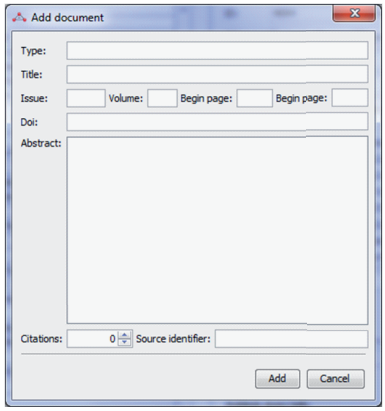


Figura B.5: Add new Document.

As an example, the manager to perform the manual set of the *Words Groups* is shown in Figure B.7. It can be seen that a particular word group with the name *GROUP-DECISION-MAKING* has been defined (left-side). It can also be observed that this word group collects four different word names or variants (top right-side) for the concept. The lower right-side allows the user to add more variants of the concept *GROUP-DECISION-MAKING*.

The manual set group manager allows us to add a new group, delete a set of groups, join a set of groups under other, and finally edit them. Furthermore, this manager allows us to add a set of entities to a selected group, or delete a set of entities from a group. This can be done using the up-row and down-row from the middle of the right-panel.

The groups can also be added from a set of entities without group through the buttons “to new group” and “to different group”. The former build a new group adding the selected entities to it. The name of the group can be chosen from the entities or can be given by the user (Figure B.8). In the latter, each entity will be associated with

a group and the group's name will be the main field of the entity.

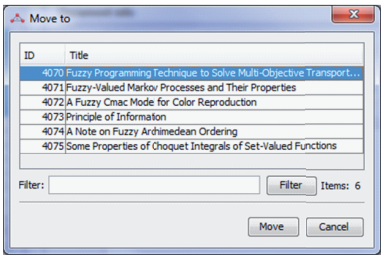


Figura B.6: Move to or join to Documents.

Additionally, this module incorporates methods to help the analyst in the de-duplicating process, such as, finding similar items by plural or by Levenshtein distance, or importing the groups and theirs associated items from a file (in XML format).

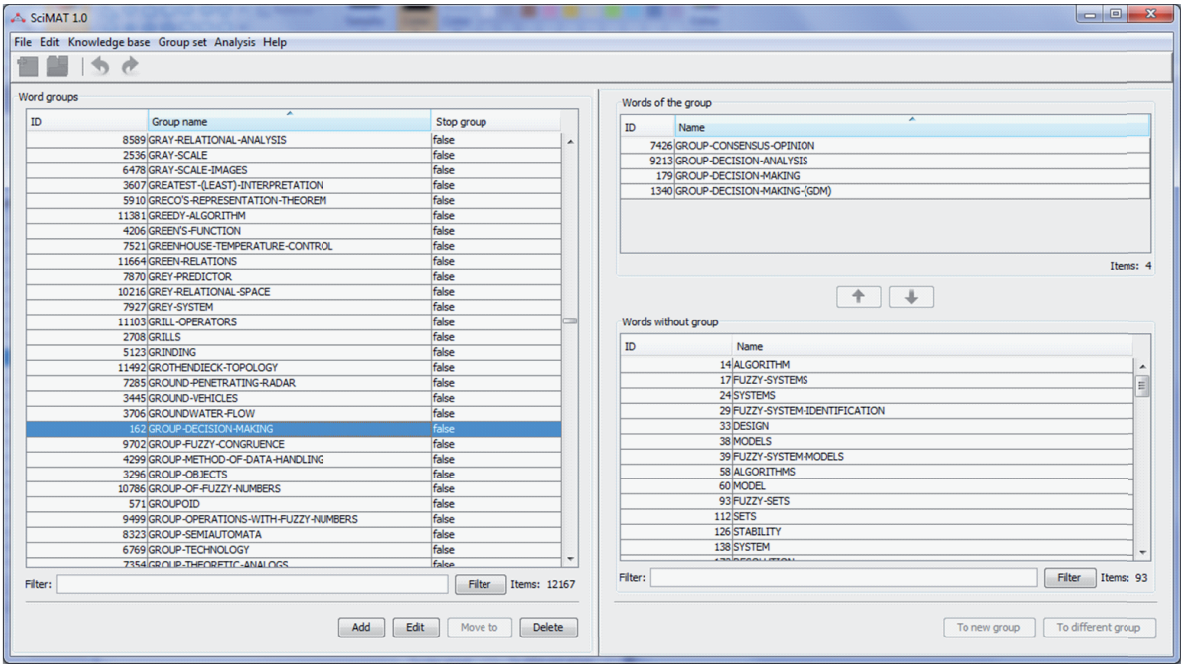


Figura B.7: Words Groups manual set manager.

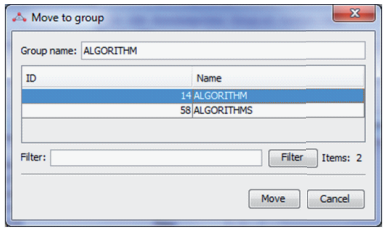


Figura B.8: Build new groups from words without group.

B.2. Science mapping analysis wizard

Once the knowledge base is ready for the science mapping analysis, this module helps the user to configure the process, choosing the methods and algorithms that have to be used by SciMAT to build the maps.

This module is implemented through a wizard and it is composed of eleven consecutive steps:

In the **first step** (Figure B.9) the user has to select the periods that he/she wants to analyze. Each period will produce a map. These periods will be used in the longitudinal or temporal analysis in order to study the structural evolution of the field. The position of the period will indicate the order in which the period will be used in the process. So, the period with lowest position will be processed first and will be the first in the longitudinal results.

The **second step** (Figure B.10) is the selection of the unit of analysis. As the unit of analysis the user can select any of the five groups existing in the knowledge base: Author Group, Author-Reference Group, Source-Reference Group, Reference Group, or Word Group. Only one of them can be selected. If the Word Group has been selected, the role of the word with which the user wants to perform the analysis has to be chosen. In this case, the user has to select the author's word, source's word or extracted word,

or indeed, any combination of them.

