DEPARTAMENTO DE INFORMACIÓN Y COMUNICACIÓN
FACULTAD DE COMUNICACIÓN Y DOCUMENTACIÓN
GRANADA, NOVIEMBRE DE 2014

UNA APROXIMACIÓN MULTIMETODOLÓGICA PARA LA CLASIFICACIÓN DE LAS REVISTAS DE *SCIMAGO JOURNAL & COUNTRY RANK (SJR)*

TESIS DOCTORAL

DOCTORANDO: **ANTONIO JESÚS GÓMEZ NÚÑEZ**
DIRECTORES: **FÉLIX DE MOYA ANEGÓN**
BENJAMÍN VARGAS-QUESADA

Editor: Universidad de Granada. Tesis Doctorales
Autor: Antonio Jesús Gómez Núñez
ISBN: 978-84-9125-167-5
URI: http://hdl.handle.net/10481/39955

UNA APROXIMACIÓN MULTIMETODOLÓGICA PARA LA CLASIFICACIÓN DE LAS REVISTAS DE SCIMAGO JOURNAL & COUNTRY RANK (SJR)

Memoria que presenta

Antonio Jesús Gómez Núñez

Para optar al grado de Doctor

Dirigida por:

Dr. Félix de Moya Anegón

Dr. Benjamín Vargas-Quesada

Granada, Noviembre de 2014

A mis padres, mi verdadera fuente de conocimiento

"There have been many authorities who have asserted that the basis of science lies in counting or measuring, i.e. in the use of mathematics. Neither counting nor measuring can however be the most fundamental processes in our study of the material universe—before you can do either to any purpose you must first select what you propose to count or measure, which presupposes a classification."

"Existe una gran cantidad de autoridades que han afirmado que la base de la ciencia reside en el conteo o medición, es decir, en el uso de las matemáticas. Sin embargo, ni el conteo ni la medición pueden ser los procesos más básicos del estudio del universo material—antes de poder llevarlos a cabo para cualquier fin, primero debemos seleccionar lo que nos proponemos contar o medir, lo cual presupone una clasificación."

Roy A. Crowson

ÍNDICE DE CONTENIDOS

PARTE III: ABSTRACT

I. Prefacio

Esta tesis doctoral se ha llevado a cabo gracias a la financiación otorgada por el Consejo Superior de Investigaciones Científicas (CSIC) a través del programa denominado *Junta para la Ampliación de Estudios*, destinado a la formación de personal investigador y técnico. Dentro de este programa, el subprograma *JAEPredoctoral* ofrece becas de cuatro años de duración para el desarrollo de tesis doctorales en líneas estratégicas de los centros que componen el CSIC.

La concesión de una de las becas *JAEPredoc* ha hecho posible la realización de esta tesis doctoral en la Unidad Asociada Grupo SCImago CSIC-UGR. Sus resultados, conclusiones y hallazgos más destacados y significativos se presentan detalladamente en los cuatro trabajos de investigación elaborados por el doctorando en colaboración con sus codirectores de tesis, así como con otros miembros del grupo SCImago o reconocidos investigadores de su área.

II. Agradecimientos

Durante el tiempo de elaboración de esta tesis doctoral, muchas han sido las personas que me han influido positivamente en las distintas facetas necesarias para su exitosa realización y finalización. En primer lugar, me gustaría expresar mi más profundo agradecimiento al Dr. Félix de Moya Anegón y al Dr. Benjamín Vargas Quesada por su confianza en mi trabajo y por darme la oportunidad de poder llevarlo a cabo. Sin sus acertadas recomendaciones y consejos, este trabajo no hubiera sido posible.

Como si de un viaje se tratara, numerosas han sido las personas que me han acompañado en su realización, bien en parte o en la totalidad del trayecto. Especialmente, vaya mi recuerdo para mi tía Encarnación Núñez Ramírez, la que fue mi segunda madre y la que siempre ha estado y sigue estando junto a mí apoyándome. También, quiero dar las gracias a María Montserrat Posse Moure, por su ayuda con las interminables traducciones y correcciones del inglés, pero más aún, por su paciencia infinita y por aguantar, probablemente, los momentos más duros, haciéndome el camino mucho más llevadero.

Igualmente, son muchas otras las personas que, de una u otra forma, han estado a mi lado y a los que quiero agradecer sobremanera su ayuda y ánimo incondicional. Numerosos han sido los amigos de Granada, Osuna, Madrid y de otros lugares que me han ofrecido su apoyo, mostrándose siempre dispuestos a escuchar y a aguantar estoicamente mi casi único y exclusivo tema de conversación durante los últimos años. Para no olvidar a nadie, evitaré nombrarlos individualmente, en mi convencimiento de

que ellos mismos se sentirán partícipes y sabrán reconocer su inestimable papel en el desarrollo de este trabajo.

Quiero dar las gracias también a mis compañeros de profesión, a muchos de los cuales también tengo la suerte de considerarlos mis amigos, por poner a mi disposición su conocimiento, experiencia y tiempo de forma desinteresada. De entre todos ellos, me gustaría hacer mención especial a todo el personal de la biblioteca de la Facultad de Comunicación y Documentación de la Universidad de Granada, así como a mis compañeros del grupo de investigación SCImago, especialmente a Antonio González Molina, Diego Guzmán Morales y, por supuesto, a Zaida Chinchilla Rodríguez.

Por último, me gustaría transmitir mi gratitud a toda mi *familia* en general, dejando para el final el momento en el que mi agradecimiento y reconocimiento se vuelven eternos, como no, hacia la figura de mis padres, Antonio Gómez Vega y Josefa Núñez Ramírez, a quienes les debo todo lo bueno que pueda ser o tener y de los que he aprendido las lecciones más importantes y más enriquecedoras de mi vida. GRACIAS a ellos por enseñarme el significado de valores como la humildad, la perseverancia, el esfuerzo, la lucha, la amistad, o el apoyo incondicional, valores importantes y necesarios no sólo para el desempeño de este trabajo, sino también para afrontar los diferentes retos que acontecen a lo largo de la vida.

PARTE I: ASPECTOS GENERALES DEL TRABAJO

1. Introducción

En las sociedades de las economías avanzadas, el desarrollo científico-tecnológico está considerado un símbolo inequívoco de avance, progreso y bienestar económico y social tanto para los gobiernos como para sus ciudadanos. La interacción entre ciencia, tecnología y sociedad (CTS) ha sido ampliamente analizada por la denominada corriente de estudios CTS, que expone el incuestionable influjo de la investigación científica y la innovación tecnológica en la sociedad. Ésta, a su vez, se presenta como un elemento mediador con una gran capacidad para influir en el diseño y la configuración de las políticas científico-tecnológicas.

López Cerezo y Luján López (1997) describen el contexto del actual desarrollo científico-tecnológico haciéndose eco de una concienciación social colectiva con respecto a los peligros y potenciales impactos derivados de un mal uso de la ciencia y la tecnología. Por su parte, Moreno Rodríguez (1997) expone que la sociedad está estrechamente ligada a la idea de progreso, y que en su interrelación con éste, es capaz de transformarlo para su propio beneficio. Contrariamente, la ciencia y la tecnología también son capaces de ejercer su influencia sobre la sociedad. Así lo manifiesta Cañedo Andalia (2001) al afirmar que tanto la ciencia como la tecnología "constituyen un poderoso pilar del desarrollo cultural, social, económico y, en general, de la vida en la sociedad moderna", así como una "fuerza productiva inmediata" que influye sobre todas las esferas de la actividad humana.

La generación de conocimiento científico-tecnológico, que resulta esencial para el avance, el desarrollo y el bienestar económico y social, se fundamenta en un proceso

previo de investigación por el que se adquiere, (re)organiza y/o transforma el conocimiento existente en conocimiento nuevo, como norma general, de acuerdo a un método extendido, aceptado, sistemático y concluyente: el denominado método científico. Después, el ciclo de la investigación continúa con la difusión y presentación a la comunidad del nuevo conocimiento originado para su comprobación, legitimación y su potencial utilización en futuros experimentos e investigaciones.

Es habitual que tanto los principales resultados de investigación como el nuevo conocimiento científico-tecnológico generado se difundan a través de publicaciones científicas, especialmente, en forma de artículos o patentes, aunque esto depende en gran medida de factores como el área científica, el objeto o la fuente de financiación (pública o privada) de la investigación llevada a cabo. Si partimos de esta premisa, resulta evidente que la Bibliometría, que podríamos definir como una disciplina encargada de analizar y cuantificar la literatura y sus principales características mediante técnicas estadísticas, emerja entonces como una herramienta destacada en el ejercicio de la evaluación de la ciencia y la investigación, al diseñar indicadores a partir de la literatura científica compilada y tratada por los servicios de indización y resumen de las grandes bases de datos bibliográficas.

Entre estas bases de datos, destacan especialmente dos: Web of Science (WoS) (Thomson Reuters, 2009) y Scopus (Elsevier, 2004) que, actualmente, están reconocidas como los servicios de acceso e indización de información científica más prestigiosos y reconocidos a nivel internacional. Ambas bases de datos ofrecen las prestaciones básicas de cualquier otra base de datos de literatura científica, como por

ejemplo, búsqueda y recuperación de información, acceso a la información bibliográfica, descarga de contenidos, etc. Pero además, WoS y Scopus incluyen índices de citas construidos a partir de la inmensa red de citación establecida en base a las referencias bibliográficas citadas por los ítems recopilados, convirtiéndose así, en fuentes de información esenciales para el desarrollo de indicadores de impacto que resultan de gran utilidad para el ejercicio de la evaluación de la ciencia y la investigación o la toma de decisiones en política científica.

En los países desarrollados, gran parte de la investigación se articula en torno a la política científica diseñada por sus gobiernos, mientras que su posterior ejecución se canaliza a través de una serie organizaciones de diversa índole que son financiadas con fondos públicos. La financiación pública destinada a I+D+i en estos países establece el reparto de los fondos en función de variables como la relevancia temática, la producción, la calidad, la visibilidad o los impactos de la investigación de las instituciones, grupos e investigadores que integran el sistema. Weinberg (1962) ya trató este asunto afirmando que el gran crecimiento de la ciencia exigía cada vez más aportación de una sociedad cuyos recursos resultaban limitados. En su opinión, era necesario tomar decisiones que permitieran escoger entre los diversos campos científicos (*scientific choice*) y las diferentes instituciones (*institutional choice*) que reciben fondos del estado. Para ello sugirió una serie de criterios que consideró útiles para establecer prioridades en la selección, y que dividió en *criterios internos*, generados dentro del propio campo científico, y *externos*, más relevantes y originados fuera del campo, entre los que se incluirían el *mérito tecnológico*, *científico* y *social*.

Ante el cada vez más complejo panorama al que se enfrentan los profesionales encargados del diseño de la política científica y la evaluación de la investigación en los países, especialmente, en momentos de crisis económica como el actual, la toma de decisiones debe estar bien fundamentada y contrastada con datos y herramientas coherentes y consistentes. Los indicadores bibliométricos se presentan como instrumentos eficientes y efectivos en los procesos de evaluación de la ciencia y de la investigación desarrollada a nivel de investigadores, grupos, instituciones, áreas geográficas, países, disciplinas o áreas científicas, proporcionando valoraciones cuantitativas y cualitativas, realistas y fiables sobre el estado de la ciencia y la investigación analizada.

Entre el elenco de herramientas bibliométricas existentes con carácter internacional, SCImago Journal & Country Rank (SJR) destaca como un portal científico de acceso abierto desarrollado por el grupo SCImago (2007) para la generación de indicadores bibliométricos referentes a revistas o países sobre la base de la información incluida en la base de datos Scopus. Entre otras funciones, SJR permite elaborar rankings de países y revistas en función de diferentes indicadores de impacto (*indicador SJR*), producción y calidad, así como crear informes, mapas y gráficos comparativos. Por ello, este portal se ha constituido como una herramienta realmente interesante y relevante para el diseño y desarrollo de análisis bibliométricos y de dominio.

En sus orígenes, SJR adoptó el sistema de clasificación original de Scopus que, siguiendo el modelo utilizado en WoS, se compone de un esquema jerárquico orientado a la clasificación de revistas y dividido en dos niveles. El primer nivel o nivel

superior, estaría formado por las denominadas *áreas temáticas*, con una cobertura y un nivel de agregación de revistas más amplio. A continuación, en el segundo nivel, se situarían las *categorías temáticas*, que implican agrupaciones temáticas de revistas más pequeñas y, por lo tanto, con un alcance temático más específico y reducido.

Todas las revistas indizadas en WoS o Scopus son luego asignadas a una o varias categorías temáticas (y por extensión, a una o varias áreas) en función de diferentes criterios definidos por sus desarrolladores. Aunque la asignación de revistas y el diseño del sistema de clasificación de SJR fueron mejorados a posteriori mediante un análisis pormenorizado tanto del ámbito temático (*scope*) definido por las propias revistas como de sus patrones de citación, la necesidad de una nueva optimización en la clasificación de las revistas de SJR quedó patente por medio de un estudio detallado de las categorías que componen su esquema de clasificación (en el momento de realización de este trabajo, asciende a 308 categorías englobadas dentro de 27 áreas temáticas) y el desequilibrio presente en la distribución de las revistas a lo largo de las mismas, generado por la presencia de grandes concentraciones de revistas en ciertas categorías temáticas. Otros aspectos importantes a tener en cuenta fueron, por ejemplo, el desajuste entre las adscripciones temáticas definidas para las revistas de SJR y el criterio de los propios editores o las diferencias resultantes al comparar la adscripción de las revistas comunes incluidas en SJR y WoS, puesto que a pesar de tratarse de bases de datos con un sistema de clasificación y unas características similares, las diferencias tanto en las adscripciones como en la denominación y la composición de las áreas y categorías temáticas de ambas sistemas eran notorias.

La clasificación y la organización del conocimiento científico reflejado en las publicaciones recopiladas por las bases de datos, repositorios y otras fuentes de información resultan de vital importancia en la elaboración de indicadores y análisis bibliométricos. En opinión de Albert Roy Crowson, destacado biólogo británico del siglo XX centrado en el estudio de la taxonomía, la clasificación en sí misma es un proceso básico para cualquier propósito y, por supuesto, para la ciencia en general. En su obra *Classification and Biology* (1970), Crowson afirma que eran muchos los autores que consideraban que la base de la ciencia radicaba en contar o medir, es decir, en las matemáticas. Sin embargo, en el estudio objetivo y racional del universo material externo, que es como él define la ciencia, ni siquiera estos procedimientos matemáticos pueden considerarse tan básicos y esenciales como la propia clasificación que se presupone en el momento de seleccionar aquello que se pretende contar o medir.

Atendiendo a las ideas de Crowson, la significación de la clasificación parece ir más allá de una disciplina o ámbito científico concreto, más aún, cuando asegura que "clasificar cosas es, tal vez, la actividad más fundamental y característica de la mente humana, y subyace a todas las formas de ciencia". Otros autores y estudiosos anteriores ya habían resaltado la importancia de la clasificación para la ciencia. El filósofo George Henry Lewes (Lewes, 1893), por ejemplo, definió la ciencia como "la clasificación sistemática de la experiencia", mientras que el también filósofo Ernest Nagel (Nagel, 1961), afirmaba que "es el deseo por las explicaciones al mismo tiempo sistemáticas y controlables por la evidencia factual lo que genera la ciencia; y es la organización y la

clasificación del conocimiento sobre la base de principios explicativos lo que es el objetivo distintivo de la ciencia".

La clasificación, por lo tanto, se antoja esencial en cualquier procedimiento o actividad vinculada a la ciencia. Así sucede, por ejemplo, con la investigación, actividad inherente a la creación de nuevo conocimiento científico y, por extensión, con su principal producto derivado: la literatura científica. Desde el punto de vista de la Bibliometría, disciplina encargada del análisis cualitativo y cuantitativo de la literatura y sus características mediante técnicas estadísticas, parece obvio que la clasificación de la información derivada de la ciencia y la investigación y reflejada en las publicaciones debería resultar lo suficientemente precisa y apropiada como para facilitar su fin último: la confección de indicadores y herramientas (rankings, análisis de dominios, grafos…) de utilidad en la evaluación y en la toma de decisiones. La correcta clasificación de la literatura posibilitará la elaboración de indicadores consistentes y fiables. El resultado de su aplicación servirá para representar de forma concluyente y efectiva el estado de la cuestión analizado, evitando llegar a conclusiones erróneas y obtener interpretaciones incorrectas de la realidad.