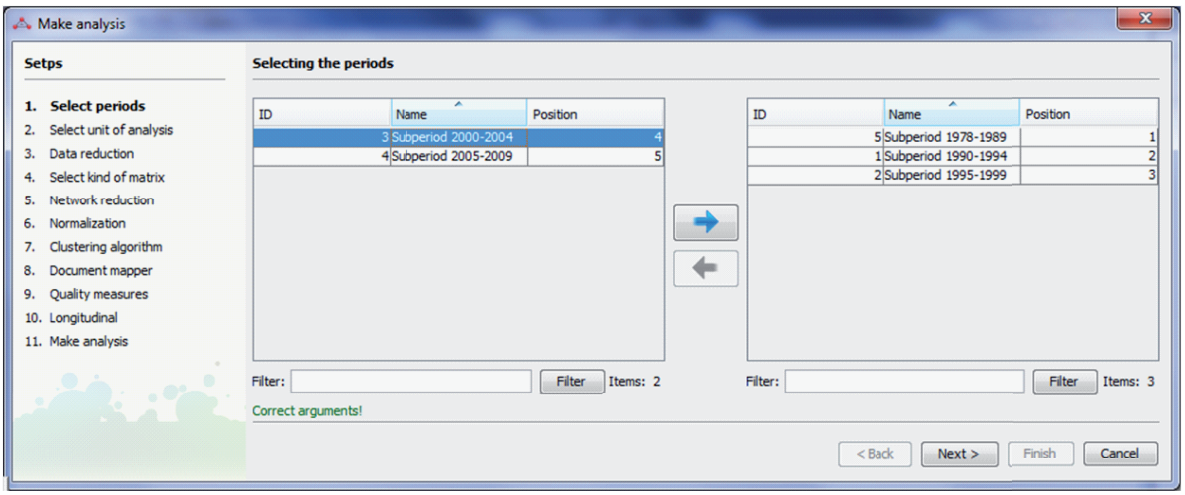


Figura B.9: Select periods (Step 1).

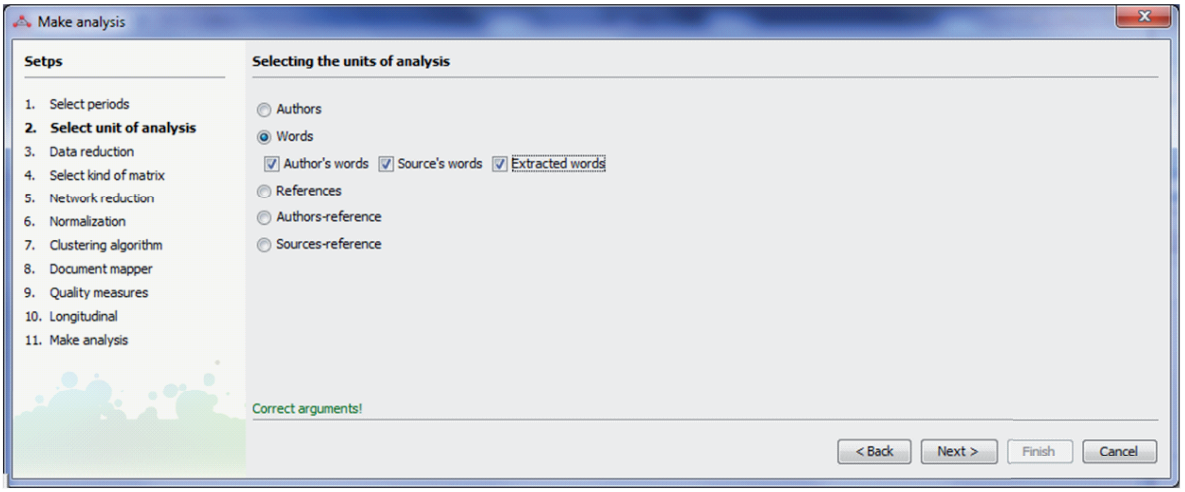


Figura B.10: Select unit of analysis (Step 2).

The **third step** (Figure B.11) is the data reduction. SciMAT allows the data to be filtered using a minimum frequency threshold. For each selected period, a threshold must also be selected. That is, only the item that appears in almost n documents in a given period will be taken into account.

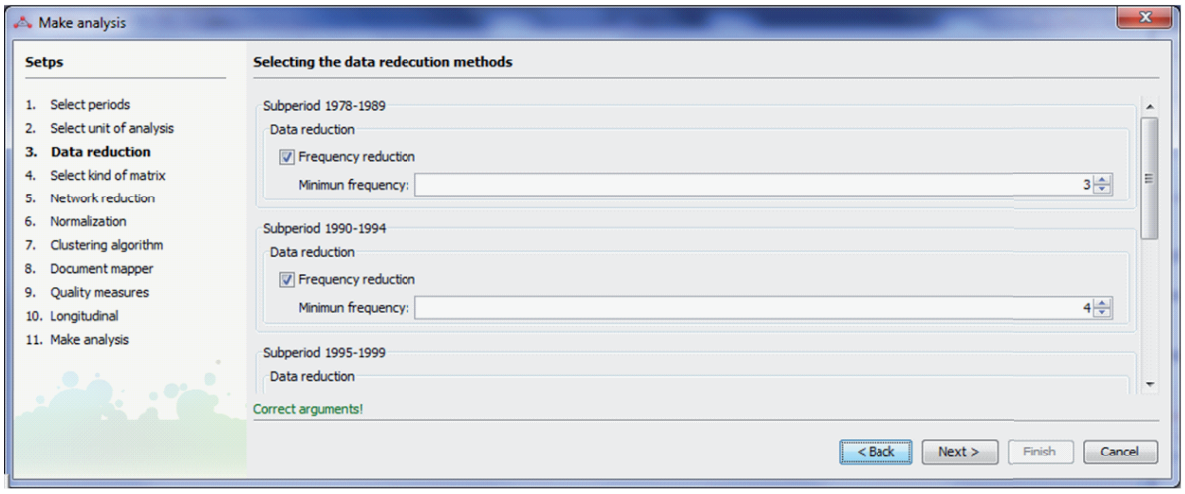


Figura B.11: Select data reduction method (Step 3).

The **fourth step** (Figure B.12) is the selection of the way in which the network will be built: co-occurrence or coupling. Using co-occurrence, co-author, co-word, co-citation (using the references), author co-citation (using the authors-reference), and journal co-citation (using the sources-reference) network can be built. Otherwise, the coupling can be used in a basic or aggregated way. In the former, a document coupling network can be built using the selected unit of analysis as coupled items. That is, if the Reference has been chosen, a document bibliographic coupling network will be built. In the latter, an author or journal coupling network can be made, and again, the coupled item will be the selected unit of analysis.

The **fifth step** (Figure B.13) is the network reduction. SciMAT allows the network to be filtered using a minimum edge value threshold. For each selected period, a threshold value must be set. That is, only the edges with a value greater or equal to n in a given period will be taken into account.

The **sixth step** (Figure B.14) is the selection of the similarity measure used to normalize the network. SciMAT allows the user to choose the similarity measures com-

monly used in the literature to normalize networks: Association Strength, Equivalence Index, Inclusion Index, Jaccard's Index and Salton's Cosine.