A la vista del trascendental papel que la clasificación adquiere en el ámbito de la Bibliometría, este trabajo de investigación surge como una oportunidad para intentar mejorar la clasificación actual de la plataforma bibliométrica SJR. A través de distintas propuestas metodológicas, se recogen una serie de procedimientos semi-automáticos destinados, en primer lugar, a aumentar la precisión en la asignación temática de las revistas de SJR. Así, se persigue, por ejemplo, reducir el solapamiento de las categorías

temáticas derivado de la multi-asignación de revistas o las elevadas concentraciones de revistas que se producen en ciertas categorías temáticas. En segundo lugar, se pretende también optimizar y actualizar el esquema de clasificación temático de SJR mediante la generación de nuevas categorías, la eliminación de categorías obsoletas o difusas y la transformación o adaptación de otras existentes mediante fusión, desagregación, cambio de denominación, etc. Para ello, se han elaborado varios experimentos basados en diferentes técnicas estadísticas (*clustering*) y bibliométricas (análisis de referencias bibliográficas) que se presentan en forma de publicaciones científicas y que pueden ser utilizados tanto de forma individual como combinados. Su propósito más inmediato es la consecución de una mejora tanto cuantitativa como cualitativa en la clasificación de las revistas y en el diseño del esquema jerárquico de clasificación en uso, especialmente, a nivel de las categorías temáticas.

2. Delimitación del estudio

2.1. Terminología empleada

2.1.1. Definición de clasificación

Anteriormente, se ha puesto de manifiesto el papel fundamental que la clasificación desempeña dentro de las actividades y procesos relacionados con la ciencia y la investigación en general. Paralelamente, y circunscribiéndonos al ámbito temático de este trabajo hemos destacado también la importancia de la clasificación en el ámbito de la Bibliometría, donde la correcta clasificación de la literatura científica utilizada como base de sus experimentos se antoja esencial para la creación de indicadores y otros productos con un nivel de fiabilidad y solidez adecuados.

Sin embargo, no hemos definido todavía qué se entiende por clasificación y cuáles son sus características y peculiaridades más significativas. De forma genérica, Chan (1981) definió la clasificación como “el proceso de organizar el conocimiento de acuerdo a algún tipo de orden sistemático” para continuar añadiendo que “este proceso ha sido considerado la actividad más fundamental del pensamiento humano”, tal y como ya había afirmado Crowson (1970) con anterioridad.

Desde una perspectiva mucho más técnica, Moravcsik (1986) ofrece su particular visión asegurando que “clasificación significa ordenar cualquier cosa a través de un conjunto de categorías discretas”. Como consecuencia de la naturaleza multidimensional de los sistemas, dichas categorías deben ser luego etiquetadas mediante descriptores, al menos uno por cada dimensión. Por lo tanto, clasificar sería una forma de transformar variables continuas en discretas. Así, por cada dimensión de

un sistema, todo indicador continuo debe ser transformado en un conjunto de etiquetas discretas, tal y como se muestra en la figura 1.

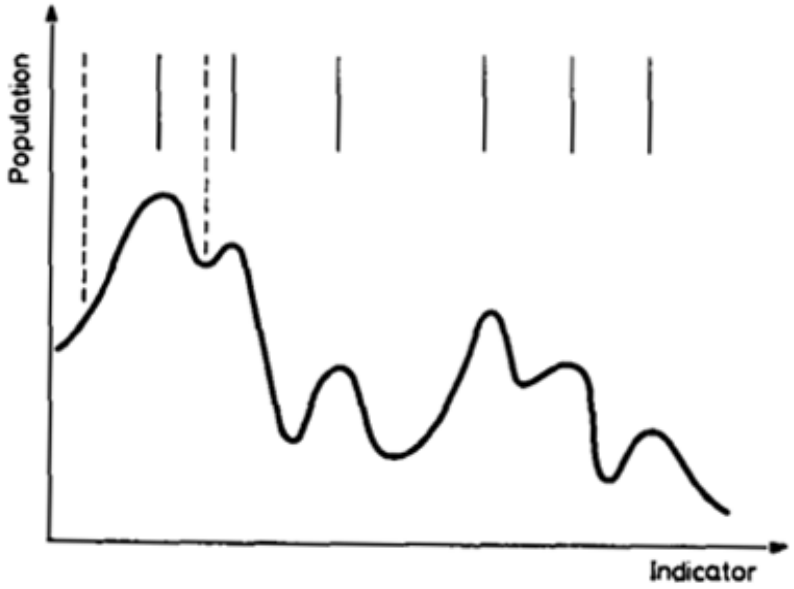


Figura 1: Representación de la "discretización" de un indicador continuo en una dimensión dada (Moravcsik, 1986).

De acuerdo con esta representación, en el eje de abscisas se recogen los distintos valores del indicador continuo, mientras que en el eje de ordenadas se representan las frecuencias de aparición de esos valores en el sistema, por ejemplo, el número de personas con un peso determinado en un país. Según Moravcsik, para que la clasificación resulte viable y sea efectiva, la distribución continua debe estar compuesta por un número moderado de valles y picos más o menos pronunciados. La decisión sobre el número final de categorías discretas que formarán parte de esa dimensión del sistema es totalmente arbitraria. Igualmente, la población contenida entre los altos de la distribución puede abordarse bien, definiendo de forma

conveniente límites arbitrarios o, por el contrario, estableciendo porcentajes para dividir la población comprendida entre dos categorías.

Para Jacob (2004), sin embargo, dentro de las Ciencias de la Información el término *clasificación* tiene tres acepciones distintas pero relacionadas, que son:

1. Un *sistema* de clases ordenadas de acuerdo a una serie de principios predeterminados y encaminado a organizar un conjunto de entidades.
2. Un *grupo* o *clase* dentro de un sistema de clasificación
3. El *proceso* de asignación de entidades a clases o grupos temáticos en un sistema de clasificación.

Como *sistema*, Jacob define la clasificación como un conjunto de clases mutuamente exclusivas y no solapadas ordenadas dentro de una estructura jerárquica que refleja un orden predeterminado de la realidad. Mientras, como *proceso*, la clasificación implicaría la asignación sistemática y ordenada de cada entidad en una única clase dentro de un sistema de clases mutuamente exclusivas y no solapadas.

Dentro del campo de la Organización del Conocimiento y las Ciencias de la Información Slavic (2007) utiliza el término clasificación para denotar esquemas de conceptos lógicamente organizados así como jerárquica y semánticamente estructurados que se crean con propósitos de indexación de contenidos e intermediación del conocimiento. Según su criterio, analizando un esquema de clasificación bien estructurado cualquiera puede percibir las relaciones semánticas entre los conceptos y encontrar conceptos o clases más genéricas, específicas o relacionadas. A partir de estas relaciones, por lo

tanto, sería posible establecer diferentes niveles de agregación para las publicaciones o documentos integrados en un sistema de información permitiendo, por ejemplo, delimitar sub-disciplinas, disciplinas, o grandes áreas temáticas.

La definición de clasificación de Hjorland, al que puede considerarse como uno de los grandes teóricos e investigadores de las Ciencias de la Información en la actualidad, no difiere demasiado con la definición proporcionada por Jacob anteriormente. En su opinión, la clasificación se puede definir como "la ordenación de objetos, procesos, ideas o cualquier otra cosa (incluidos los documentos) en clases a partir de algunas propiedades o características de esos objetos" (Hjorland, 2008a). Los trabajos e investigaciones desarrollados por estos dos autores han resultado también fundamentales para afrontar el siguiente apartado de esta tesis, donde se reflejan los diversos problemas relacionados con el uso de la terminología específica de la clasificación dentro de la literatura científica.

2.1.2. Problemas terminológicos

Es evidente que este trabajo de investigación tiene un eminente carácter aplicado y que, por lo tanto, su objetivo principal no reside en discernir o profundizar en aspectos teóricos y terminológicos referentes a la clasificación. Ahora bien, las dificultades y problemas acerca del uso de la terminología sobre clasificación y otros términos adyacentes a lo largo de la literatura científica nos han llevado a dedicar unas pocas líneas a este asunto en un intento por despejar las posibles dudas que pudieran generarse durante su lectura. Nuestro propósito no es otro que tratar de delimitar, argumentar y justificar el uso que se ha hecho de algunos de estos términos claves en

nuestra investigación, proporcionando indicaciones y definiciones precisas que faciliten su comprensión y sirvan para concienciar a los potenciales lectores sobre los problemas que pueden derivarse por la falta de una normalización terminológica.

La falta de normalización aumenta la importancia del proceso de selección de términos durante la creación de estrategias para la búsqueda y recuperación de información. Dicha selección debe estar orientada a la elección de términos apropiados y relevantes que representen de forma efectiva las necesidades de información de los usuarios y que permita encontrar un cierto equilibrio entre exhaustividad y precisión. De esa forma, puede reducirse la aparición de fenómenos como el ruido o el silencio documental, al tiempo que se minimiza la introducción de sesgos y tendencias o una posible distorsión en los resultados de la investigación.

La revisión de la literatura efectuada en determinadas fases de este trabajo sirvió para revelarnos un uso incontrolado, indistinto y poco ortodoxo de términos significativos relacionados con la clasificación y la gestión del conocimiento, como por ejemplo, área, categoría, disciplina, campo, especialidad, ámbito, tema, materia, etc., en numerosos trabajos de disciplinas como Bibliometría, Estadística o Ciencias de la Computación. En muchas ocasiones, estos términos se introducen como sinónimos aun no siéndolos en realidad. Otras veces, los términos se presentan, simplemente, sin una delimitación y una definición previa que permita identificar y diferenciar los conceptos o acepciones a los que hacen referencia.

Para solucionar muchos de estos problemas es recomendable tener en cuenta los trabajos teóricos desarrollados por autores como Hjorland (1992, 2001, 2011) donde numerosos conceptos, teorías y paradigmas relacionados con las Ciencias de la Información, la clasificación y la gestión del conocimiento son analizados en profundidad. Hjorland no sólo se ha preocupado por definir y delimitar conceptos básicos de las Ciencias de la Información, sino que ha extendido también el objeto de sus investigación al desarrollo y exposición de los paradigmas, modelos teóricos y métodos apropiados para el diseño e implementación de sistemas de clasificación y organización del conocimiento (Hjorland & Pedersen, 2005; Hjorland, 1998, 2003, 2008b, 2012). Como veremos más adelante en el apartado de Antecedentes, el término *organización del conocimiento* aparece estrechamente ligado al de clasificación, pero en opinión de Hodge (2000), se trata de un término mucho más amplio que abarcaría todos los sistemas de organización de información y gestión del conocimiento, incluyendo los sistemas de clasificación y categorización.

El uso de los términos *clasificación* y *categorización* es probablemente el mayor problema terminológico con el que nos hemos encontrado durante la realización de este trabajo y en él hemos centrado nuestro estudio. Dentro del amplio y variado corpus de literatura científica en el que se tratan estos conceptos es fácil encontrar ejemplos sobre el uso impreciso y laxo de los términos clasificación y categorización (Olivera Betrán, 2011; Qu, Cong, Li, Sun, & Chen, 2012; Rak, Kurgan, & Reformat, 2007; Roitblat, Kershaw, & Oot, 2010; Torres-Salinas et al., 2010, etc.), ignorando determinados matices y peculiaridades que los caracterizan y los distinguen. Entre las posibles causas que podrían explicar este fenómeno destaca la diversidad y la

heterogeneidad en la autoría de estos trabajos. Así, muchos de sus autores cuentan con una formación diversa y proceden de disciplinas con una base teórica variada y dispar como, por ejemplo, Bibliometría, Ciencias de la Computación, Medicina, Filosofía, Estadística, etc. Otra posible razón puede tener que ver con la traducción y la correspondencia incorrecta de los términos traducidos de trabajos científicos en otros idiomas, principalmente, del inglés.

Jacob (1991) afirma que a pesar de representar conceptos distintos, categorización y clasificación son términos generalmente empleados como sinónimos en la literatura sobre categorización. En una de sus investigaciones (Jacob, 2004) desarrolla un análisis profundo y detallado de ambos términos con la intención de mostrar sus diferencias y particularidades y facilitar su comprensión. Como parte de dicha investigación, desarrolla la siguiente tabla, donde se definen de forma resumida los principales rasgos caracterizadores y distintivos de los términos categorización y clasificación.

CATEGORIZATION		**CLASSIFICATION**
	Process	
Creative synthesis of entities based on context or perceived similarity		Systematic arrangement of entities based on analysis of necessary and sufficient characteristics
	Boundaries	
Because membership in any group is non-binding, boundaries are “fuzzy”		Because classes are mutually-exclusive and non-overlapping, boundaries are fixed

	Membership	
Flexible: category membership is based on generalized knowledge and/or immediate context		Rigorous: an entity either is or is not a member of a particular class based on the intension of a class
	Criteria for Assignment	
Criteria both context-dependent and context-independent		Criteria are predetermined guidelines or principles
	Typicality	
Individual members can be Rank-ordered by typicality (graded structure)		All members are equally representative (ungraded structure)
	Structure	
Clusters of entities; may form hierarchical structure		Hierarchical structure of fixed classes

Tabla1: Comparison of Categorization and Classification (Jacob, 2004)

Analizando los trabajos de Jacob llegamos a dos conclusiones básicas:

1. De acuerdo a sus ideas, los diferentes procesos desarrollados en este trabajo parecen estar orientados más hacia la categorización que a la clasificación propiamente dicha de revistas de SJR. Así, atendiendo a sus principios, se ha desarrollado un proceso sistemático de ordenación y organización en el que se parte del análisis de las similaridades entre las revistas en base a sus patrones de citación y/o sus derivados para, posteriormente, generar diferentes grupos o clústeres equiparables a categorías temáticas que pueden organizarse mediante una estructura jerárquica en dos niveles: (1) áreas y (2) categorías. La pertenencia a

cada grupo o clúster es flexible de acuerdo con el conocimiento general del clasificador y a su contexto y, finalmente, los límites entre las categorías son difusos puesto que la adscripción de una revista en una categoría no es exclusiva, es decir, las revistas pueden ser multi-asignadas, dando lugar a cierto grado de solapamiento. No obstante, otras características y especificaciones propias de la clasificación, como el uso de un esquema jerárquico previamente determinado, así como el mismo nivel de pertenencia o representatividad de las revistas incluidas en las diversas categorías, aparecen también a lo largo de los procedimientos ejecutados en nuestras diferentes propuestas.

2. Si entendemos que en este trabajo de investigación desarrollamos propuestas para la clasificación de revistas de SJR en el campo de la Bibliometría (concebida ésta como una sub-disciplina de las Ciencias de la Información[1,2]) dichas propuestas incluirían la clasificación desde la triple perspectiva descrita por Jacob:
   a. Así, como *proceso*, los distintos procedimientos metodológicos ejecutados permitieron llevar a cabo una reorganización o reordenación de las revistas de SJR por medio de la generación de nuevas agrupaciones que se establecieron en función de las relaciones (citas) existentes entre las revistas, es decir, en base a sus patrones de citación. Nuestro proceso de clasificación implica también la delimitación de grupos (categorías temáticas) y la adscripción de revistas de forma lógica y sistemática utilizando diversas técnicas bibliométricas

[1] Pérez Matos, N.E. (2002). La bibliografía, bibliometría y las ciencias afines. *ACIMED* v.10, n.3. Disponible en: http://eprints.rclis.org/5141/1/bibliografia.pdf [Fecha de consulta: 02-09-2014].
[2] Araujo Ruiz, J.A & Arencibia Jorge, R. (2002). Informetría, bibliometría y cienciometría: aspectos teórico-prácticos. *ACIMED* v.4. Disponible en: http://eprints.rclis.org/5000/1/aci040402.pdf [Fecha de consulta: 02-09-2014].

y estadísticas, así como la creación, modificación, actualización e incluso eliminación de algunos de estos grupos o categorías temáticas.

b. Cada uno de los *grupos* generados como resultado de los procesos de clasificación ejecutados junto con las revistas que los integran fueron analizados. En los experimentos de *clustering* fue necesario tener en cuenta determinadas propiedades de las revistas (citación, texto, o ambos) para asignar una etiqueta o nombre a cada uno de estos grupos y, por consiguiente, adscribir sus revistas a la categoría temática pertinente. Estas categorías temáticas podían ser reutilizadas del sistema de clasificación original de SJR, o bien ser de nueva creación. Los grupos de revistas generados fueron también validados mediante la comparación con otros sistemas de clasificación, por ejemplo, con la propia clasificación original de SJR, con el sistema de categorías temáticas de WoS, o ya través de otros métodos alternativos, como el uso de técnicas de visualización de información.

c. Finalmente, el *sistema* de clasificación de SJR se vio afectado por los diversos cambios introducidos en el conjunto de categorías temáticas incluidas en el mismo. Se analizaron estos cambios y su posible incidencia en la organización y en las relaciones jerárquicas establecidas entre áreas y categorías temáticas de SJR. Entre estos cambios pueden enumerarse permutaciones en las relaciones jerárquicas, aparición de poli-jerarquía y creación o desaparición de algunas relaciones entre los dos niveles jerárquicos implementados en el sistema.

Dentro del campo de la minería de textos y datos Meunier, Forest y Biskri (2005) también analizaron las similitudes y diferencias entre categorización y clasificación desde cuatro facetas distintas:

1. Como división (*partition*), determinan que clasificación y categorización son completamente conceptos idénticos.
2. Como etiqueta (*label*), la categoría no es más que el nombre asignado al conjunto de objetos clasificados (clase), el cual, permanecerá invariable con independencia del nombre o etiqueta finalmente asignado.
3. Como estado cognitivo (*cognitive state*), la categorización es más compleja que la clasificación al requerir algún tipo de operación cognitiva efectuada por algún agente como la memoria, la percepción, o la comparación.
4. Como "morfismo" (*morphism*), término que en el ámbito de las matemáticas hace referencia a una representación entre dos objetos en una categoría abstracta[3], consideran que las categorías implícitamente sitúan a las clases en alguna estructura o esquema jerárquico.

En su opinión, a pesar de que representan conceptos distintos, *clase* y *categoría* aparecen habitualmente fusionados y utilizados de forma similar en la literatura científica. Pero mientras el término clasificación (de texto) se refiere principalmente a la clasificación entendida como clúster o conjunto de entidades, es decir, como un proceso de agrupación o agregación de entidades conforme a ciertos criterios, la categorización (de texto) conlleva un plus o añadido. En concreto, la categorización, se fundamenta en un proceso de clasificación previo que, seguidamente, conlleva un

[3] Weisstein, E.W. (1995). *Wolfram MathWorld: the web's most extensive mathematics resource*. Disponible en: http://mathworld.wolfram.com/Morphism.html [Fecha de consulta: 02-09-2014].

proceso de etiquetado que consiste en atribuir nombres o etiquetas predefinidas. La elección de estas etiquetas es lo que requiere ese plus, que no es otro que la dimensión estructural y cognitiva necesarias en para categorizar.

Una vez analizada la problemática derivada del uso de los términos clasificación y categorización, tomamos la decisión de utilizar *clasificación* como término preferente en este trabajo, limitando en la medida de lo posible el uso del término categorización, al que consideramos sinónimo a pesar de sus connotaciones y rasgos distintivos. Las principales razones para llegar a esta determinación fueron:

- En primer lugar, el trabajo antes citado de Meunier, Forest y Biskri (2005) donde se afirma que, como *división* o *procedimiento* ambos términos son prácticamente idénticos: "For some, classification and categorization are identical concepts. There is no real difference between them. Here classification and categorization mean a type of procedure by which, on certain objects, entities, or elements is applied a regrouping, sorting or *clustering* process. The abstract entity produced by this process can be called a class, a kind, a category, a group, a set, a collection, an aggregate, a cluster, etc. For instance, in doing one's weekly market, one will regroup objects and form the class of "apples", as opposed to classes of "oranges" or "soft drinks". Por lo tanto, dentro de la faceta más utilizada a lo largo de este trabajo, es decir, como proceso para la generación de grupos de entidades y su asignación a clases o grupos temáticos, ambos términos pueden ser admitidos casi sin lugar a dudas.
- En segundo lugar, se ha tratado de mantener un criterio de homogeneidad en relación con la literatura científica de corte bibliométrico, donde el término

clasificación aparece normalmente en la mayoría de estudios encaminados a organizar la literatura científica de bases de datos científicas analizadas y utilizadas como referencia y fuente de datos de esos estudios, principalmente WoS o Scopus.

- Por otra parte, algunas técnicas de clasificación automáticas, como por ejemplo el análisis de *clustering*, permiten identificar grupos de documentos u objetos relacionados internamente (*intra-clúster*) y establecer además relaciones o asociaciones *inter-clúster*, por ejemplo mediante su cercanía o por medio de relaciones jerárquicas. No obstante, estas agregaciones tienen que ser posteriormente etiquetadas. En las diferentes propuestas desarrolladas en este trabajo, parte de ese etiquetado se ha realizado conforme a las etiquetas procedentes delas categorías temáticas del esquema de clasificación de SJR. La reutilización de estas etiquetas ha permitido clasificar cada una de las revistas mediante su asignación o adscripción a las categorías originales de dicho esquema.
- En relación con el punto anterior, hemos tratado de utilizar, como norma general, los términos *área* y *categoría* temática para referirnos a los dos posibles niveles de agregación de revistas presentes en las diferentes propuestas de clasificación desarrolladas en nuestra investigación. Esta decisión no responde a otro criterio que continuar con la coherencia, en este caso, con respecto al sistema de clasificación existente en SJR.

2.2. Fuentes y datos utilizados

En el año 2004 Elsevier[4] introdujo en el mercado de la información la mayor base de datos de resúmenes y citas de literatura científica y académica en la actualidad: Scopus. A día de hoy, esta base de datos cuenta con más de 20.000 títulos activos de revistas procedentes de más de 5.000 editores, lo que se traduce en una amplia cobertura temática en las áreas de Ciencia, Tecnología, Medicina, Ciencias Sociales y, desde 2009, también en Artes y Humanidades, gracias a la inclusión de las más de 2.700 revistas procedentes del European Reference Index for Humanities (ERIH) de la European Science Foundation llevada a cabo por Elsevier. Consecuentemente, Scopus se ha convertido en una potente herramienta caracterizada por un marcado carácter multidisciplinar e internacional, una fuerte orientación tecnológica y la expansión de ciertas áreas temáticas, como las Humanidades.

Pero además de una extensa cobertura de la literatura científica, Scopus incorpora información precisa sobre las relaciones que se establecen entre los documentos a través de la incorporación y el control de sus listas de referencias bibliográficas. Por todo ello, pronto se convirtió en una fuente de información alternativa para el desarrollo de estudios y análisis bibliométricos que, hasta su aparición, se habían fundamentado básicamente en la información de las bases de datos del Institute for Scientific Information (ISI) liderado por Eugene Garfield. Mediante un proceso de fusión empresarial estas bases de datos pasaron luego a ser gestionadas por la compañía Thomson Reuters y se integraron en el portal Web of Science (WoS), que actualmente indexa alrededor de unos 12.500 títulos de revistas.

[4] Elsevier (2014). *Elsevier*. Disponible en: http://www.elsevier.com/ [Fecha de consulta:02-09-2014]

WoS y Scopus presentan similitudes en lo referente al control de la información y a su orientación. A pesar de incluir grupos o paquetes de revistas exclusivos las dos bases de datos comparten, sin embargo, más de 11.000 títulos de revistas comunes[5]. Las diferencias esenciales tienen un componente temporal que afecta al control de las referencias citadas y de la filiación de los autores (Jacsó, 2005a, 2005b). Así, mientras WoS controla todas las referencias citadas en los documentos sin importar su año de publicación, Scopus únicamente recoge las referencias de los documentos publicados desde 1996 en adelante (Fingerman, 2006; Jacsó, 2009). Aun así, la red de citación de Scopus resulta más extensa que la de WoS como consecuencia del mayor número de documentos citables que indiza. Respecto a la filiación, en sus inicios, Scopus sólo incluía los datos de la institución del primer autor, siguiendo el procedimiento utilizado en otras bases de datos como MEDLINE/PubMed, pero a partir del año 2003[6] el registro se extendió al resto de autores de forma exhaustiva.

A las características específicas de la base de datos Scopus información hay que añadir las singularidades propias de SCImago Journal & Country Rank[7]. Esta herramienta de acceso abierto puede definirse como un sistema de información científica basado en los contenidos de Scopus entre 1996 y 2012, que facilita la generación de listados ordenados de revistas y países convirtiéndose en un recurso dirigido a la evaluación de

[5] El dato de los títulos comunes y exclusivos se ha obtenido de: Center for Research Libraries. (2012). *Academic Database Assessment Tool*. Disponible en: http://adat.crl.edu/databases [Fecha de consulta: 02-09-2014].

[6] Afirmación tomada de una presentación de: Horrocks, G. (2009). Battle of the giants: a comparison of Web of Science, Scopus… & Google Scholar. *ICIC 2006*. Disponible en: http://www.haxel.com/icic/archive/2006/programme/oct23#battle-of-the-giants-a-comparison-of-scopus-web-of-science-and-google-scholar [Fecha de consulta: 02-09-2014].

[7] SCImago Research Group (2007). *SCImago Journal and Country Rank (SJR)*. Disponible en: http://www.scimagojr.com/ [Fecha de consulta: 02-09-2014].

la ciencia a nivel mundial. La posibilidad de acceder gratuitamente a los indicadores de referencia tanto a nivel mundial como regional o nacional, la hacen óptima para su uso como referente en el contexto internacional. La información que se proporciona resulta similar a la ofrecida en los Essential Science Indicators de Thomson Reuters nacionales. La diferencia más importante tiene que ver con la inclusión de indicadores sobre producción primaria, auto-citación y h-index además de los ya tradicionales (número de documentos, número de citas y total de citas por documento). Los usuarios pueden acceder a SJR y replicar estos indicadores en cualquier momento, lo que facilita su comparación con una región o un conjunto de países en un período temporal definido.

En lo referente a la clasificación, la plataforma SJR ha adoptado el esquema de clasificación implementado en la base de datos Scopus. Dicho esquema responde a un modelo de clasificación con dos niveles jerárquicos que se corresponden con 27 áreas y 308 categorías temáticas procedentes del sistema de clasificación disciplinar de la ciencia, conocido como *All Science Journal Classification (ASJC)*[8]. No obstante, en comparación con Scopus la plataforma SJR presenta un valor añadido que tiene que ver con la mejora en la adscripción de las revistas a las categorías del sistema. Esta mejora hace referencia a potenciales nuevas asignaciones de las revistas a categorías temáticas distintas de las asignadas inicialmente por Scopus en base al criterio de los editores. Las propuestas son minuciosamente estudiadas por miembros del grupo SCImago y aceptadas en el caso de corroborarse mediante un análisis detallado del ámbito temático y los patrones de citación de las revistas. De esta forma, a la categoría

[8] Elsevier (2014). *Scopus Journal Title List*. Disponible en: http://www.elsevier.com/online-tools/scopus/content-overview [Fecha de consulta: 02-09-2014].

originalmente asignada en Scopus, se añaden las nuevas categorías consideradas. Por último, las propuestas de cambio aceptadas son comunicadas a Elsevier proporcionando así el *feedback* necesario para optimizar también la clasificación inicial de las revistas de Scopus con las nuevas adscripciones temáticas.

Si atendemos al objetivo principal perseguido en este trabajo, parece obvio que la fuente de información seleccionada no puede ser otra que el propio SJR. En el caso concreto de los diferentes experimentos elaborados durante nuestra investigación, casi un total de 19000 títulos de revistas activas de Scopus formaron parte de nuestro conjunto de datos. Más concretamente, dos conjuntos de datos han sido utilizados como base de dichos experimentos: (i) el primero de ellos recopila los datos de citación de un total de 17.158 revistas Scopus activas sobre las que se aplicó un procedimiento de análisis de las referencias citadas; (ii) el segundo conjunto, por su parte, es más amplio y se compone de los datos de citación de 18.891 revistas Scopus activas que se utilizaron para la ejecución de distintos procedimientos de *clustering*. Una descripción más detallada sobre ambos conjuntos de datos ha sido proporciona en la correspondiente sección de Materiales y Métodos de este trabajo.

2.3. Estructura del trabajo

Esta tesis doctoral consta de tres partes claramente diferenciadas:

- *Parte I:* Integra los contenidos comunes a todo el proceso de investigación reflejado en esta tesis doctoral. Tras una breve *introducción* para contextualizar la investigación y su significación, se procede a delimitar el tema de la investigación haciendo especial énfasis en aspectos claves como la terminología empleada y los

posibles problemas derivados de su uso, o la fuente de información escogida como base para la extracción de los datos necesarios para la investigación llevada a cabo. A continuación, se *justifica* la oportunidad y los diversos motivos que conducen a la realización de este trabajo junto con sus *objetivos* generales y específicos, así como su interrelación con las cuatro propuestas o experimentos de clasificación descritos más adelante. Seguidamente, se desarrollan los *antecedentes*, donde se representa el marco teórico a través de la revisión de trabajos de investigación relacionados con la temática que se aborda, es decir, la problemática de la clasificación (especialmente mediante técnicas automáticas) en el ámbito de la Bibliometría Una vez analizadas las principales *limitaciones* del trabajo, se continua con la presentación y la definición de los *materiales* o conjuntos de datos utilizados en los diferentes experimentos prácticos llevados a cabo junto con el diseño de las diferentes *metodologías* implementadas para la mejora de la clasificación de SJR. El análisis de los principales *resultados* obtenidos como consecuencia de la aplicación de nuestras propuestas metodológicas desemboca en la posterior *discusión*, generada en relación con los objetivos previamente marcados, y en la extracción de las *conclusiones* más destacadas. Finalmente, para superar las posibles limitaciones, inconvenientes o problemas encontrados a lo largo de nuestra investigación se esbozan una serie de *perspectivas futuras de investigación*.

- *Parte II:* Está compuesta por cuatro publicaciones científicas donde se recogen los diferentes procedimientos metodológicos desarrollados para conseguir los objetivos inicialmente planteados además de los principales resultados derivados de su aplicación y las conclusiones más relevantes extraídas. Las tres primeras corresponden a artículos en inglés publicados en revistas de impacto del área. La

cuarta publicación se trata de una contribución en español para un congreso científico de ámbito internacional.

- Artículo 1: *Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis*
- Artículo 2: *Optimizing SCImago Journal & Country Rank classification by community detection*
- Artículo 3: *Updating the SCImago Journal & Country Rank classification: a new approach using Ward's clustering and alternative combination of citation measures*
- Comunicación: *Visualización y análisis de la estructura de la base de datos Scopus*

- *Parte III:* En el último apartado de la tesis, se ha incluido un amplio y estructurado resumen en inglés con los aspectos más destacados del trabajo, incluyendo, una breve introducción, los objetivos fundamentales del trabajo, el material utilizado para el desarrollo de las propuestas metodológicas expuestas, los resultados más interesantes y, como no, la discusión y las principales conclusiones extraídas del análisis detallado de los resultados.

3. Justificación y Objetivos

3.1. Justificación

Al comienzo de este trabajo afirmábamos que, en la actualidad, la ciencia, la tecnología y la investigación son fenómenos asociados a la idea de avance, competitividad, desarrollo, poder, ventaja o bienestar social. La importancia de la ciencia, la tecnología y la investigación en el mundo actual se traduce también en un constante crecimiento de la inversión pública y privada, una incesante generación de productos científicos (congresos, patentes, publicaciones…) o en un paulatino aumento de los programas de formación en ciencia y tecnología en los países desarrollados y emergentes. Cifras detalladas sobre estos indicadores de desarrollo pueden encontrarse en las bases de datos de organismos como la OCDE[9], EUROSTAT[10], los institutos nacionales y regionales de estadísticas, las oficinas de patentes (European Patent Office[11], US Patent and Trademark Office[12], etc.) o las bases de datos comerciales de publicaciones científicas, en especial, WoS y Scopus.

En el ámbito de la clasificación y la organización del conocimiento, tanto el dinamismo como el crecimiento característicos de la ciencia y la investigación acarrean ciertas implicaciones a tener en cuenta en los procesos de diseño, elaboración e implementación de los sistemas de clasificación. Desde una perspectiva bibliométrica, estos aspectos resultan claves. Por una parte, es necesario crear sistemas de clasificación sólidos y estables que permitan organizar la literatura científica

[9] OECD (2014). *OECD Statistics*. Disponible en: http://stats.oecd.org/ [Fecha de consulta: 02-09-2014].
[10] Eurostat (2014). *Database - Eurostat*. Disponible en: http://ec.europa.eu/eurostat/data/database [Fecha de consulta: 02-09-2014].
[11] European Patent Office (2014). *EPO Searching for Patents*. Disponible en: http://www.epo.org/searching.html [Fecha de consulta: 02-09-2014].
[12] United States Patent and Trademark Office (2014). *Search for Patents*. Disponible en: http://www.uspto.gov/patents/process/search/ [Fecha de consulta: 02-09-2014].

almacenada en las bases de datos a través de un sistema de clasificación compuesto por agregaciones bien definidas de dicha literatura (como las áreas y categorías temáticas de SJR). Estas agregaciones han de presentar cierta correspondencia con las diferentes disciplinas o campos de la ciencia. Por otra parte, la correcta adscripción de la literatura a estos grupos resulta clave para su organización, facilitando los procesos de búsqueda y recuperación de información y, al mismo tiempo, la correcta delimitación de las diferentes disciplinas de la ciencia y la investigación a través de su propia producción científica. Estos procesos, por lo tanto, son esenciales para el desarrollo y producción de indicadores bibliométricos.