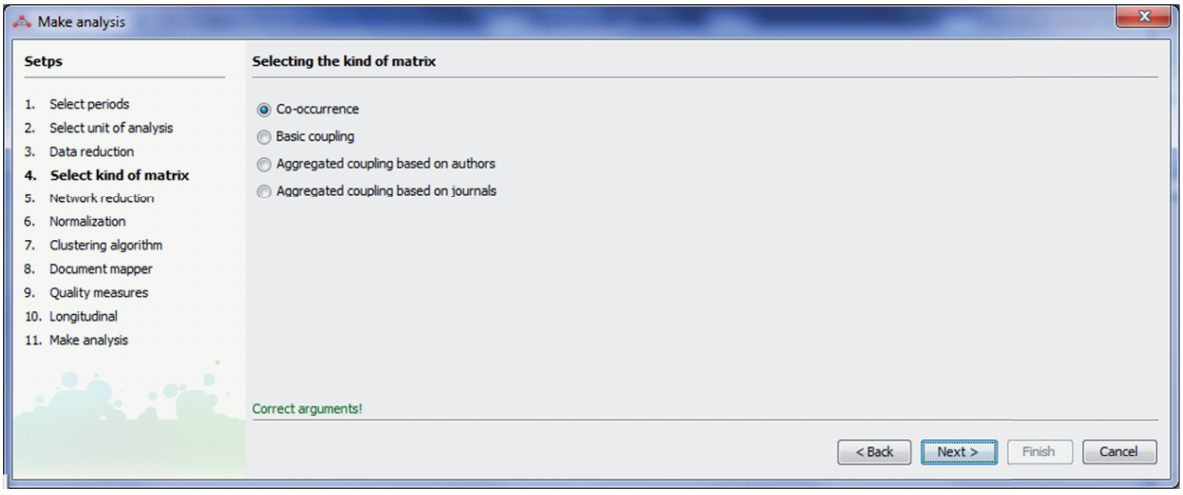


Figura B.12: Select kind of bibliometric network (Step 4).

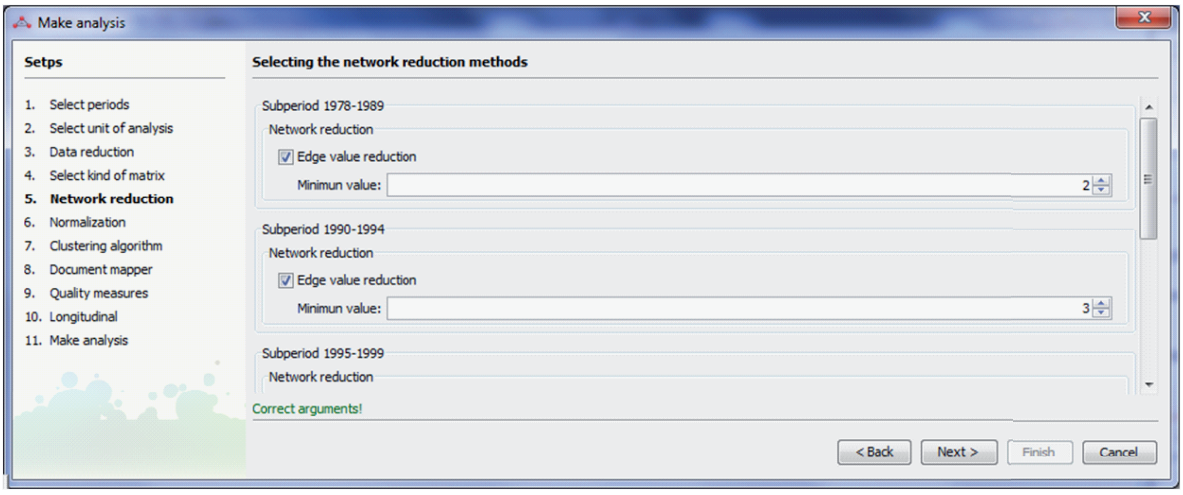


Figura B.13: Select network reduction threshold (Step 5).

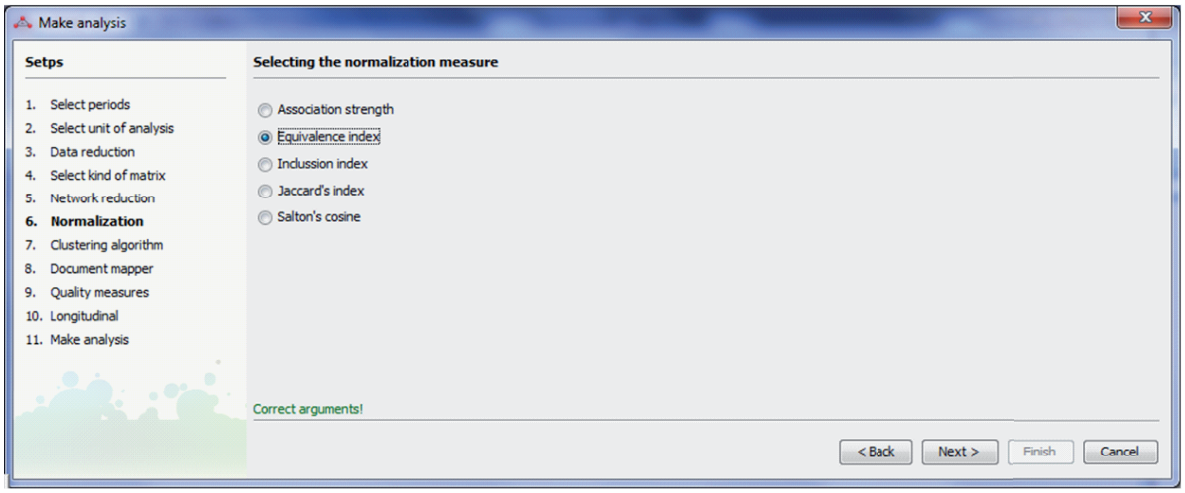


Figura B.14: Select normalization measure (Step 6).

The **seventh step** (Figure B.15) is the selection of the clustering algorithm used to get the map and its associated clusters or subnetworks. Different clustering methods are available in SciMAT, such as, the Simple Centers Algorithm (Coulter et al., 1998), Single-linkage (Small & Sweeney, 1985) and variants such as Complete-linkage, Average-linkage and Sum-linkage.