Narin (1976) había apreciado la existencia de notables diferencias entre los sistemas de clasificación utilizados en distintas bases de datos e intentó solucionar los inconvenientes ocasionados por la falta de un esquema de clasificación apropiado. Para alcanzar cierta homogeneidad, propone clasificar los artículos de las bases de datos asignándolos a las categorías temáticas de las revistas fuentes a las que pertenecen, pero siempre y cuando la cobertura (número de trabajos) de las revistas fuera suficientemente amplia.

Posteriormente, Gómez, Bordons, Fernández y Méndez (1996) llevaron a cabo un análisis profundo de las clasificaciones utilizadas en las bases de datos científicas, su problemática, y las diferentes propuestas desarrolladas hasta el momento en el ámbito de la Bibliometría. Uno de los principales problemas que encontraron durante dicho análisis fue la falta de normalización o, en otras palabras, la utilización de sistemas de clasificación adaptados a las características y particularidades de las diferentes bases

de datos científicas existentes. De entre la variedad de propuestas utilizadas para clasificar la literatura científica y delimitar las disciplinas de la ciencia los autores destacan tres enfoques habituales:

- La clasificación de los documentos mediante códigos o palabras claves, como en el caso de la base de datos MEDLINE/PubMed, cuyas publicaciones son representadas temáticamente por descriptores del tesauro Medical Subject Headings (MeSH)[13].
- La clasificación de revistas por medio de un sistema de categorías temáticas similar al de la base de datos WoS (y actualmente, también Scopus).
- La clasificación a partir de los datos de entrada o inputs (por ejemplo, las líneas de investigación o los objetivos socio-económicos) de los proyectos de investigación clasificados conforme al sistema de codificación de la UNESCO.

Entre sus principales conclusiones resaltan que es posible abordar eficientemente la delimitación temática en los estudios bibliométricos mediante el uso de *keywords* o códigos temáticos, revistas y publicaciones, filiación institucional, áreas disciplinares de centros e instituciones, formación profesional de los autores, etc., dejando a criterio del bibliómetra la selección de la forma más apropiada en función de las singularidades de cada estudio. Concluyen su trabajo, sin embargo, apuntando a la imperante necesidad de normalizar los procesos de clasificación de cara a mejorar la comparabilidad y la consistencia tanto de los indicadores como de los análisis bibliométricos en general.

[13] National Center for Biotechnology Information (2014). *Medical Subject Headings*. Disponible en: http://www.ncbi.nlm.nih.gov/mesh [Fecha de consulta: 02-09-2014].

La ausencia tanto de un modelo como de un esquema de clasificación estándar en el ámbito bibliométrico ha sido reseñada no sólo en los trabajos anteriormente citados, sino también por otros mucho más recientes (Archambault, Beauchesne, & Caruso, 2011a; Waltman & van Eck, 2012), lo que demuestra que este asunto continua siendo un problema de interés pendiente de ser resuelto. Esta falta de homogeneidad, no obstante, parece que viene ocasionada por el propio perfil y las características propias de cada una de las bases de datos, o lo que es lo mismo, por sus diferencias en cuanto a cobertura temática y temporal, tipo de información compilada, sistema de indexación utilizado, o servicios disponibles para el usuario.

A este respecto, Glänzel y Schubert (2003) llegan a afirmar que, "tras siglos enteros con numerosos y variados intentos de desarrollo de un esquema de clasificación perfecto, la solución más sensata y razonable se fundamenta en un enfoque pragmático". En esa misma dirección, Hampel (2002) considera que en la investigación actual sobre clasificación y, sobre todo, en el uso de técnicas de *clustering*, conviene tener presente que toda clasificación es, en cierta forma, más o menos arbitraria y que sus límites resultan difusos. Jacob (2004), por su parte, cree que todo esquema de clasificación es artificial y arbitrario: *artificial* porque no es más que un instrumento creado con el propósito expreso de instaurar una organización efectiva y significativa, y *arbitrario* porque los criterios utilizados para definir las clases en el esquema reflejan una perspectiva única del dominio que excluye todas las demás.

Con respecto a SJR, Jacsó (2013) indicó algunas posibles mejoras a implementar en diferentes aspectos de la plataforma con objeto de optimizar las búsquedas, la

comparación de las revistas o las preferencias en cuanto a la predominancia de las instituciones en los rankings. Una de estas mejoras apuntaba la clasificación de las revistas de SJR. Sin embargo, los trabajos destinados a la mejora de la clasificación de las revistas de la plataforma SJR habían comenzado años atrás, tanto con experimentos y procedimientos *ad hoc* desarrollados por el grupo SCImago, como a través de trabajos de investigación convertidos en publicaciones científicas.

De ese modo, en 2011 se publicó en la revista *Scientometrics* el artículo titulado "Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis" (Gómez-Núñez, Vargas-Quesada, Moya-Anegón, & Glänzel, 2011) encaminado a la mejora del sistema de clasificación y de la adscripción temática de revistas de SJR mediante el análisis de referencias bibliográficas. Más adelante, otras dos propuestas metodológicas basadas en el uso de algoritmos de *clustering* sobre diferentes combinaciones de medidas de citación han sido también llevadas a cabo para: 1) mejorar la precisión en la asignación de las revistas a las categorías temáticas del esquema de clasificación de SJR, y 2) actualizar y reajustar las categorías temáticas de SJR de acuerdo a las evoluciones experimentadas por la investigación y la ciencia y reflejadas en la literatura científica compilada por el sistema.

El esquema de clasificación de SJR es una evolución del sistema de categorías temáticas de la base de datos Scopus. Brevemente, podemos definirlo como un esquema jerárquico con dos niveles de agregación (áreas y categorías) orientado a la clasificación de revistas. Si bien es cierto que este esquema ha sufrido pocas modificaciones con respecto al modelo original, con la excepción de la inclusión de

categorías como ‘Nanoscience and Nanotechnology’ (Munoz-Ecija, Vargas-Quesada, Chinchilla-Rodríguez, Gómez-Nuñez, & Moya-Anegón, 2013) o ‘Social Work’, no lo es menos que la asignación de revistas a sus categorías temáticas ha sido objeto de continuas revisiones por parte del equipo de SCImago, de acuerdo a los criterios y opiniones expertas de los editores, el análisis de los objetivos y ámbito temático de las revistas, el estudio de los patrones de citación o la evaluación comparativa con otros esquemas de clasificación similares.

A pesar de los numerosos ejemplos de autores del ámbito de la Bibliometría que cuestionan la validez de la clasificación de la literatura científica y la delimitación de las disciplinas tomando como unidad de medida las revistas (Aksnes, Olsen, & Seglen, 2000; Börner et al., 2012; Gómez et al., 1996; Waltman & van Eck, 2012), las clasificaciones temáticas orientadas a revistas son una propuesta común, puesto que sus agregaciones pueden tomarse como índices de actividad, es decir, como indicadores de la organización intelectual de las ciencias y del intercambio de conocimiento que se produce entre los investigadores y académicos de diferentes disciplinas (Loet Leydesdorff, 2004). Archambault, Beauchesne y Caruso (2011) destacan como principal fortaleza de este tipo de clasificación los bajos requerimientos de costes y recursos, así como la facilidad para su implementación y posterior uso. Por su parte, Waltman y van Eck (2012) resaltan que las clasificaciones diseñadas a nivel de revistas tienen la ventaja de que las nuevas publicaciones recogidas por la base de datos pueden ser integradas directamente en el sistema.

Lo cierto es que el modelo basado en la clasificación de las revistas continua siendo de aplicación, por ejemplo, en los dos servicios de información científica comerciales de mayor prestigio y reconocimiento para la comunidad internacional: WoS y Scopus. El análisis de la literatura a este nivel permite también crear agregaciones de revistas (clústeres) para delinear las diferentes disciplinas científicas de una base de datos, determinar sus relaciones e interacciones (mediante texto o citación) y averiguar qué papel desempeñan determinadas revistas dentro de uno o varios clústeres o en la totalidad de una red.

Estos aspectos resultan claves a la hora de confeccionar indicadores bibliométricos, análisis de dominios y rankings basados en las revistas compiladas por las bases de datos, especialmente a niveles meso y macro, donde grandes agregaciones de revistas son necesarias, por ejemplo, para analizar la producción investigadora de un área científica concreta dentro de un país (por ejemplo, la medicina), o en todo un continente. La propia medida o indicador de impacto de referencia a nivel mundial para la evaluación de la ciencia y la investigación, es decir, el *Impact Factor (IF)*, que según Rossner, Van Epps y Hill (2007) tiene una gran in fluencia en la comunidad científica e influye en decisiones importantes como dónde publicar, a quién promocionar o contratar, la concesión de ayudas o subvenciones e incluso la asignación de pluses económicos, se construye en base al recuento de las citas de los documentos individuales que luego son agrupadas por revistas.

En base a las evidencias antes señaladas y teniendo en cuenta además que el principal objetivo perseguido por SJR no es otro que la generación de indicadores de revistas y

países para evaluar y analizar dominios científicos, consideramos que el sistema de clasificación orientado a revistas implementado en esta plataforma resulta adecuado para sus propósitos. No obstante, tal como ocurre con cualquier clasificación, es necesario diseñar un proceso continuado para su mejora, revisión y actualización que permita reajustar y reorganizar la literatura de acuerdo a los cambios y la evolución experimentados por la propia ciencia y la investigación. Como ya hemos mencionado, el fin no es otro que la actualización y mejora del sistema junto con la correcta adscripción de las revistas a lo largo de sus diferentes categorías temáticas para poder así, desarrollar indicadores estables, fiables y coherentes.

3.2. Objeto general de la investigación

El objetivo principal de este trabajo de investigación es el diseño e implementación de una serie de propuestas metodológicas basadas en la aplicación e integración de técnicas estadísticas, bibliométricas y de computación, orientadas a establecer procedimientos semi-automáticos para la mejora y optimización de la clasificación de la plataforma SJR. Dicha mejora persigue un doble fin:

1. La revisión y actualización periódica del esquema de clasificación de la plataforma SJR para ajustar las áreas y las categorías temáticas definidas en dicho esquema de acuerdo a las necesidades derivadas del dinamismo y de la constante evolución de la ciencia y la investigación. Este proceso permitirá, por ejemplo, la adaptación a los cambios ocasionados por las tendencias en la investigación. Nuestras propuestas se centran en la actualización y modificación del esquema fundamentalmente al nivel de las categorías temáticas, mediante:
    - La introducción de nuevas categorías.

- La eliminación de categorías obsoletas.
- La reestructuración de categorías existentes ya en uso a través de procesos como la fusión, la desagregación o, simplemente, el cambio de denominación.
- La eliminación de categorías demasiado genéricas y difusas, especialmente las etiquetadas como 'Miscellaneous'.

2. Refinar y ajustar la adscripción temática actual de las revistas incluidas en la plataforma SJR, clasificar adecuadamente las nuevas revistas incorporadas al sistema a posteriori y reubicar ciertas revistas previamente clasificadas en las nuevas categorías aceptadas e incorporadas al esquema de clasificación. El refinado en la asignación de las revistas incluiría, entre otros procesos:
   - La reducción de la multi-asignación de las revistas y, por consiguiente, el solapamiento entre las categorías del sistema.
   - La mejora en la distribución de revistas a lo largo de las diferentes categorías del sistema, evitando, en la medida de lo posible, grandes concentraciones en unas pocas categorías concretas.
   - El análisis y la evaluación de los cambios de asignación sufridos por las revistas, tanto por adición, pérdida o reemplazo de categorías temáticas.

En las diferentes propuestas metodológicas diseñadas se utilizarán técnicas de clasificación automáticas con objeto de reducir al máximo, y en la medida de lo posible, la intervención humana durante los procesos de clasificación, evitando así posibles sesgos y tendencias derivados de la misma. Igualmente, se procederá a implantar algún tipo de control conforme a varios criterios para tratar de evaluar las nuevas clasificaciones generadas en las dos vertientes de interés expuestas, es decir, en lo

referente a la actualización del esquema y en lo que concierne a la mejora en la adscripción de las revistas a sus categorías.

3.3. Objetivos Específicos

Los objetivos específicos que nos planteados giran en torno a una hipótesis clara de partida, que no es otra que nuestro convencimiento de que *las distintas técnicas o propuestas de clasificación expuestas en los artículos de investigación desarrollados e incluidos en esta tesis supondrán una sensible mejora de la clasificación de SJR*, entendiéndose ésta como un reajuste tanto del esquema jerárquico de clasificación como de la asignación o adscripción de las revistas en las categorías temáticas del sistema.

3.3.1. El análisis de referencias bibliográficas como técnica de clasificación de revistas

El análisis de referencias bibliográficas es una técnica bibliométrica que permite clasificar los documentos científicos en base a su lista de referencias bibliográficas. Para ello es necesario examinar alguna de las características distintivas de los documentos citados, como por ejemplo, sus palabras claves o las categorías temáticas asignadas originalmente en el sistema. En nuestro caso particular, las listas de referencias bibliográficas de los documentos fueron agregadas por revistas y se sometieron a un análisis cuantitativo para determinar el número de veces que una revista citaba a otra. Posteriormente, una nueva agregación por categorías de SJR de las revistas citadas permitió clasificar las revistas citantes a través de un simple recuento de citas, siendo las categorías que mejor representan a cada revista aquellas

con valores o frecuencias de citación más altos una vez ponderadas mediante el uso de porcentajes.

A nivel de artículo, Glänzel y sus colaboradores desarrollaron dos trabajos que se apoyan en el análisis de referencias bibliográficas para: (i) clasificar artículos publicados en revistas generales y multidisciplinares (Glänzel, Schubert, & Czerwon, 1999), así como (ii) clasificar artículos de revistas del área de Ciencias Sociales indizados por el SSCI (Glänzel, Schubert, Schoepflin, & Czerwon, 1999). Aunque ambos trabajos arrojaron resultados aceptables, sus autores creían oportuno la necesidad de aplicar ciertas mejoras metodológicas. Así, en el primero de los dos trabajos determinaron que la iteración del proceso de asignación de categorías podía servir para mejorar notablemente la eficiencia del método y, por consiguiente, los resultados finales obtenidos.

Utilizando niveles de agregación más amplios, Pinski y Narin (1976) introdujeron las medidas de influencia para medir la importancia de las revistas, las categorías o las áreas temáticas en función del análisis de las citas y las referencias bibliográficas. Archambault y otros autores (2011) diseñaron e implementaron un complejo método para elaborar una ontología de alrededor de 34.000 revistas y actas de congreso tanto de WoS como Scopus mediante un procedimiento algorítmico iterativo donde se integra un proceso de análisis y conteo de las citas y referencias bibliográficas de los artículos por categorías o disciplinas científicas.

Para el caso concreto que exponemos, esencialmente, hemos seguido los trabajos e ideas de Glänzel para tratar de diseñar una propuesta de clasificación fundamentada en el análisis de referencias como eje de la mejora de la clasificación de las revistas de SJR. Además, planteamos la posibilidad de ejecutar el proceso de forma iterada con objeto de refinar el procedimiento hasta alcanzar una clasificación apropiada y ajustada a nuestros fines. Partiendo de estos principios, trataremos de responder a la primera pregunta de investigación que nos planteamos en esta tesis: *¿Se puede actualizar el esquema de clasificación de SJR y, al mismo tiempo, refinar o ajustar la adscripción de las revistas a las categorías de dicho esquema utilizando una metodología basada en el análisis iterativo de referencias bibliográficas de las revistas?*

3.3.2. Algoritmos de *clustering* sobre redes basadas en citación de revistas

Una de las técnicas estadísticas multivariante más utilizada con propósitos de clasificación es el *clustering*. Pero además, dentro de la inmensa variedad de algoritmos y métodos de *clustering* existentes, el *clustering* jerárquico es uno de los más habituales. Este tipo de algoritmos se caracterizan por establecer agrupaciones de entidades u objetos (en base a una serie de características comunes o a su similaridad) que se organizan mediante una estructura jerárquica multinivel creada conforme a procedimientos *top-down* (*clustering* jerárquico divisivo), que van creando subdivisiones desde lo genérico a lo específico, o *bottom-up* (*clustering* jerárquico aglomerativo), que transcurren creando agrupaciones desde lo específico a lo genérico. En el caso que aquí nos ocupa, serán las propias revistas las que ocuparán el nivel más específico de la estructura jerárquica generada. En función de los intereses y objetivos que se persigan, esta estructura puede analizarse a diferente niveles de agregación o

jerárquicos, proporcionando agrupaciones de revistas más o menos extensas y equiparables a las disciplinas o áreas temáticas que forman parte de la ciencia y la investigación.

De entre los distintos tipos de algoritmos jerárquicos, los algoritmos de detección de comunidades han sido utilizados en diferentes estudios encaminados a la clasificación de la literatura científica. Dentro de este grupo de algoritmos destacan especialmente dos: (i) el método de Louvain (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) y (ii) el de VOS (Waltman, Eck, & Noyons, 2010). Ambos algoritmos tienen como objetivo principal la división de una red en diferentes módulos o comunidades a través de la optimización de la función de modularidad de Newman y Girvan (2004). Esto se traduce en la asignación de valores de modularidad altos para las mejores divisiones posibles de la red, que serán aquellas caracterizadas por una elevada densidad de enlaces entre los nodos de cada comunidad (*intra-clúster*) y reducido número de conexiones o links entre las diferentes comunidades (*inter-clúster*).

Ambos algoritmos son también capaces de descomponer una amplia tipología de redes, incluyendo redes de gran tamaño y densidad. Además, los dos algoritmos permiten parametrizar diferentes elementos claves para su funcionamiento, como el parámetro de resolución, el número máximo de iteraciones, el número máximo de niveles en cada iteración, o el número máximo de repeticiones en cada nivel. Analizando su uso con fines de clasificación en el ámbito bibliométrico, así como su alto potencial y eficiencia sobre grandes redes (en opinión de Waltman, Van Eck y Noyons (2010) si se introduce un valor suficientemente alto en el parámetro de resolución los clústeres de pequeño

tamaño pueden ser siempre identificados, siendo mayor el número final de clústeres cuanto mayor es el valor de dicho parámetro), decidimos adoptar la detección de comunidades por el método de Louvain y VOS como parte de nuestras propuestas de clasificación de revistas SJR.

Otro método de *clustering* utilizado con resultados contrastados en la clasificación de la literatura científica es el método de Ward (Janssens, Zhang, Moor, & Glänzel, 2009a; Zhang, Janssens, Liang, & Glänzel, 2010). Más concretamente, se trata de un algoritmo de *clustering* jerárquico aglomerativo que, a partir de un procedimiento *bottom-up*, es capaz de crear grupos de entidades en distintos niveles jerárquicos generados a lo largo de varias iteraciones. El procedimiento comienza por las unidades más específicas (*singletons*), que en nuestro caso serán las revistas y continuaría generando fusiones entre aquellos grupos de revistas que aparecen menos distantes entre sí de acuerdo con el criterio de la varianza mínima dentro de cada clúster. El proceso crea una estructura con diferentes niveles jerárquicos que finaliza con la formación de un único clúster en el nivel jerárquico superior, que estaría formado por todas las revistas incluidas en el análisis.

Para llevar a cabo nuestros experimentos y comprobar su rendimiento y eficiencia, ejecutamos los tres algoritmos de *clustering* sobre la red de revistas SJR resultante de combinar tres medidas basadas en la citación de revistas, como son la citación directa, la co-citación y el coupling, siguiendo para ello la combinación de medidas denominada *Weighted Direct Citation* propuesta por Persson (2010). Partimos de la suposición de que la integración de las tres medidas en una medida conjunta servirá

para incrementar la fortaleza de las relaciones establecidas entre las revistas, al tiempo que se maximizará el efecto clúster al ejecutar los algoritmos de *clustering*.

Sin embargo, la mezcla de medidas basadas en la citación (citación directa, co-citación y coupling) creada por Persson admite diversas interpretaciones en lo que a la propia integración de las medidas se refiere. Así, en el experimento basado en algoritmos para la detección de comunidades, la red de revistas final se construyó integrando las tres medidas mencionadas de acuerdo a esta expresión lógica:

Weighted Direct Citation = Citación Directa OR Co-citación OR Coupling

De esta forma, la matriz final de revistas estaría compuesta por aquellos pares de revistas donde al menos una de las tres medidas estuviera presente. A esta forma de integrar las tres medidas la hemos denominado *soft combination*.

Por el contrario, en el experimento basado en Ward, se utilizó una variante que incluía la fraccionalización de los valores de las tres medidas incluidas en la red de revistas, lo que Persson denominó *Normalized Wedighted Direct Citation*. Luego, la integración de las tres medidas se realizó siguiendo esta otra expresión lógica:

Normalized Weighted Direct Citation = Citación Directa con valores fraccionalizados AND Co-citación con valores fraccionalizados AND Coupling con valores fraccionalizados

A diferencia de la expresión anterior, utilizando esta forma de integración, la matriz final estaría compuesta sólo por aquellos pares de revistas que forzosamente dispongan de valores para cada una de las medidas incluidas, originándose así una integración de medidas que podríamos definir como *hard combination*.

Tomando como punto de partida los resultados obtenidos en el primer experimento mediante el uso del análisis de referencias bibliográficas, la aplicación de diferentes algoritmos de *clustering* sobre redes de revistas basadas en diferentes combinaciones de medidas derivadas de la citación debería traducirse en resultados y clasificaciones distintas, lo que nos lleva a plantearnos las siguientes cuestiones: *¿Se puede clasificar la totalidad de las revistas de SJR mejorando su adscripción a las categorías y, al mismo tiempo, actualizar el esquema de clasificación de SJR a partir de otros métodos alternativos al análisis de referencias como, por ejemplo el clustering, utilizando combinaciones de medidas basadas en la citación? En cualquier caso, ¿existen marcadas diferencias entre las clasificaciones generadas por los diferentes algoritmos utilizados y en base a las dos formas adoptadas para la combinación de las medidas de citación?*

3.3.3. Evaluación de los resultados de la clasificación

Un paso fundamental para medir la efectividad de los diferentes procedimientos metodológicos implementados de cara a la mejora de la clasificación de SJR es la validación y contrastación de los resultados en base a algún mecanismo de evaluación fiable y consistente. En las diferentes propuestas abordadas, dos han sido básicamente los métodos empleados para llevar a cabo esta tarea. Para ello, se han teniendo

siempre muy en cuenta los recursos necesarios y el tiempo disponible en cada momento para poder acometer la validación final de los resultados de cada una de las propuestas desarrolladas.

3.3.3.1. Evaluación por comparación

Uno de los sistemas de evaluación más sencillos de ejecutar consiste en efectuar un análisis por comparación de los nuevos sistemas de clasificación generados en relación con otros sistemas de clasificación ya existentes, sobre todo, con el sistema original de la plataforma SJR. También cobra especial relevancia la comparación con sistemas con una larga trayectoria, tradición y aceptación, como en el caso del sistema de categorías de WoS, que además es bastante cercano y parecido en cuanto a su estructura disciplinar jerárquica. El procedimiento a seguir puede, por ejemplo, fundamentarse en el diseño de un conjunto de indicadores aplicables a los diferentes sistemas evaluados. Un ejemplo claro de evaluación por comparación entre diferentes sistemas de clasificación generados tanto de forma automática como manual (en base al criterio de expertos) fue desarrollado por Rafols y Leydesdorff (Rafols & Leydesdorff, 2009). Esto nos lleva, por lo tanto, a formular la siguiente pregunta de investigación: *¿Resulta la evaluación por comparación un mecanismo eficiente y suficiente para poder contrastar y validar los resultados obtenidos en las diferentes propuestas de clasificación elaboradas en este trabajo de investigación?*

3.3.3.2. Evaluación mediante técnicas de visualización

Otro sistema viable para la evaluación de los resultados de clasificación obtenidos es la utilización dela visualización de la información, que incluye un conjunto de técnicas

que posibilitan representar gráficamente las entidades clasificadas junto con sus relaciones. A partir de ellas pueden luego identificarse las diferentes agregaciones de revistas generadas durante el procedimiento de clasificación. Esto resulta especialmente interesante y útil tras la aplicación de técnicas de *clustering* puesto que algunas herramientas como VOSViewer (van Eck & Waltman, 2010) o Pajek (Batagelj & Mrvar, 1997) integran diferentes algoritmos de *clustering* que pueden combinarse de forma simultánea con diferentes técnicas para la directa visualización de sus resultados.

Como se verá más adelante, muchos de los artículos y publicaciones sobre clasificación citados en el apartado de Antecedentes de este trabajo utilizaron de forma paralela algún procedimiento para la visualización final de sus resultados. Este hecho resultó fundamental para el desarrollo de una nueva pregunta de investigación bastante relacionada con la anterior: *¿Resulta la visualización de información una herramienta eficiente y suficiente para poder contrastar y validar los resultados obtenidos en las diferentes propuestas de clasificación puestas en marcha en este trabajo de investigación?*

4. Antecedentes

4.1. La clasificación de la ciencia a lo largo de la historia

La clasificación de la ciencia y el conocimiento ha sido un asunto ampliamente abordado a lo largo de la historia. Glänzel y Schubert (2003) no dudaron en afirmar que "la clasificación de la ciencia por medio de una estructura disciplinar es tan antigua como la propia ciencia en sí misma". Efectivamente, una mera revisión de fuentes de información de ámbitos diversos como la Historia, la Filosofía de la Ciencia o las Ciencias de la Información, resulta suficiente para localizar un número considerable de propuestas relacionadas con la clasificación y la organización del conocimiento en base a una estructura disciplinar. Este tipo de estructuras se constituyen sobre la base de un conjunto de disciplinas temáticas y de sus relaciones, las cuales se establecen en función de criterios como la subordinación, la similitud o la interacción entre disciplinas, dando lugar a diferentes modelos de organización: jerárquico, asociativo, facetado, etc. Su objetivo más inmediato es tratar de conseguir una imagen verosímil de la estructura subyacente de la ciencia y el conocimiento.

A este respecto, resulta interesante, el enfoque sistémico expuesto por Iyer (2012), que presenta el conocimiento como un sistema que puede subdividirse o fragmentarse en varios subsistemas interrelacionados o, lo que es lo mismo, en disciplinas. Éste precisó también que la evolución y el desarrollo histórico de la ciencia y el conocimiento, ha dado lugar a constantes subdivisiones de sus ramas y disciplinas, principalmente, debido a fenómenos como el crecimiento exponencial de la literatura científica, la especialización o la interdisciplinaridad. En este punto, resulta especialmente interesante el enfoque de Henry Bliss que, según Hjorland (2008b), se

caracteriza por su visión tradicional de la organización del conocimiento. En dicho enfoque, las ciencias tienden a representar el orden de la naturaleza. Este orden se refleja luego en las clasificaciones bibliotecarias, que tratarán de representar el orden del conocimiento tal y como se revela en la misma ciencia. Todo este proceso puede resumirse por medio del este sencillo esquema lógico:

NATURAL ORDER → SCIENTIFIC CLASSIFICATION → LIBRARY CLASSIFICATION (KO) [14]

Para los profesionales de las Ciencias de la Información y la Documentación, esto implica poseer un cierto conocimiento sobre la ciencia y el dinamismo que la caracteriza para que, de esta forma, puedan tener presente sus continuos cambios y avances a la hora de clasificar la literatura donde se recoge el conocimiento científico.

Clasificación de las ciencias y *organización del conocimiento* son conceptos que aparecen estrechamente ligados a las Ciencias de la Información, pero también a la *epistemología de la ciencia*, que se encarga del estudio del conocimiento científico y de los métodos, fundamentos, condiciones y procesos necesarios para su generación. Su origen se remonta a la época de la Antigua Grecia y se circunscribe al ámbito de la filosofía, destacando figuras influyentes como Aristóteles en Grecia, o Porfirio en Roma. Sin embargo, la clasificación de las ciencias también resultó de enorme interés para científicos y estudiosos del área de las Ciencias Naturales, donde sobresalen propuestas como la clasificación biológica ideada por Linneo (s. XVIII) o la posterior clasificación de las especies diseñada por Darwin ya en el siglo XIX.

[14] KO es la abreviatura normalmente utilizada para el término Gestión del Conocimiento, en inglés, *Knowledge Organisation*.

Pero, sin lugar a dudas, una de las áreas donde mayor importancia adquirió el estudio de la clasificación y la organización del conocimiento fue dentro de las ciencias de la Información y la Documentación. Así, ya en el siglo XVIII, las clasificaciones desarrolladas por los enciclopedistas y bibliógrafos franceses se convirtieron en el punto de partida de posteriores sistemas clasificatorios de gran notoriedad e importancia como, por ejemplo, las clasificaciones de Cutter o Dewey (s. XIX). Sin embargo, conviene destacar por encima de todos el proyecto desarrollado en los inicios del siglo XX por Otlet y Lafontaine, tanto por su envergadura como por las implicaciones que conlleva, puesto que la aparición de la *Clasificación Decimal Universal (CDU)* pasó a ser considerada por muchos teóricos y estudiosos de la materia como el origen de la documentación científica propiamente dicha.

Los autores antes citados, son sólo algunos de los ejemplos del amplio espectro compilado por San Segundo Manuel (1996) en un exhaustivo trabajo sobre la organización del conocimiento. En el citado trabajo se recogen figuras relevantes involucradas en el estudio de la clasificación de la ciencia y el saber desde la Antigüedad hasta nuestros días, pasando por las diversas etapas de la historia y por distintas áreas o campos del saber.

La tabla 2 presenta algunas de estas figuras agrupadas en tres áreas de conocimiento amplias y bien diferenciadas, como son: 1) *Filosofía y Teología*, donde se enumeran figuras claves del pensamiento filosófico, religioso y teológico; 2) *Ciencias Naturales o Experimentales*, que incluye autores relacionados con disciplinas como la biología o la

física y, finalmente, 3) *Ciencias de la Información,* que recoge enciclopedistas, bibliógrafos, bibliotecarios y otros profesionales de la información.

Por supuesto, esta clasificación debe considerarse artificial y relativamente precisa ya que su único objeto es presentar de forma clara y concisa algunos de los personajes destacados de la historia de la humanidad que, de algún u otro modo, han estudiado el fenómeno de la clasificación de las ciencias y el saber. Somos totalmente conscientes de que muchos de estos autores podrían ser incluidos en áreas distintas a las aquí propuestas. Igualmente, si atendemos a criterios como su formación multidisciplinar, el alcance temático de sus obras e investigaciones o su interés por diversos campos de la ciencia y objetos de estudio, entendemos que muchos de ellos podrían ser situados en más de un área de conocimiento a la vez. Por lo tanto, no dudamos de que otras clasificaciones y adscripciones alternativas serían completamente posibles y viables.

	Edad Antigua		Edad Media		Edad Moderna			Edad Contemporánea	
	Grecia	**Roma**	**Alta EM**	**Baja EM**	**S. XVI**	**S. XVII**	**S. XVIII**	**S. XIX**	**S. XX**
Filosofía y Teología	*Aristóteles*	*Porfirio*		*Ramón Llull*		*Francis Bacon*		*Auguste Comte*	*Rudolf Carnap*
		Plinio "El Viejo"				*Thomas Hobbes*			
						John Locke			
			San Isidoro de Sevilla	*Juan de Fidanza "S. Buenaventura"*					
				Roger Bacon					
Ciencias Naturales					*Konrad Gesner*		*Carlos Linneo*	*André M. Ampère*	
								Charles R. Darwin	
Ciencias de la Información	*Calímaco*				*François Grudé "La Croix du Maine"*	*Gottfried Leibniz*	*Jean le Rond D'Alembert*	*Melvil Dewey*	*James D. Brown*
					Alejo Venegas		*Denis Diderot*	*Charles A. Cutter*	*Henry E. Bliss*
							Jacques C. Brunet		*Paul Otlet*
									Henry Lafontaine

Tabla 2: Figuras relevantes de occidente involucradas en la clasificación de la ciencia y el saber

Analizando detenidamente la tabla, puede constatarse que todos los autores y estudiosos incluidos en ella pertenecen al mundo o cultura occidental. Sin embargo, el interés por la organización del conocimiento científico no sólo tuvo lugar en el seno de la cultura y la ciencia de Occidente. Así, en Oriente, surgieron también importantes figuras relacionadas con diferentes ámbitos de la ciencia que se interesaron en el estudio de la clasificación de la ciencia y el saber. Tal es el caso de figuras tan reconocidas como la del filósofo chino Confucio (ss. VI-V a.C.), el político e ideólogo Vladímir I. Lenin en la antigua URSS (ss. XIX-XX), o el matemático y bibliotecario de origen indio Shiyali R. Ranganathan (s. XX), que destaca por ser el creador de unos de los sistemas de clasificación más notorios y recientes en el ámbito de Ciencias de la Información y la Documentación: la *Clasificación Colonada*.

4.2. Métodos de clasificación y organización del conocimiento

Muchos autores del área de las Ciencias de la Información han apuntado un evidente y progresivo desuso del término *clasificación* en favor del término *organización del conocimiento*, que intrínsecamente posee unas implicaciones y una cobertura temática más amplias. Por ejemplo, Hodge (2000), estima que los denominados sistemas de organización del conocimiento abarcarían todos los tipos de esquemas de organización de información y gestión del conocimiento, incluyendo sistemas de clasificación y categorización, vocabularios controlados y estructurados, tesauros, redes semánticas y ontologías. No obstante, San Segundo Manuel (1996) considera que “el sentido que se da al nuevo concepto de organización del conocimiento proviene del concepto de clasificación de las ciencias”.