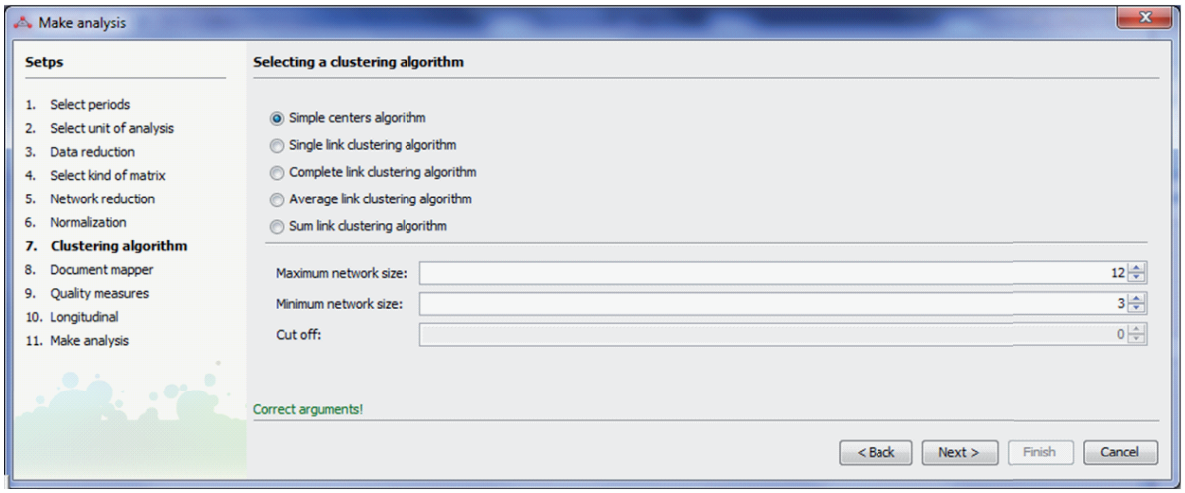


Figura B.15: Select clustering algorithm (Step 7).

The **eighth step** (Figure B.16) is the selection of the documents mapper used in the performance analysis. SciMAT incorporates five different document mappers for co-occurrence networks: i) core mapper, ii) intersection mapper which adds the documents that have all the items of the cluster, iii) k-core mapper which adds the documents that have at least k items in common with the cluster, iv) secondary mapper, and v) union mapper which adds documents that have at least one item in common with the cluster (this is the union of the documents associated with the core and secondary mappers). For coupling networks, SciMAT has two kinds of document mappers depending on the kind of coupling used. That is, if a basic coupling has been selected (each item of the cluster will be a document), the basic coupling document mapper is the only one available, which adds the items of the cluster as documents. If an aggregated coupling is selected, the aggregated coupling document mapper can be selected, which adds the documents associated with its items to each cluster (author's or journal's oeuvres). We should point out that each item or node of the cluster also has a set of associated documents. These documents correspond to the set of items associated with the item in the corresponding dataset.

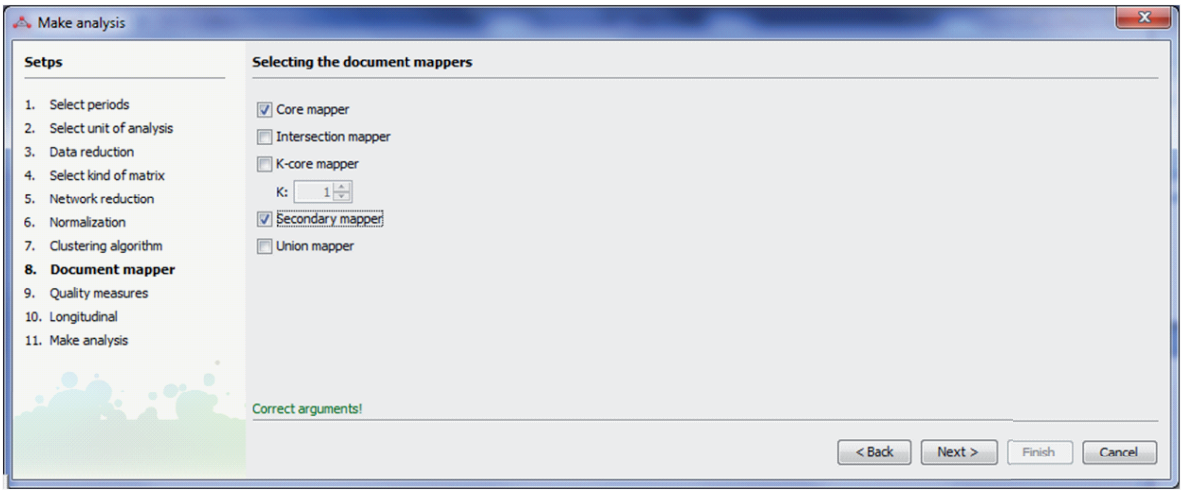


Figura B.16: Select documents mapper (Step 8).

The **ninth step** (Figure B.17) is the selection of the performance and quality bibliometric measures. SciMAT adds by default the number of documents as performance measure. Moreover, the citations of a set of documents are used in order to assess the quality and impact of the clusters. In this sense, basic measures such as the sum, minimum, maximum and average citations, or complex measures such as the h-index (Alonso et al., 2009; Hirsch, 2005), g-index (Egghe, 2006), hg-index (Alonso et al., 2010) or q2-index (Cabrerizo et al., 2010) can be selected.

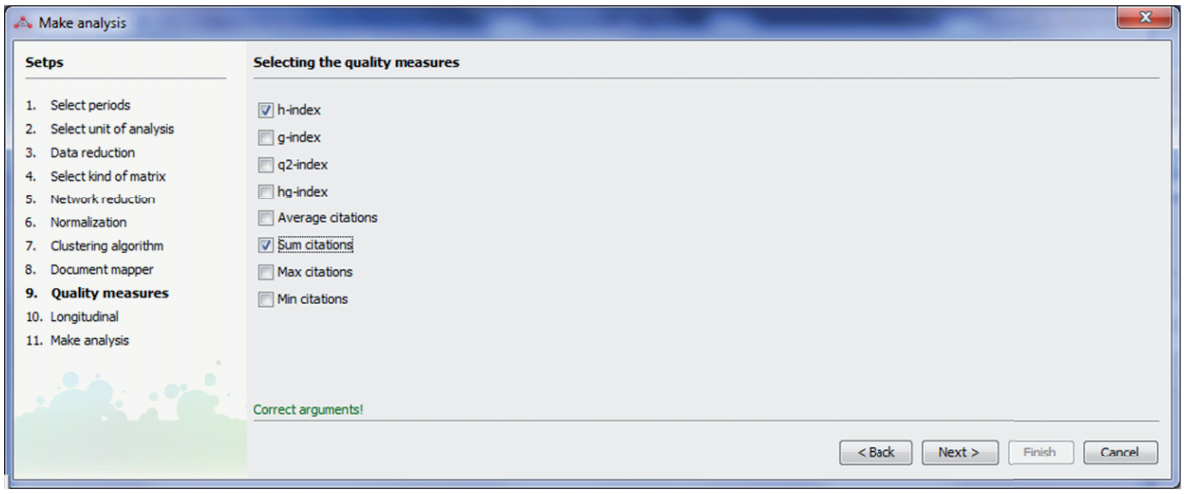


Figura B.17: Select bibliometric quality measures (Step 9).