Hjorland, cuyos trabajos teóricos resultan fundamentales para el estudio y la comprensión de estas cuestiones, asimiló los métodos de clasificación desarrollados en las ciencias a los métodos de organización del conocimiento empleados en el área de las Ciencias de la Información y la Documentación, al considerar que comparten paradigmas epistemológicos esenciales (Hjorland, 2003). Entre los métodos de organización del conocimiento enumerados por Hjorland estrían:

- Estandarización [Standardization]
- Organización automatizada del conocimiento [Computer based knowledge organization]
- Métodos manuales o intelectuales [Manual or intelectual methods]
- Métodos cuantitativos [Quantitative methods]
- Métodos Cualitativos [Qualitative methods]
- Métodos basados en el texto [Text based methods]
- Métodos basado en personas [People based methods]
- Métodos basados en la organización institucional [Institutional based methods (university organisation)]
- Métodos bibliométricos [Bibliometrical methods]
- Métodos basados en frecuencia de las palabras [Word frequency based methods]
- Métodos sociológicos [Sociological methods]
- Métodos históricos [Historical methods]
- métodos críticos, epistemológicos y pragmáticos [Pragmatic, epistemological and critical methods]

Las clasificaciones o sistemas de organización obtenidos como resultado de aplicar estos métodos pueden ser clasificados a su vez en función de diversos principios o criterios, entre los que podemos señalar, por ejemplo:

1. El *procedimiento lógico* empleado en su desarrollo, lo que da lugar a clasificaciones deductivas o apriorísticas (de lo general a lo específico) e inductivas o a posteriori (de lo particular a lo general);
2. La *sintaxis o tipo de coordinación entre los términos* incluidos en la clasificación, que permitirá diferenciar, por ejemplo, entre sistemas precoordinados y postcoordinados;
3. El *modelo organizativo o tipo de relaciones* que se establecen en el esquema clasificatorio, que puede generar sistemas jerárquicos, asociativos, facetados y combinados o mixtos;
4. El *nivel u objeto de aplicación de la clasificación*, que posibilita la generación de clasificaciones orientadas a documentos fuentes (libros, revistas…), partes de documentos (artículos, ponencias…), o sustitutos del documento (resúmenes, reseñas…);
5. El *nivel de especialización* de la clasificación, que resulta esencial para distinguir entre clasificaciones universales, multidisciplinares, o especializadas;
6. El *grado de intervención o acción humana* en su desarrollo, que permite elaborar clasificaciones automáticas, manuales e híbridas o semiautomáticas (que combinan procedimientos automáticos y manuales al mismo tiempo).
7. La *inclusión* o *pertenencia* de los objetos o elementos a clasificar en las distintas clases o categorías temáticas provoca una distinción clara entre clasificaciones exclusivas y solapadas.

4.3. Clasificaciones automáticas: orígenes y propuestas

Como ya se ha dicho antes, este trabajo de investigación tiene como principal objetivo el desarrollo de metodologías para la mejora de la clasificación y la consiguiente adscripción temática de las revistas de SJR, plataforma que permite el desarrollo de análisis de dominios y rankings conforme a diversos indicadores bibliométricos. Por este motivo hemos considerado oportuno centrar nuestro estudio en la revisión de aquellas propuestas de clasificación que se han abordado en el campo de la Bibliometría, así como en otras áreas temáticas o campos afines, por ejemplo, la Cienciometría, la recuperación de información o la Estadística.

Un acontecimiento clave para el devenir y la evolución de estas disciplinas fue el lanzamiento de la primera edición de Science Citation Index (SCI) en 1963, primera gran base de datos e índice de citas multidisciplinar que fue desarrollada en el seno del ISI encabezado por Eugene Garfield (Cronin & Atkins, 2000). La información bibliográfica junto con la citación de los trabajos incluidos en esta base de datos permitió la generación de nuevos indicadores de gran valor para la evaluación y la toma de decisiones en política científica y de investigación.

Paralelamente, la cada vez más ingente producción científica y los crecientes problemas relacionados con su organización y almacenamiento dieron lugar a numerosos esfuerzos orientados a la optimización y mejora de la clasificación de la literatura en las bases de datos de literatura académica y científica. Estas mejoras perseguían, a su vez, un incremento de la efectividad en la búsqueda y recuperación

de información y en la delimitación temática de disciplinas a través de las publicaciones, con el propósito inmediato de poder diseñar e implementar medidas e indicadores consistentes, efectivos y fiables. La opinión de Todorov (1989) resume perfectamente el papel de las clasificaciones, su importancia y su utilidad dentro de las grandes bases de datos: "las bases de datos bibliográficas de ciencia (o sus versiones impresas) están creadas y desarrolladas sobre la base de clasificaciones más o menos elaboradas para categorizar los productos de la literatura. Los esquemas de clasificación en uso están orientados a disciplinas en la mayoría de los casos, y sus (sub)divisiones reflejan de algún modo las (sub)categorías o áreas científicas de la investigación actual".

Junto al nacimiento del SCI, la constante y acelerada evolución sufrida por la computación repercutió sobremanera en el diseño y la implementación de nuevos ensayos de clasificación basados en técnicas automáticas. Especialmente interesantes fueron los avances experimentados por el hardware, que se tradujeron en la consiguiente comercialización e industrialización de los primeros computadores personales ya en la década de los 70. A raíz de estos avances, dentro de las Ciencias de la Información comenzaron a ejecutarse nuevas pruebas y estudios enfocados, especialmente, a la recuperación de información y a la clasificación de documentos utilizando técnicas automáticas procedentes de la Computación, la Estadística o la Bibliometría. Pero el gran desarrollo, no sólo tuvo lugar en materia de hardware (aumento de la capacidad de almacenamiento, velocidad de procesamiento, etc.), sino también en lo que concierne al software (aparición de nuevos lenguajes, sistemas operativos, aplicaciones, etc.) y a la accesibilidad y disponibilidad de la información,

que alcanzaría su apogeo con el nacimiento de las redes de comunicación e Internet. Todo ello fomentó la puesta en marcha de experimentos cada vez más complejos y ambiciosos con conjuntos de datos más extensos y menos recursos de sistema y tiempo de ejecución.

Como ya hemos señalado, las clasificaciones de la literatura científica generadas mediante la aplicación de técnicas automáticas basadas en algoritmos y otros métodos de computación y estadísticos, como el cálculo de la probabilidad, el análisis factorial o el *clustering*, ha sido un tema de estudio recurrente y habitual en campos como la recuperación de información y la Bibliometría. En muchas ocasiones, estas clasificaciones han sido comparadas con aquellas generadas de forma subjetiva conforme a criterios heurísticos o empíricos y que se fundamentan en la asignación manual de documentos conforme al criterio o la opinión de expertos. Ya a finales de los años 50 comenzó a forjarse el interés en este asunto de la mano de autores como Luhn (1957). No obstante, tal y como afirma Garland (1982) este interés seguirá bastante vivo a lo largo de los años 60 y posteriores.

Archambault, Beauchesne y Caruso (2011) resumieron la opinión de diversos autores con respecto a los puntos fuertes y debilidades tanto de clasificaciones manuales como automáticas. Si bien es cierto que las clasificaciones basadas en el análisis subjetivo del contenido son más flexibles, su variabilidad, falta de uniformidad, y sobre todo, la cada vez más numerosa literatura científica, acabó derivando en una apuesta casi generalizada por el desarrollo de clasificaciones automáticas, que se caracterizan por necesitar menos recursos y tiempo de ejecución, gracias al descomunal avance de las

tecnologías de la información, la comunicación y la computación. No obstante, este tipo de clasificación tampoco está exento de problemas, como puede ser falta de proporción en el tamaño de los grupos temáticos generados o la asignación exclusiva de los documentos en las categorías, es decir la adscripción de cada documento en una única categoría temática del sistema.

4.3.1. Probabilidad

Existen numerosas definiciones y aproximaciones a la definición de probabilidad, siendo una de las más habituales la desarrollada en 1814 por Laplace, según la cual, la probabilidad se define como la razón o cociente entre el número de casos favorables y el de todos los casos posibles (Campos, 2004). La probabilidad de un suceso es un número, comprendido entre 0 y 1 que indica las posibilidades que tiene de confirmarse al realizar un experimento aleatorio, siendo poco probable si está cercano a 0 y bastante probable si está próximo a 1.

Maron puede considerarse como una de las figuras destacadas y uno de los pioneros en la introducción de técnicas estadísticas aplicadas a la clasificación automática. En 1961, realizó una investigación para comprobar el éxito de la clasificación automática en un conjunto de 405 documentos. Para ello, parte de un proceso de indexación encaminado a extraer los términos más representativos de cada documento para, posteriormente, asignarlos mediante cálculos probabilísticos a una de las 32 categorías temáticas de un esquema de clasificación creado a priori de forma empírica. En opinión del propio autor, “esta técnica estadística implica, 1) la determinación de relaciones de cierta probabilidad entre los términos significativos extraídos y las

categorías temáticas, y 2) el uso de esas relaciones para predecir la categoría a la que pertenece el documento que contiene los términos significativos" (Maron, 1961).

Las ideas y el trabajo de Maron sirvieron de fuente de inspiración a otros muchos autores como Borko y Bernick, quienes continuaron con la utilización de cálculos probabilísticos aplicados a la clasificación. Como fruto de su colaboración vieron la luz diversos trabajos sobre clasificación automática de documentos. En uno de ellos (Borko & Bernick, 1963), utilizaron análisis factorial para obtener un conjunto de categorías temáticas a partir de las palabras significativas de una colección de resúmenes de 405 documentos sobre computación. A continuación, diseñaron un procedimiento para la clasificación automática de los documentos sobre la base de una función de predicción bayesiana de las palabras de los documentos y una ponderación por factores (*factor loading*). Tras confrontar los resultados de la asignación automática con un proceso de adscripción elaborado por expertos, concluyeron que la clasificación automática de documentos era viable y efectiva y destacaron la utilidad del análisis factorial en el proceso de detección de categorías temáticas.

En base al trabajo anterior Borko y Bernick (1964) desarrollaron experimentos adicionales que les sirvieron para determinar que, si bien es cierto que no existen grandes diferencias en cuanto a la eficiencia en la adscripción de documentos establecida por medio de los métodos basados en la predicción bayesiana y el *factor scoring* (ordenación por factores), la clasificación automática de documentos aumenta

su efectividad cuando se dispone de un esquema de categorías previo generado mediante análisis factorial, tal y como se proponía en el trabajo de Maron citado arriba.

Años más tarde, Kar y White (1978) diseñaron un algoritmo que permitía la clasificación automática de documentos de forma secuencial a través de un indicador o medida de distancia bayesiana. Mientras el algoritmo iba recorriendo las partes del texto de los documentos y extraía los *keywords* (representados como vectores), la fórmula bayesiana predecía si a partir de la porción de texto leída el documento podía ser clasificado. En caso negativo, el algoritmo continuaba ejecutándose hasta llegar a clasificar el documento en una o varias categorías. El procedimiento, por lo tanto, podía constituirse en una o varias fases y su objetivo prioritario era el ahorro de tiempo y recursos destinados a la clasificación. Los resultados obtenidos en el uso del algoritmo con colecciones de documentos más o menos extensas, demostraron una buena proporción entre precisión y tiempo empleado.

Por otra parte, interesado en encontrar un grado de similitud fiable entre los documentos de cualquier colección, Kwok desarrolla diversas medidas de similaridad documental basadas, primero, en la indexación de términos de los títulos citados en los documentos científicos (Kwok, 1984) y, posteriormente, a partir las relaciones establecidas entre documentos citantes y citados (Kwok, 1985). En ambos casos Kwok aplicó diversas técnicas, que incluían cálculos de probabilidades basados, por ejemplo, en el teorema de Bayes. Esta medida resultó especialmente útil en el campo de la recuperación de información aunque el autor incide en su utilidad para la clasificación automática y el *clustering* de documentos.

4.3.2. Análisis Factorial

El análisis factorial es una técnica de análisis multivariante para la reducción de la dimensionalidad de los datos mediante la generación de grupos de variables observables y correlacionadas que pueden ser explicadas mediante un número menor de variables no observables, denominadas factores. Según Montoya Suárez (2007) mientras que las variables incluidas en cada grupo aparecerán altamente correlacionadas, paralelamente, los grupos deberían estar poco correlacionados o relativamente incorrelacionados entre sí, es decir, deberían ser independientes.

Por medio del análisis factorial, Borko (1962) intentó demostrar cierta equiparación entre las clasificaciones de documentos obtenidas de forma manual y las derivadas mediante procedimientos automáticos. Para ello, elabora un sistema de clasificación inicial basado en técnicas matemáticas y empíricas, aplicando análisis factorial a una matriz cuadrada de correlaciones compuesta por 90 términos índice (con una ocurrencia de 20 o más) extraídos de los resúmenes de una muestra de 618 documentos de psicología. La clasificación obtenida se comparó con el sistema de clasificación (Psychological Index) desarrollado por la American Psychological Association y resultó bastante similar. Por ello, los autores destacaron la viabilidad de las técnicas automáticas de clasificación y la utilidad del análisis factorial para la detección de las dimensiones (clases) principales de los sistemas de clasificación.

Dos años más tarde (Borko, 1964) lleva a cabo un nuevo estudio para determinar la fiabilidad y la efectividad de las clasificaciones establecidas por especialistas y medir la

precisión de las técnicas de clasificación automática de documentos por medio de su posterior comparación. El proceso incluía la aplicación de análisis factorial sobre una matriz de términos extraídos de un conjunto de abstracts de 997 artículos de psicología a partir del cual se identificaron 11 categorías temáticas. Utilizando estas 11 categorías, se encargaron diferentes clasificaciones manuales generadas por tres licenciados en psicología. Seguidamente, estas clasificaciones fueron contrastadas con otra automática derivada de la aplicación de una ecuación de ordenación por factores (factor-score) que, asignando pesos a los términos significativos que componían cada vector, posibilitaba la clasificación de los documentos en una de las 11 categorías. La fiabilidad de la clasificación automática resultó bastante alta, como demostraba su alto grado de correlación (0,766) con las clasificaciones manuales.

Por su parte, Van Cott y Zavala (1968) consideraron necesario el desarrollo de un método que permitiese agrupar la literatura científica en diferentes disciplinas y fraccionarlas, a su vez, en áreas temáticamente homogéneas para evitar problemas como el solapamiento en este tipo específico de literatura. Con el propósito de definir una estructura temática eficiente, aplican análisis factorial sobre los datos extraídos de un conjunto de abstracts de 405 revistas científicas del área de la física y correspondientes al año 1961. El trabajo incluía varios análisis que permitieron delimitar, por un lado, grupos consistentes de revistas y, por otro, una serie de áreas temáticas bien definidas a través de los factores obtenidos.

En el campo de la bibliometría, Leydersdorff puede considerarse como uno de los autores más prolíficos en lo referente al uso del análisis factorial tanto con fines

clasificatorios como de visualización. Con objeto de detectar la estructura temática subyacente del SCI y el dinamismo del sistema, aplicó análisis factorial de forma iterativa sobre una red de citación agregada al nivel de revistas a lo largo de varias series temporales (Leydesdorff & Cozzens, 1993). Respecto a la aplicación de análisis factorial con propósitos de visualización destaca su trabajo para la creación de un mapa global de la ciencia a partir de la matriz de citación agregada al nivel de las categorías temáticas de ISI (Leydesdorff & Rafols, 2009), utilizando como base el trabajo previo desarrollado por Moya-Anegón en colaboración con otros autores del grupo de investigación SCImago (2007).

4.3.3. *Clustering*

El *clustering* o análisis de conglomerados es otra técnica multivariante para la reducción de datos que permite generar grupos o clústeres de entidades u objetos a partir de sus atributos o similaridades. Börner, Chen y Boyack (2005) definen el *clustering* como una técnica para clasificar una montaña de información en un montón significativo y manejable. Partiendo de esta definición, resulta evidente que el *clustering* puede desempeñar un papel destacado en los procesos automáticos de clasificación de la literatura científica, dando lugar a la generación de grupos de documentos temáticamente relacionados que, una vez analizados con detenimiento, pueden ser identificados y definidos como categorías temáticas en un sistema de clasificación. Estas categorías, a su vez, pueden ser equiparadas o entendidas como las diferentes disciplinas en las que se subdividen la ciencia y la investigación. Por otra parte, gracias a la optimización de los cada vez más avanzados algoritmos de *clustering* y al incesante desarrollo de la computación, la descomposición de grandes redes,

como es el caso de las redes de citación de revistas o documentos, resulta factible sin necesidad de disponer de una gran cantidad de recursos para su ejecución.

Existen numerosos métodos y algoritmos de *clustering* que han sido clasificados de muchas y diversas formas. En un trabajo orientado a la aplicación de procedimientos de *clustering* para la recuperación de información en la web, Mateos Sánchez y Garcia-Figuerola Paniagua (2009) utilizaron la clasificación desarrollada anteriormente por He (1999) para dividir las diferentes técnicas de *clustering* en:

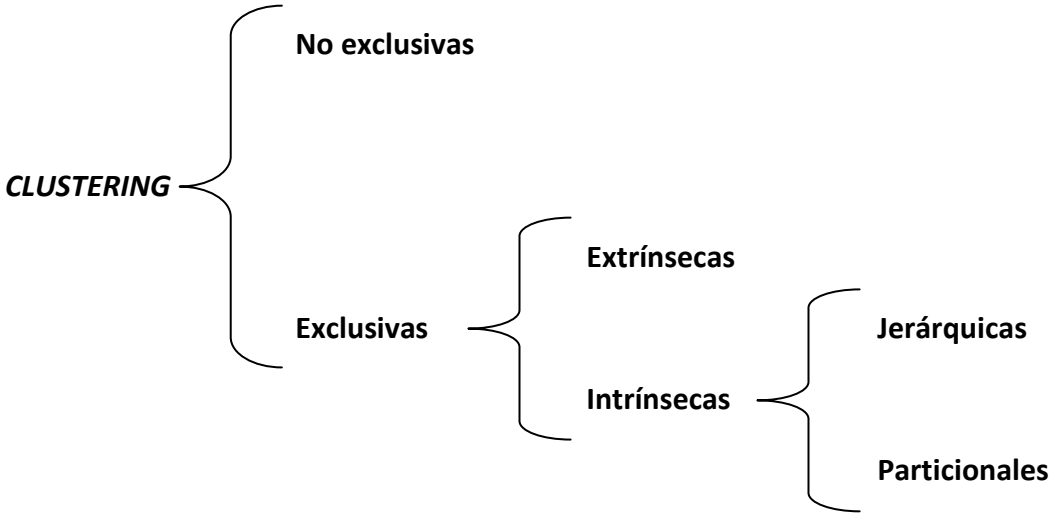


Figura 2: Clasificación de las distintas técnicas de *clustering* según Q He (1999)

En lo relativo a la aplicación de técnicas automáticas y estadísticas empleadas con fines clasificatorios, Cormack (1971) elaboró una exhaustiva revisión bibliográfica donde recopiló algunos conceptos útiles fundamentales para el diseño de clasificaciones. Cormack, no obstante, centró su atención en el extenso número de técnicas de *clustering* (en especial, aquellas con modelos matemáticos mejor formulados), así como en las distintas medidas de similaridad y disimilaridad en uso, proporcionando

una amplia visión de sus principios fundamentales, sus ventajas y sus inconvenientes, más que detalles propios de su implementación.

La aplicación de *clustering* en procesos automáticos de clasificación ha sido una de las propuestas más extendidas y frecuentes a lo largo de la investigación llevada a cabo, por ejemplo, en campos como la Bibliometría, la recuperación de información o la visualización de información. Doyle (1965) realizó una revisión de artículos y presentó algunos aspectos básicos sobre la clasificación automática, exponiendo sus posibles ventajas e inconvenientes. Además, introdujo una propuesta para mejorar la clasificación automática que parte de la hipótesis de que la clasificación de un documento mejorará cuanto mayor sea la cantidad de información relevante que se extrae del mismo. Para tal fin, utiliza un conjunto de 100 documentos que representa a través de una lista de 36 términos representativos ordenados por su frecuencia. A continuación, un programa automático se ejecutaría en seis ocasiones tomando consecutivamente 12, 15, 19, 24, 30 y 36 términos de la lista. Finalmente, el programa agrupaba jerárquicamente los documentos formando una estructura de árbol que facilitaba la integración de los documentos en grupos.

A finales de los 60 Price y Schiminovich (1968) proponen la creación de un sistema de clasificación basado en el análisis de la relación entre los documentos de una muestra de 240 artículos sobre física teórica de altas energías utilizando bibliographic coupling. En función de sus relaciones, agrupan los documentos mediante un procedimiento de *clustering* manual que podía ser fácilmente implementado a través de un algoritmo o software. Tres años más tarde, Schiminovich (1971) realiza una nueva propuesta de

clasificación automática utilizando un algoritmo de detección de patrones de bibliographic coupling entre los documentos. El proceso, que mostró tiempos razonables de ejecución, permitió la autogeneración de clústeres de documentos significativos que se correspondían con temas de interés en la investigación sobre física. Luego, a partir de cada clúster construyó una bibliografía con los documentos citados en cada agrupación. Los nuevos documentos eran posteriormente asignados a estos clústeres de forma automática en función de un determinado umbral de citación.

Carpenter y Narin (1973) fueron pioneros en tomar las revistas como unidad de análisis al aplicar *clustering* sobre un conjunto de 288 revistas relevantes pertenecientes a las disciplinas de física, química y biología molecular para obtener grupos temáticamente interrelacionadas con objeto de facilitar la creación de sub-disciplinas y la agrupación de las revistas a un nivel más preciso. Para ello, asumieron dos premisas básicas: 1) la existencia de patrones de citación similares entre las revistas de una misma área temática, y 2) la existencia de reciprocidad en la citación de las revistas que pertenecen a la misma área temática. En su opinión, la clasificación o asignación temática asignada a una revista podía ser luego transmitida a los artículos y documentos que forman parte de ella.

Volviendo a los documentos como unidad de análisis, Small y Griffith (1974) desarrollaron una técnica automática de detección de clústeres de documentos altamente citados (umbral ≥ 10) de SCI tomando la co-citación entre pares de documentos como unidad de medida. Su principal objetivo era representar la estructura de la ciencia a través de especialidades con altos niveles de actividad,

partiendo de la evidencia de que el rápido crecimiento y desarrollo de una especialidad suele estar unido a la aparición de determinados documentos clave que se citan de forma rápida y frecuente.

Ese mismo año, con objeto de mejorar la recuperación de información, Bichteler and Parsons (1974) presentaron una versión modificada del algoritmo de Schiminovich (1971) para la clasificación automática de los artículos de una base de datos a partir de las relaciones bibliográficas establecidas entre los documentos. Aquellos documentos con patrones de citación similares dieron lugar a la formación de grupos o clústeres homogéneos. Los resultados generados por medio de la clasificación automática fueron luego comparados con los obtenidos mediante análisis de contenido tradicional. Finalmente, realizaron una combinación de ambos métodos para mejorar y refinar los resultados mediante la complementariedad de ambas técnicas.

Kwok (1975), por su parte, realizó una selección de documentos con un alto nivel de interrelación para lanzar un algoritmo de *clustering* con capacidad para clasificar automáticamente los documentos en base a las representaciones generadas a partir de sus títulos y de los títulos citados en sus listas de referencias bibliográficas. De esta forma, pretendía equiparar los grupos de documentos con temas de investigación que, posteriormente podrían ser reagrupados en áreas temáticas.

En la década de los 80, destacan los trabajos de algunos autores como Croft (1980), que presenta un modelo de búsqueda aplicable a los sistemas de recuperación de información en el que se propone la clasificación de las consultas (*queries)* enviadas a

los sistemas mediante *clustering*. De esta forma, la consulta se encuadraba dentro del clúster de documentos con una mayor probabilidad de resultar relevante, procediéndose a continuación, a la recuperación de ese clúster concreto. Así pues, se trata de un modelo de clúster probabilístico basado en la denominada regla de decisión de Bayes. Los resultados obtenidos en la recuperación demostraron tener mayor eficiencia que las búsquedas basadas en otros modelos de clúster como el heurístico.

Por su parte, Garland (1983), partiendo de la premisa de que los *keywords* pueden considerarse indicadores temáticos representativos del contenido de los documentos, desarrolla un método de clasificación jerárquica automatizado utilizando la técnica de clúster denominada single-link, que aplica a una muestra de 416 monografías pertenecientes a la categoría de ciencias en la clasificación desarrollada por la Biblioteca del Congreso de los Estados Unidos. El proceso de agrupación de los documentos se estableció mediante el emparejamiento de los *keywords* y la aplicación de seis umbrales de corte distintos. Los resultados obtenidos demostraron la existencia de relación entre los clúster de documentos generados y las clases asignadas por la Biblioteca del Congreso a los mismos documentos, aunque el método presentaba ciertas limitaciones susceptibles de ser mejoradas.

A finales de los 80, Todorov (1989) propone un método alternativo a la co-citación para representar las relaciones entre las subcategorías temáticas en un esquema de clasificación. Para ello diseña una matriz que recogía la frecuencia o número de artículos de una revista (filas) siendo clasificados en una determinada categoría

(columnas). Seguidamente, aplica un proceso de *clustering* iterativo que persigue representar las relaciones entre las revistas y las subcategorías analizadas para medir el grado de asociación entre las subcategorías. Como alternativa a la co-citación propone utilizar la coocurrencia de los artículos de las 95 revistas especializadas de la categoría de física de la materia condensada como unidad de medida.

Mucho más reciente resultan propuestas como la elaborada por Spasser (1997) que, con la intención de visualizar y explorar el área de farmacia a través de su literatura, utiliza técnicas estadísticas multivariantes como MDS y análisis de clúster aplicadas sobre matrices de coocurrencia de códigos y encabezamientos temáticos utilizados en la categorización de las publicaciones de la base de datos International Pharmaceutical Abstracts (IPA). Las conclusiones del trabajo destacan la facilidad del método para ofrecer una visión global de una categoría o área temática y de su estructura. Por lo tanto, la propuesta de Spasser podría aplicarse para estructurar y delimitar categorías siempre y cuando se disponga de un esquema o proceso de clasificación previo.

Bassecoulard y Zitt (1999) desarrollaron una propuesta para la clasificar y agrupar revistas de SCI/CMCI (Science Citation Index/Computer and Mathematics Citation Index) en distintos niveles (categorías y áreas temáticas) utilizando técnicas de *clustering*. Para ello, propone tres pasos: 1) la selección del conjunto de revistas más citadas dentro de SCI/CMCI, 2) la construcción de una matriz de similaridad con las revistas nucleares y su agrupación en 150 especialidades mediante clúster jerárquico y, finalmente, 3) la reasignación de todas las revistas a estas categorías.

4.3.4. Métodos bibliométricos

Hasta el momento, hemos visto como muchas de las propuestas automáticas descritas anteriormente se caracterizan por una mezcla de diferentes técnicas y métodos, principalmente técnicas estadísticas y de computación, con el único fin de desarrollar clasificaciones y asignaciones automáticas de documentos a categorías o disciplinas temáticas efectivas y sólidas. Así, se han revisado, por ejemplo, trabajos donde se han utilizado algoritmos de *clustering* o análisis factorial en combinación con cálculos probabilísticos. No obstante, la mayoría de estos trabajos necesitan de un componente bibliométrico previo que se fundamenta en el estudio o análisis de las relaciones o el texto de una colección de documentos, dando lugar a listas de adyacencia o matrices, por ejemplo, de citación o términos sobre las que luego se han ejecutado los procedimientos de clasificación automáticos diseñados. Más concretamente, este componente puede estar formado por un análisis de los patrones de citación o de sus derivados, es decir, citación directa, co-citación, o bibliographic coupling (véanse los trabajos de Persson (2010) o Boyack y Klavans (Boyack & Klavans, 2010)), el conteo de la frecuencia de los términos (Boyack et al., 2011), o el análisis de la coocurrencia de palabras (Cantos-Mateos, Vargas-Quesada, Chinchilla-Rodríguez, & Zulueta, 2012) o de publicaciones (Todorov, 1989b), etc.

El análisis de las citas y las referencias bibliográficas de los documentos ha sido una propuesta habitual en el ámbito bibliométrico de cara a generar indicadores y medidas de impacto o de influencia, pero también en los procedimientos automáticos de adscripción temática, clasificación y delineación de la literatura científica. Pinski y Narin (1976) introdujeron las medidas de influencia para medir la importancia de las

revistas, las categorías o las áreas temáticas en función del análisis de las citas y las referencias. Así afirman que "en el continuo intento de evaluar, categorizar y medir el rápido crecimiento del conocimiento, la habilidad para medir la influencia de un área de actividad sería una ventaja valiosa. Un esquema para evaluar la influencia de la investigación científica en una categoría de la ciencia o dentro de una institución podría servir como una ayuda en la gestión para valorar la efectividad de la labor o empresa científica, así como para proporcionar datos para los estudios sobre política científica". Ese mismo año Narin, Pinski y Gee (1976) realizan un estudio de la literatura de biomedicina y clasifican una muestra de 900 revistas a través de un esquema de casi 50 categorías independientes y cuatro niveles de investigación. Las categorías se establecen en función de una asignación subjetiva derivada del examen visual y los patrones de citación establecidos entre las revistas y obtenidos a través del análisis de las referencias bibliográficas. Este trabajo serviría también de inspiración a otros trabajos posteriores, como el de Lewison y Paraje (2004) que proponen la clasificación de las revistas de biomedicina de SCI en función de tres niveles de investigación: clínica, básica e intermedia.

Glänzel, junto a diversos colegas, publicó dos artículos centrados en el análisis de referencias como método de clasificación de artículos científicos. Partiendo de la premisa de que el tema o materia de cualquier artículo viene determinado por su lista de referencias, propone un método de clasificación que denomina *item-by-item* y que pretendía clasificar los artículos incluidos en las revistas científicas en función del análisis detallado de sus listas de referencias bibliográficas. De esta forma, las

categorías temáticas de las fuentes que aparecen con mayor frecuencia entre las referencias de cada artículo son consideradas las categorías que mejor los definen.

En el primero de estos dos artículos (Glänzel, Schubert, & Czerwon, 1999), los autores proponen la utilización del análisis de referencias aplicado a la clasificación de los trabajos publicados en revistas generales y multidisciplinares. Su estudio se centra en trabajos recogidos por el SCI en el periodo temporal comprendido entre 1989-1992. Si bien el método resultó eficiente en la delimitación de las áreas y categorías temáticas principales, es cierto que no está exento de problemas, como en el caso de las referencias realizadas a publicaciones fuentes no recogidas por la base de datos, los documentos carentes de referencias bibliográficas, o las auto-citas. Estos inconvenientes se tradujeron, en la imposibilidad de asignación temática para un elevado número de artículos. No obstante, los autores consideran que la iteración del proceso de asignación de categorías, podría mejorar notablemente la eficiencia del método.

En el segundo artículo (Glänzel, Schubert, Schoepflin, & Czerwon, 1999), se centran en la aplicación del análisis de referencias sobre los artículos del SSCI. A diferencia del SCI, donde la mayoría de los tipos documentales referenciados en las revistas están incluidos en la base de datos, las revistas del SSCI se caracterizan por tener una cobertura parcial, mucho más limitada en cuanto al número de documentos referenciados y registrados. Esta cuestión dificulta enormemente la transferencia de las categorías a los artículos puesto que, generalmente, este método sólo es aplicable a las revistas con una cobertura total en la base de datos. También la proporción de

trabajos que no disponen de referencias bibliográficas u otros elementos de interés como la filiación de los autores, es mucho más alto en el ámbito de las Ciencias Sociales y Humanidades que en el de las "ciencias puras" debido a los hábitos de publicación y citación característicos y específicos de estas dos áreas. Una vez definidos los problemas más importantes, los autores proceden a identificar siete grandes áreas temáticas en Ciencias Sociales que se corresponderían con agrupaciones de varias categorías temáticas de ISI y, a continuación, proceden a la adscripción de los artículos a estas categorías en función de la adscripción temática previa de sus fuentes.

Mucho más reciente es el trabajo de López-Illescas junto a otros colegas (2009), donde se establece una propuesta para la delimitación de las categorías de Oncología y Cardiología a partir de una estrategia combinada basada en categorización de las revistas especializadas de WoS y un proceso de análisis de referencias inspirado en los dos trabajos de Glänzel citados justo antes. De forma más precisa, la delimitación de las categorías se llevó a cabo mediante la identificación de las revistas especializadas de las categorías de Oncología y Cardiología (*subfield's specialist journals*) a las que, posteriormente, se añade un grupo adicional de revistas no especializadas (*additional journals*) cuyos artículos aparecen citando a trabajos pertenecientes al grupo de revistas especializadas a partir de un umbral determinado.

Son muchas las ocasiones en las que el análisis de las relaciones que se producen entre las revistas científicas de una red de citación ha sido utilizado para la estructuración y delimitación de categorías científicas. Leydesdorff (2002) plantea el diseño de esquemas dinámicos y evolutivos que permitan actualizar y definir a posteriori los

posibles cambios que se producen en las categorías de la ciencia y la investigación, representadas mediante conjuntos de revistas obtenidos en base a las relaciones establecidas a través de sus citas. De esta forma, como afirma Wagner (2005) "los científicos pueden identificar por sí mismos su categoría o campo observando las revistas que citan, más que a través de una estructura preconcebida del área".