The **tenth step** (Figure B.18) is the selection of the similarity measure used to build the evolution map and the overlapping map. SciMAT allows us to choose between: Association Strength, Equivalence Index, Inclusion Index, Jaccard's Index and Salton's Cosine.

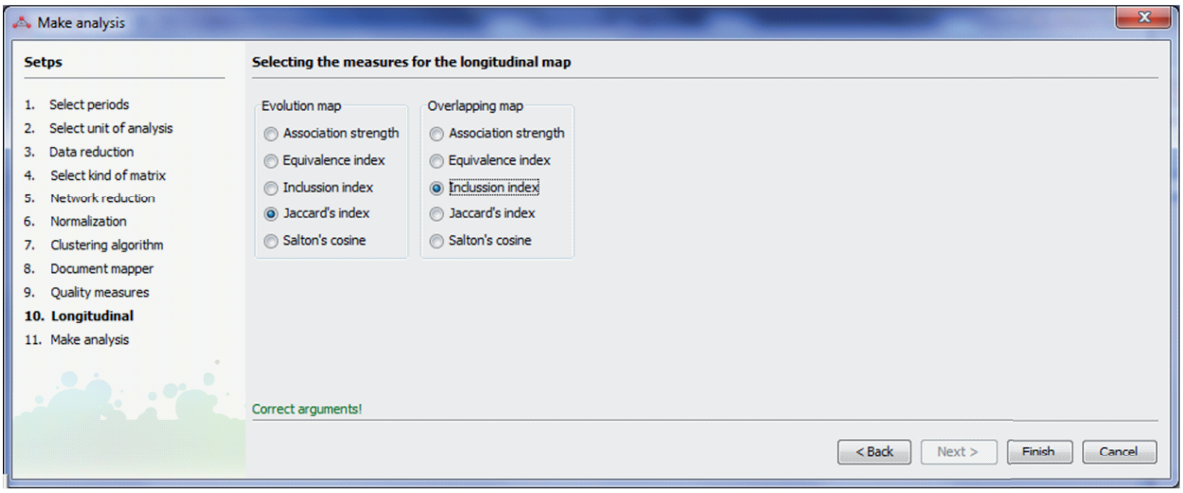


Figura B.18: Select similarity measures for the longitudinal analysis (Step 10).

Finally, the **eleventh step** (Figure B.19) is responsible to perform the science mapping analysis. This process can be cancelled at any time.

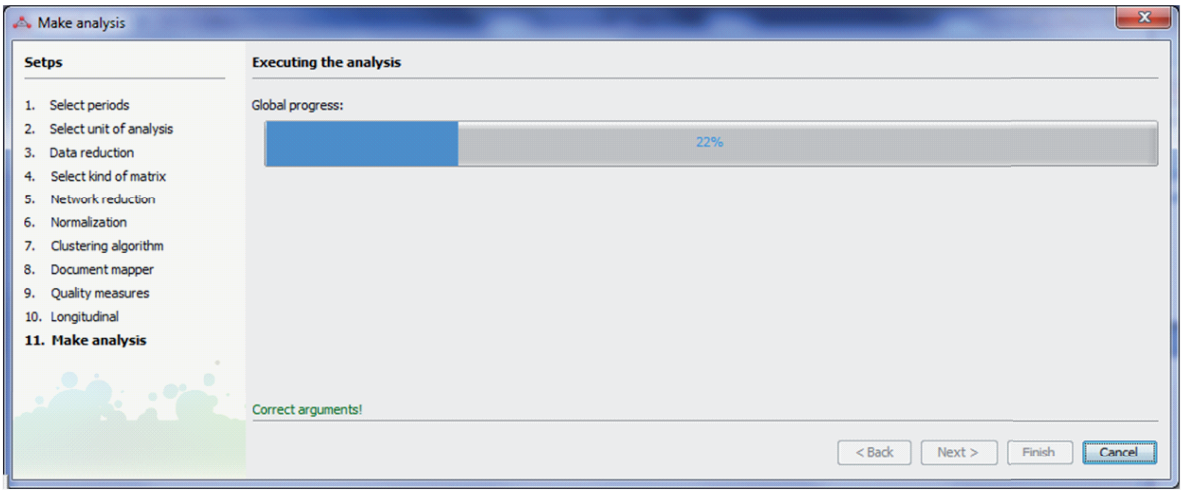


Figura B.19: Make the analysis (Step 11).

At the end, the analysis has to be saved (a new save window will be open when the process end), and then the results are visualized in the visualization module.

B.3. Visualization module

Once the science mapping analysis has been done, the results are visualized through this module. Furthermore, we can load an analysis previously done (to do that it is not necessary that a knowledge base is loaded).

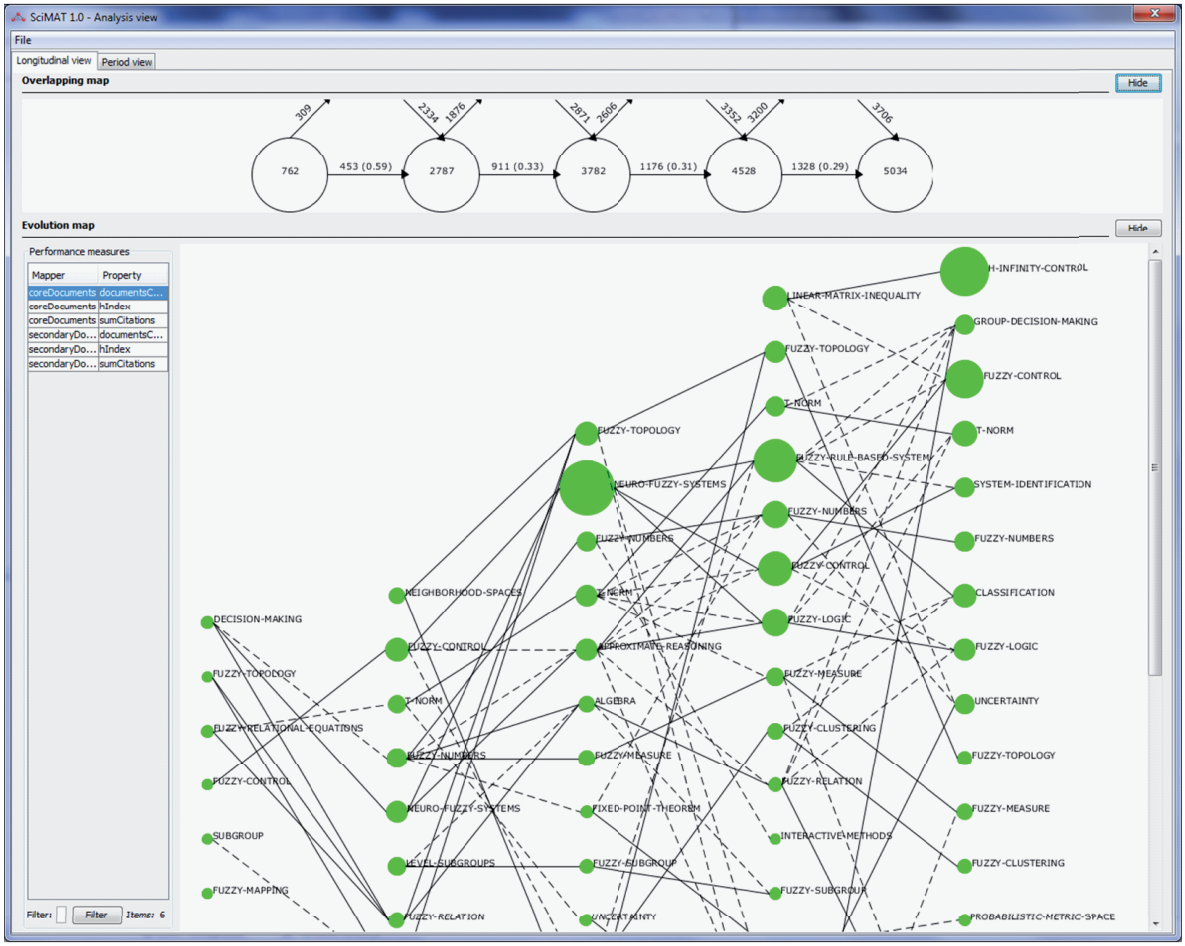


Figura B.20: Longitudinal view.

The visualization module has two views: longitudinal view and period view.

In the Longitudinal view (Figure B.20) the overlapping map (top) and evolution map (down) are shown. This view helps us to detect the evolution of the clusters throughout the different periods, and to study the transient and new items of each period and

the items shared by two consecutive periods. The right-table allows us to choose the measure used to draw the nodes in the evolution map.

Finally, the Period view (Figure B.21) shows detailed information for each period, its strategic diagram, and for each cluster, the bibliometric measures, the network and their nodes.

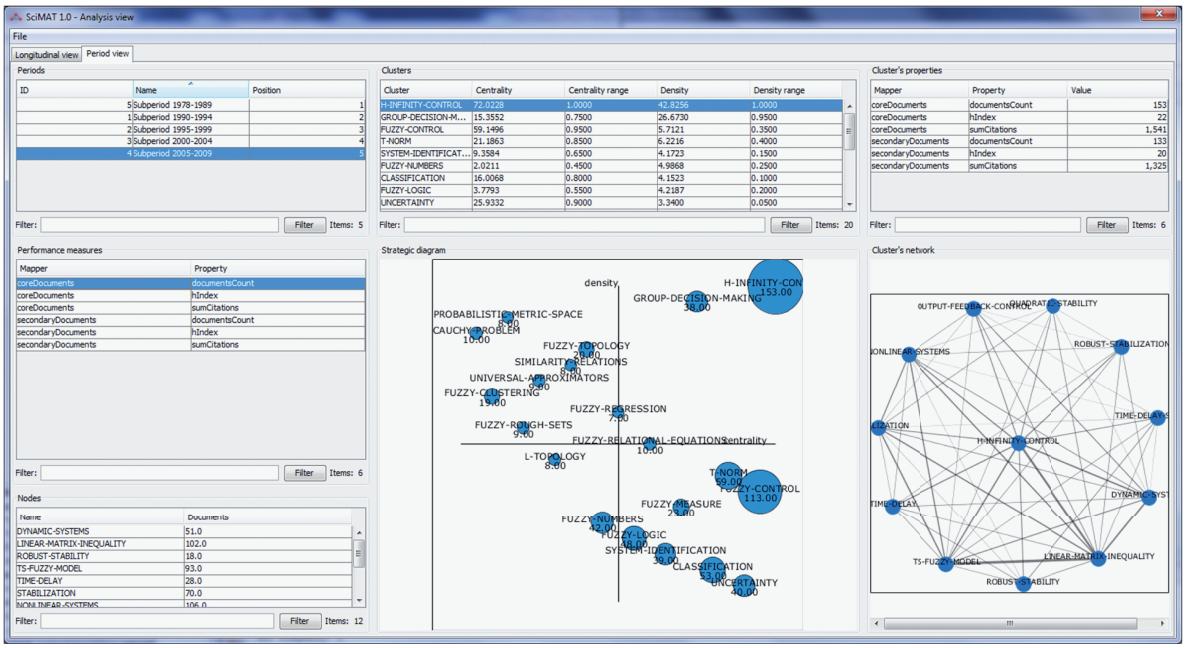


Figura B.21: Period view.

The visualization module is able to build a report in HTML or LaTeX format. The images (strategic diagrams, overlapping items map, etc.) are exported in PNG and SVG format, so the user can edit them. Furthermore, the cluster networks and evolution maps are exported in Pajek format.

Bibliografía

[1] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, and F. Herrera, Journal of Informetrics, 3:4 (2009), 273–289.

[2] ———, Scientometrics, 82:2 (2010), 391–400.

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, Addison-Wesley, 1999.

[4] R. Bailón-Moreno, *Ingeniería del conocimiento y vigilancia tecnológica aplicada a la investigación en el campo de los tensioactivos. desarrollo de un modelo ciencimétrico unificado*, Ph.D. thesis, Universidad de Granada, 2003.

[5] R. Bailón-Moreno, E. Jurado-Alameda, and R. Ruíz-Baños, Journal of the American Society for Information Science and Technology, 57:7 (2006), 949–960.

[6] R. Bailón-Moreno, E. Jurado-Alameda, R. Ruíz-Baños, and J. P. Courtial, Scientometrics, 63:2 (2005), 259–276.

[7] J. Bar-Ilan, Scientometrics, 82:3 (2010), 495–506.

[8] V. Batagelj and A. Mrvar, Connections, 21:2 (1998), 47–57.

[9] M. Batty, Environment and Planning A, 35:5 (2003), 761–765.

[10] M. Bordons and I. Gómez, *The web of knowledge: A festschrift in honor of eugene garfield*, vol. 57, ch. Collaboration network in science, pp. 688–690, Information Today, 2000.

[11] S. P. Borgatti, M. G. Everett, and L. C. Freeman, *Ucinet 6 for windows: Software for social network analysis, analytic technologies, Harvard, MA. http://www.analytictech.com*, 2002.

[12] K. Börner, C. Chen, and K.W. Boyack, Annual Review of Information Science and Technology, 37 (2003), 179–255.

[13] K. Börner, W. Huang, M. Linnemeier, R.J. Duhon, P. Phillips, N. Ma, A. Zoss, H. Guo, and M.A. Price, Scientometrics, 83:3 (2010), 863–876.

[14] K. W. Boyack, B .N. Wylie, and G. S. Davidson, Journal of the American Society for Information Science and Technology, 53:9 (2002), 764–774.

[15] R. R. Braam, H. F. Moed, and A. F. J. van Raan, Journal of the American Society for Information Science, 42:4 (1991), 252–266.

[16] S.C. Bradford, Engineering, 137:137 (1934), 85–86.

[17] F. J. Cabrerizo, S. Alonso, E. Herrera-Viedma, and F. Herrera, Journal of Informetrics, 4:1 (2010), 23–28.

[18] T. Cahlik, Scientometrics, 49:3 (2000), 373–387.

[19] M. Callon, J. P. Courtial, W. A. Turner, and S. Bauin, Social Science Information, 22:2 (1983), 191–235.

[20] M. Callon, J.P. Courtial, and F. Laville, Scientometrics, 22:1 (1991), 155–205.

[21] M. Callon, J.P. Courtial, and H. Penan, *El estudio cuantitativo de la actividad cientifica: de la bibliometria a la vigilancia tecnologica*, Ediciones TREA, 1995.

[22] P. J. Carrington, J. Scott, and S. Wasserman, *Models and methods in social network analysis*, Structural Analysis in the Social Sciences, Cambridge University Press, 2005.

[23] C. Chen, Proceeding of the National Academy of Science, 101:1 (2004), 5303–5310.

[24] ———, Journal of the American Society for Information Science and Technology, 57:3 (2006), 359–377.

[25] C. Chen, F. Ibekwe-SanJuan, and J. Hou, Journal of the American Society for Information Science and Technology, 61:7 (2010), 1386–1409.

[26] P. Chen and S. Redner, Journal of Informetrics, 4:3 (2010), 278–290.

[27] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, *A note on the its topic evolution in the period 2000-2009 at t-its*, IEEE Transactions on Intelligent Transportation Systems (in press).

[28] D. J. Cook and L. B. Holder, *Mining graph data*, Wiley-Interscience, 2006.

[29] N. Coulter, I. Monarch, and S. Konda, Journal of the American Society for Information Science, 49:13 (1998), 1206–1223.

[30] J. P. Courtial, Scientometrics, 19:1-2 (1990), 127–141.

[31] J. P. Courtial and B. Michelet, Scientometrics, 31:3 (1994), 251–260.

[32] G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N Wylie, Journal of Intelligent Information Systems, 11:3 (1998), 259–285.

[33] G. S. Davidson, B. N. Wylie, and K. W. Boyack, *Cluster stability and the use of noise in interpretation of clustering*, Proceedings of the IEEE Symposium on Information Visualization, 2001, pp. 23–30.

[34] L. Egghe, Scientometrics, 69:1 (2006), 131–152.

[35] S. I. Fabrikant, D.R. Montello, and D. M. Mark, Journal of the American Society for Information Science and Technology, 61:2 (2010), 253–270.

[36] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, FASEB Journal, 22:2 (2008), 338–342.

[37] W. Gänzel, Scientometrics, 51:1 (2001), 69–115.

[38] X. Gao and J. Guan, Scientometrics, 80:1 (2009), 283–302.

[39] E. Garfield, Current Contents: Social & Behavioural Sciences, 7:45 (1994), 5–10.

[40] A.-W Harzing, *The publish or perish book*, Tarma Software Research Pty Ltd, 2010.

[41] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, IEEE Transactions on Visualization and Computer Graphics, 8:1 (2002), 9–20.

[42] Q. He, Library Trends, 48:1 (1999), 133–159.

[43] M. Henderson, S. Shurville, and K. Fernstrom, Campus-Wide Information Systems, 26:3 (2009), 149–167.

[44] B.W. Herr, W. Huang, S. Penumarthy, and K. Börner, *Progress in convergence: Technologies for human wellbeing*, vol. 1093, ch. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence, pp. 161–179, Boston: Annals of the New York Academy of Sciences, 2007.

[45] E. Hetzler and A. Turner, IEEE Computer Graphic and Applications, 24:5 (2005), 22–26.

[46] J. Hirsch, Proceedings of the National Academy of Sciences, 102 (2005), 16569–16572.

[47] B. Jarneving, Scientometrics, 65:2 (2005), 245–263.

[48] V. Kandylas, S. P. Upham, and L. H. Ungar, ACM Transactions on Knowledge Discovery from Data, 4:2 (2010), art. no. 7.

[49] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, Neurocomputing, 21:1-3 (1998), 101–117.

[50] M. M. Kessler, American Documentation, 14:1 (1963), 10–25.

[51] J. Kleinberg, Data Mining and Knowledge Discovery, 7:4 (2003), 373–397.

[52] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela, IEEE Transactions on Neural Networks, 11:3 (2000), 574–585.

[53] J.A. Kreibich, *Using sqlite*, O'Reilly, 2010.

[54] I.V. Krsul, *Software vulnerability analysis.*, Ph.D. thesis, Purdue University, 1998.

[55] J. B. Kruskal, *Statistical methods for digital computers*, ch. Multidimensional scaling and other methods for discovering structure, pp. 19–50, Wiley, 1977.

[56] L. Leydesdorff and O. Persson, Journal of the American Society for Information Science and Technology, 61:8 (2010), 1622–1634.

[57] L. Leydesdorff and I. Rafols, Journal of the American Society for Information Science and Technology, 60:2 (2009), 348–362.

[58] L. Leydesdorff and T. Schank, Journal of the American Society for Information Science and Technology, 59:11 (2008), 1810–1818.

[59] A. G. López-Herrera, M. J. Cobo, E. Herrera-Viedma, and F. Herrera, Internacional Journal of Hybrid Intelligent Systems, 17:7 (2010), 17–32.

[60] A. G. López-Herrera, M. J. Cobo, E. Herrera-Viedma, F. Herrera, R. Bailón-Moreno, and E. Jimenez-Contreras, Information Research, 14:4 (2009), paper 421.

[61] J.M. López-Piñero, *El análisis estadístico y sociométrico de la literatura científica*, Valencia : Centro de Documentación e Informática Médica, 1972.

[62] A.J. Lotka, Journal of the Washington Academy of Sciences, 16:12 (1926), 317–323.

[63] K.W. McCain, Journal of the American Society for Information Science, 42:4 (1991), 290–296.

[64] S. Mikki, Scientometrics, 82:2 (2010), 321–331.

[65] H.F. Moed, R.E. De Bruin, and T.N. Van Leeuwen, Scientometrics, 33:3 (1995), 381–422.

[66] C. M. Morel, S.J. Serruya, G.O. Penna, and R. Guimarães, PLoS Neglected Tropical Diseases, 3:8 (2009), art. no. e501.

[67] S.A. Morris and B. Van Der Veer Martens, Annual Review of Information Science and Technology, 42:1 (2008), 213–295.

[68] F. Moya-Anegón, B. Vargas-Quesada, Z. Chinchilla-Rodríguez, E. Corera-Álvarez, V. Herrero-Solana, and F.J. Munoz-Fernández, Information Processing and Management, 41:6 (2005), 1520–1533.

[69] F. Moya-Anegón, B. Vargas-Quesada, Z. Chinchilla-Rodríguez, E. Corera-Álvarez, F.J. Munoz-Fernández, and V. Herrero-Solana, Journal of the American Society for Information Science and Technology, 58:14 (2007), 2167–2179.

[70] E Noma, Journal of the American Society for Information Science, 35:1 (1984), 29–33.

[71] E. C. M. Noyons, H. F. Moed, and M. Luwel, Journal of the American Society for Information Science, 50:2 (1999), 115–131.

[72] E. C. M. Noyons, H. F. Moed, and A. F. J. van Raan, Scientometrics, 46:3 (1999), 591–604.

[73] O. Persson, R. Danell, and J. Wiborg Schneider, *Celebrating scholarly communication studies: A festschrift for olle persson at his 60th birthday*, vol. 5, ch. How to use Bibexcel for various types of bibliometric analysis, pp. 9–24, International Society for Scientometrics and Informetrics, 2009.

[74] H. P. F. Peters and A. F. J. van Raan, Scientometrics, 20:1 (1991), 235–255.

[75] ———, Research Policy, 22:1 (1993), 23–45.

[76] X. Polanco, C. François, and J. C. Lamirel, Scientometrics, 51:1 (2001), 267–292.

[77] A. L. Porter and S. W. Cunningham, *Tech mining: Exploiting new technologies for competitive advantage*, John Wiley & Sons, Inc., 2004.

[78] A. L. Porter and J. Youtie, Journal of Nanoparticle Research, 11:5 (2009), 1023–1041.

[79] ———, Nature Nanotechnology, 4 (2009), 534–536.

[80] D. Price, *Little science, big science*, vol. 15, Columbia University Press, 1963.

[81] D. Price and S. Gürsey, Ci. Informatics Rio de Janeiro, 4:1 (1975), 27–40.

[82] A. Quirin, O. Cordón, J. Santamaría, B. Vargas-Quesada, and F. Moya-Anegón, Information Processing and Management, 44:4 (2008), 1611–1623.

[83] A.R. Ramos-Rodríguez and J. Ruíz-Navarro, Strategic Management Journal, 25:10 (2004), 981–1004.

[84] A. Rip and J.P. Courtial, Scientometrics, 6:6 (1984), 381–400.

[85] M. Rosvall and C. T. Bergstrom, PLoS ONE, 5:1 (2010), e8694.

[86] G. Salton and M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, 1983.

[87] R. W. Schvaneveldt, F. T. Durso, and D. W. Dearholt, *Network structures in proximity data*, Psychology of Learning and Motivation, vol. 24, Academic Press, 1989, pp. 249 – 284.

[88] Sci^2 Team, *Science of Science (Sci^2) Tool. Indiana University and SciTech Strategies, http://sci.slis.indiana.edu*, 2009.

[89] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, Genome Research, 13:11 (2003), 2498–2504.

[90] P. Shapira, J. Youtie, and A. L. Porter, Scientometrics, 85:2 (2010), 595–611.

[91] D. Skillicorn, *Understanding complex datasets: Data mining with matrix decompositions*, Data Mining and Knowledge Discovery Series, Chapman & Hall, 2007.

[92] A. Skupin, Journal of Informetrics, 3:3 (2009), 233–245.

[93] H. Small, Journal of the American Society for Information Science, 24:4 (1973), 265–269.

[94] ———, Social Studies of Science, 7 (1977), 139–166.

[95] ———, Scientometrics, 38:2 (1997), 275–293.

[96] ———, Journal of the American Society for Information Science and Technology, 54:5 (2003), 394–399.

[97] ———, Scientometrics, 68:3 (2006), 595–610.

[98] H. Small and E. Garfield, Journal of Information Science, 11:4 (1985), 147–159.

[99] H. Small and M. E. D. Koenig, Information Processing and Management, 13:5 (1977), 277–288.

[100] H. Small and E. Sweeney, Scientometrics, 7:3 (1985), 391–409.

[101] H. Small and S. P. Upham, Scientometrics, 79:2 (2009), 365–375.

[102] L.L. Thurstone, Psychological Review, 38 (1931), 406–427.

[103] D. Torres-Salinas, *Diseño de un sistema de información y evaluación científica. análisis ciencimétrico de la actividad investigadora de la universidad de navarra en el área de ciencias de la salud: 1999-2005*, Ph.D. thesis, Universidad de Granada, 2007.

[104] S. P. Upham and H. Small, Scientometrics, 83:1 (2010), 15–38.

[105] N. J. van Eck and L. Waltman, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 15:5 (2007), 625–645.

[106] ———, Journal of the American Society for Information Science and Technology, 60:8 (2009), 1635–1651.

[107] ———, Scientometrics, 84:2 (2010), 523–538.

[108] N. J. van Eck, L. Waltman, R. Dekker, and J. van den Berg, CoRR, abs/1003.2551 (2010).

[109] A. F. J. van Raan, *Handbook of quantitative science and technology research*, ch. Measuring Science, pp. 19–50, Springer Netherlands, 2005.

[110] ———, Measurement, 3:1 (2005), 1–19.

[111] L. Waltman, N. J. van Eck, and E. C .M. Noyons, Journal of Informetrics, 4:4 (2010), 629–635.

[112] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Cambridge University Press, 1994.

[113] H. D. White and B. C. Griffith, Journal of the American Society for Information Science, 32 (1981), 163–172.

[114] J. A. Wise, Journal of the American Society for Information Science, 50:13 (1999), 1224–1233.

[115] L.A. Zadeh, Information and Control, 8:3 (1965), 338–353.

[116] ———, Information Sciences, 178:13 (2008), 2751–2779.

[117] D. Zhao and A. Strotmann, Journal of the American Society for Information Science and Technology, 59:13 (2008), 2070–2086.

[118] G.K. Zipf, *Human behaviour and the principle of least effort*, Addison-Wesley, 1949.

[119] M. Zitt, E. Bassecoulard, and Y. Okubo, Scientometrics, 47:3 (2000), 627–657.