Por su parte, Zitt y Bassecoulard (2006) establecieron una propuesta para la delimitación de categorías temáticas complejas basada en redes de citación y entradas léxicas que aplicaron al campo de la nanotecnología. El método se divide en distintos pasos: 1) Se construyen y diseñan consultas con la ayuda y la opinión de expertos de cara a formar una bibliografía inicial (denominada "semilla") sobre nanotecnología; 2) los trabajos citados por este grupo semilla de documentos, se denominan "núcleo citado" y todos los trabajos que citen a este núcleo, son también incorporados a al grupo semilla inicial, asumiendo que ambos grupos comparten la misma base intelectual; 3) un tercer paso de gran interés pero no implementado en su propuesta era la aplicación de *clustering* sobre redes de citación o de términos (léxico).

Existen también otras estrategias para la definición de categorías temáticas mediante la delimitación de la literatura científica, cuyo enfoque más tradicional se basa en el diseño de ecuaciones o filtros temáticos para recuperar revistas especializadas a través de palabras claves de sus títulos. Sin embargo, según Lewison (1996, 1999), dicho enfoque no resultará suficientemente apropiado ni exhaustivo si no se tiene en cuenta que, una gran cantidad de autores, publican tanto en revistas especializadas, como en revistas generales y multidisciplinares, tal como sucede en el campo de la biomedicina.

Por ello, Lewison propuso elaborar las ecuaciones o filtros temáticos utilizando, tanto el nombre de las revistas especializadas, como las palabras claves o *keywords* de los títulos de los artículos, lo que le permitió incrementar la exhaustividad y la precisión en la recuperación, que superó el 90% en algunas categorías de biomedicina. Para tal fin propuso además calibrar los filtros o ecuaciones por medio de un factor de calibración.

Partiendo de los trabajos de Lewison, otros autores como Costas y Bordons (2008) propusieron la aplicación de un filtro temático mixto basado en revistas especializadas y descriptores o palabras clave del título para delimitar el área interdisciplinar de Ciencias del Mar a partir de los datos de SCI, SSCI y AHCI. Los resultados obtenidos por el filtro mixto presentaron un 69% de precisión y un 77% de exhaustividad.

Otro método alternativo para la delimitación de categorías científicas a través de las palabras significativas extraídas de la filiación o direcciones institucionales de los autores de los trabajos científicos fue propuesto De Bruin y Moed (1993). Tomando como muestra los datos de un conjunto de artículos del SCI publicados en las revistas *Science* y *Nature* en 1985, los autores elaboran un árbol genealógico específico para descomponer cada una de las direcciones institucionales en diversos niveles con el fin de aislar las palabras significativas de aquellas vacías o ambiguas. A continuación, construyen tablas con los términos significativos más frecuentes y con aquellos términos que coocurren en un mismo trabajo, bien en una misma dirección, o en direcciones distintas. Finalmente, realizan *clustering* utilizando el método del vecino más cercano (*single-link*) y aplicando como medida de similaridad el coseno de Salton.

La metodología propuesta por De Bruin y Moed resulta de gran interés para la elaboración de mapas científicos y la visualización de las categorías obtenidas a partir de la coocurrencia de las palabras significativas del campo filiación de los artículos científicos y sus relaciones. No obstante, su método necesita una validación y una mejora para solucionar problemas, como, las diferentes formas en las abreviaturas empleadas en los términos significativos, el uso de diferentes idiomas, la ausencia de direcciones en el ámbito de las Humanidades, las categorías y términos significativos "ocultos" bajo el nombre de ciertas instituciones (como IBM o Karolinska Institutet), etc. Los mismo autores, afirmaron se trata de un método que "sólo puede ser utilizado como una herramienta adicional, combinada con otros métodos de delimitación".

Mientras que en esta sección hemos tratado de representar una visión general sobre los orígenes y el desarrollo de las diferentes técnicas y métodos automáticos para la clasificación de la literatura en el ámbito de la bibliometría, otras propuestas y trabajos de investigación más actuales destinadas a la clasificación automática y la delimitación de categorías temáticas a partir de técnicas estadísticas, computacionales y bibliométricas pueden ser localizadas en la pertinente revisión bibliográfica de la literatura científica relacionada y citada en cualquiera de los cuatro artículos científicos emplazados en la Parte II de esta tesis doctoral.

5. Limitaciones

El punto de partida para la implementación de cada uno de nuestros procedimientos de clasificación es la selección de la fuente de información a partir de la cual se extraerá el conjunto de datos necesarios para abordar cada experimento. Para el desarrollo de nuestras propuestas se seleccionaron y recuperaron los datos de Scopus incluidos en la plataforma SJR. Aun siendo la mayor base de datos de literatura científica de la actualidad (con más de 20.000 títulos de revistas de impacto) y a pesar de su inequívoco perfil multidisciplinar e internacional y su extensa cobertura temática, tal y como sucede en cualquier fuente de información de sus características, Scopus presenta diferentes tipos de sesgos (Moya-Anegón et al., 2007), entre los que destacan:

- *Sesgo temático* reflejado, por ejemplo, en una extensa representación de áreas como la Ingeniería y, en especial, las Ciencias de la Vida y de la Salud.
- *Sesgo temporal* ocasionado como consecuencia de indexar únicamente aquellas publicaciones originadas en el periodo temporal transcurrido desde el año 1966 hasta la actualidad e, igualmente, como resultado de compilar sólo las referencias bibliográficas citadas desde 1996 en adelante.
- *Sesgo geográfico* hacia ciertos países anglosajones como UK y, principalmente, USA.
- *Sesgo idiomático* claro en favor del inglés y escasa cobertura de otros idiomas.
- *Sesgo editorial* derivado principalmente de la inclusión en el sistema de amplios paquetes de revistas pertenecientes a grandes grupos editoriales, como Springer, Taylor & Francis, Blackwell y, sobre todo, del propio Elsevier.

Otras peculiaridades importantes de cualquier fuente de información como Scopus tiene que ver con el mayor o menor grado de cobertura de las referencias citadas por la literatura que indiza, o en otras palabras, con el total de referencias bibliográficas citadas que apuntan dentro y fuera de la base de datos. El primer grupo estaría integrado por los artículos, contribuciones a congresos y otros materiales específicos (*reviews*, *letters*…) que, habitualmente son indizados y recogidos por el sistema. El segundo grupo, por su parte, hace referencia a determinados tipos documentales como libros, informes, documentos de archivos, literatura gris, etc., más propios de áreas de Ciencias Sociales o Artes y Humanidades. Este tipo de literatura cuenta con una presencia mucho más limitada dentro de este tipo de bases de datos y deben ser tenidos en cuenta, por ejemplo, a la hora de interpretar los resultados de determinados análisis o estudios bibliométricos.

Pero además, la adopción de la citación y/o sus derivados como unidad de medida y de relación entre las revistas implica una premisa básica: la existencia de un alto nivel de precisión en las relaciones o asociaciones que se establecen entre los trabajos y las fuentes que se citan en ellos. En muchas ocasiones, los errores tipográficos en las referencias o, tal como hemos mencionado antes, un alto porcentaje de referencias citadas que apuntan a documentos fuera de la base de datos, puede dar lugar a la tergiversación de los resultados obtenidos en los diferentes experimentos desarrollados.

No obstante, en lo que concierne a este asunto, Scopus ofrece una ventaja clara con respecto a sus competidores, que no es otra que la posibilidad de poder generar matrices de citación ítem-ítem en lugar de revista-revista. Es decir, a partir del conjunto de datos específico recuperado para cada uno de nuestros experimentos, se calculó una matriz inicial en base a las relaciones constituidas entre los ítems (documentos) que fueron luego agregados al nivel de las revistas. A priori, esta forma de proceder asegura un mayor nivel de precisión en las asociaciones que se establecen entre las referencias bibliográficas y los documentos fuentes a los que apuntan, limitando así el número de errores posibles.

Una última cuestión de relevancia que ha generado un amplio debate entre los expertos y estudiosos de la Bibliometría tiene que ver con la elección de la revistas como unidad de análisis. Anteriormente, ya hemos señalado algunos autores que abogan por tomar los documentos de forma individual como la unidad de análisis más apropiada sobre las que ejecutar cualquier procedimiento de clasificación en lugar de seleccionar las revistas que los engloban. No obstante, de acuerdo con el tipo y el nivel de análisis que puede configurarse a través de la plataforma SJR, principalmente análisis a niveles meso y macro, por países y regiones más o menos extensas, parece que el sistema de clasificación adoptado y aplicable a las revistas surge como una solución apropiada y efectiva, más aún, cuando las propias revistas son la unidad de referencia y análisis mínima incluida en la plataforma SJR.

6. Materiales y Métodos

Como es obvio, cada una de las cuatro propuestas metodológicas enfocadas a mejorar la clasificación de la plataforma SJR e integradas dentro de esta tesis doctoral se caracteriza por tener una serie de particularidades específicas y características propias en lo referente al diseño y desarrollo del método, así como a la implementación de sus diferentes procesos. No obstante, parte de estos procesos son comunes a las cuatro propuestas diseñadas. Algo similar, sucede en relación con los datos utilizados, puesto que tres de las cuatro propuestas presentadas comparten el mismo conjunto de datos. En esta sección pretendemos centrarnos en los aspectos comunes presentes en las diferentes propuestas elaboradas. Información mucho más detallada sobre sus particularidades y características específicas puede encontrarse en los respectivos apartados de Datos y Métodos de cada una de cuatro propuestas presentadas como publicaciones científicas e incluidas en la Parte II de esta tesis.

6.1. Materiales

La evolución y actualización de los datos de la base de datos Scopus y, por extensión, de la plataforma SJR, el amplio periodo de tiempo transcurrido entre los diferentes experimentos ejecutados y, sobre todo, los requerimientos propios de los diferentes procedimientos metodológicos desarrollados en esta tesis nos exigió manipular, al menos, dos conjuntos de datos distintos.

Para el primero de los experimentos, basado en el análisis de referencias, se recopilaron los datos de las publicaciones de un total de 17.158 revistas activas de SJR

para el periodo temporal comprendido entre 2003-2008 y se recopilaron sus citas agrupadas por revistas desde 2008 hasta 1996.

En los tres experimentos restantes, dos de ellos basados en el uso de diferentes algoritmos de *clustering* y un último sobre visualización de sus resultados, el número de revistas activas se incrementó hasta llegar a un total de 18.891. En los tres casos se creó una ventana temporal de dos años, correspondientes a los años 2009 y 2010, mientras que por medio de la ventana de citación diseñada se recopilaron las citas de sus publicaciones agrupadas por revistas desde el año 2010 hasta el 2000.

En relación con estos experimentos y los datos seleccionados para su realización, es importante dejar terminantemente claro que sus resultados reflejan el estado concreto de la base de datos en el momento de su realización. Las actualizaciones con carácter retrospectivo acometidas por Elsevier en la base de datos Scopus conllevan la introducción de nuevos documentos en el sistema, modificando las relaciones existentes entre ellos. Como consecuencia, la replicación de estos experimentos en un futuro podrían arrojar resultados no coincidentes en su totalidad con los obtenidos y expuestos en esta tesis (Moya-Anegón et al., 2013).

6.2. Métodos

Desde el punto de vista metodológico, las cuatro propuestas integradas en este trabajo presentan diferentes procedimientos semi-automáticos de clasificación. No obstante, a pesar de sus singularidades, las cuatro propuestas están interrelacionadas entre sí. En

primer lugar, todas y cada una de ellas persiguen un mismo objetivo, que no es otro que la mejora de la clasificación de la plataforma SJR. Pero aparte de compartir el mismo objetivo, cada propuesta posee también un conjunto de rasgos comunes, como puede apreciarse en la información que se presenta en la tabla 3, inspirada y adaptada de acuerdo al meta-modelo propuesto por Börner, Chen y Boyack (2005) para la producción de representaciones y mapas de dominio. La tabla resume de forma organizada ocho criterios o parámetros esenciales y necesarios para el diseño de las metodologías expuestas en las diferentes propuestas de este trabajo, concretamente:

1. *Fuente de información* utilizada para la extracción de datos.
2. *Unidad de Análisis* tomada como referencia para el proceso de clasificación, es decir, qué se pretende clasificar.
3. *Ventana Temporal* diseñada para la recogida de los datos.
4. *Unidad de Medida* utilizada como base en los diversos cálculos y operaciones cuantitativas necesarias.
5. *Ventana de Citación* configurada para seleccionar y considerar sólo aquellas citas comprendidas en un periodo de tiempo concreto.
6. *Medida de similaridad* (o disimilaridad) empleada.
7. *Técnica* utilizada para llevar a cabo el proceso de clasificación.
8. *Sistema de Etiquetado* adoptado para denominar los clústeres o grupos temáticos y facilitar la consiguiente adscripción de las revistas.

	Fuente	Unidad de Análisis	Ventana Temporal	Unidad de Medida	Ventana de Citación	Similaridad	Técnica	Sistema de Etiquetado
Propuesta 1	**Scopus/SJR**	**Revistas**	2003-2008	**Citación**	1996-2008	-	Análisis de referencias bibliográficas	**Categorías SJR**
Propuesta 2	**Scopus/SJR**	**Revistas**	2009-2010	**Citación** + Co-citación + Coupling	2000-2010	Geo normalización	Detección de comunidades: VOS y Louvain	**Categorías SJR** + Términos del título de las revistas
Propuesta 3	**Scopus/SJR**	**Revistas**	2009-2010	**Citación** + Co-citación + Coupling	2000-2010	Distancia (1-coseno)	*Clustering* jerárquico aglomerativo: Ward	**Categorías SJR** + Términos del título de las revistas
Propuesta 4	**Scopus/SJR**	**Revistas**	2009-2010	**Citación** + Co-citación + Coupling	2000-2010	Geo normalización	Detección de comunidades y Visualización: VOS	**Categorías SJR** + Términos del título de las revistas

Tabla 3: Parámetros utilizados en el diseño del material y método de las cuatro propuestas de clasificación desarrolladas

Si atendemos a los datos recogidos en la tabla, podemos observar en negrita ciertos parámetros que son compartidos simultáneamente por las cuatro propuestas de clasificación. El primero de ellos hace referencia a la fuente de información utilizada para recopilar los datos sobre los que posteriormente se ejecutarán los diferentes experimentos. La elección de este criterio está totalmente fundamentado por el propio objetivo marcado en esta tesis: la mejora de la clasificación de SJR. Es por ello, que la fuente de información seleccionada en las diferentes propuestas no puede ser otra que SJR, cuyos datos son importados de Scopus, almacenados y, finalmente, procesados por SCImago Lab[15], actual propietario de la plataforma bibliométrica.

Lo mismo sucede con la elección de las revistas como unidad de análisis, ya que son éstas las que representan el nivel de agregación más bajo que puede encontrarse dentro de la plataforma SJR. Las revistas constituyen entonces las unidades básicas sobre las que se calculan los distintos indicadores o, a partir de las cuales, pueden elaborarse los rankings. Su correcta clasificación es pues fundamental para la consistencia y fiabilidad de estos productos. De nuevo, el propio objeto de investigación de esta tesis ha resultado determinante en la elección de la unidad de análisis en las cuatro metodologías de clasificación planteadas.

Ahora bien, aun habiendo definido la revista como unidad de análisis, todos los conjuntos de datos utilizados en cada una las propuestas fueron recuperados en base a

[15] SCImago Lab (2014). *SCImago Lab: Research and Web Analytics*. Disponible en: http://www.scimagolab.com/ [Fecha de consulta: 02-09-2014].

los documentos. A continuación, los datos de dichos documentos fueron agregados conforme a sus revistas fuentes. Este hecho es especialmente importante en el procedimiento de cálculo de la citación directa y sus derivados (co-citación y coupling), al asegurar una precisión bastante alta en lo relativo a la correspondencia entre las referencias y los trabajos fuentes citados. De las tres unidades de medida escogidas en las cuatro propuestas, sólo la citación directa es común a todas ellas, convirtiéndose en la base para el cálculo de las restantes medidas y en un eje transversal o común a las diferentes propuestas. La co-citación y el coupling fueron utilizadas en los restantes experimentos en combinación con la citación directa y de acuerdo a la idea de Persson (2010) que denominó a esta medida combinada como *Weighted Direct Citation*. El esquema que Persson utilizó para interrelacionar y combinar estas tres medidas, facilita bastante su comprensión y su definición. La figura 3 representa las tres medidas derivadas de la citación siguiendo el esquema de Persson, pero añade una leve modificación que hace referencia a los dos sentidos en que puede ocurrir la citación directa por tratarse de una medida de relación asimétrica.

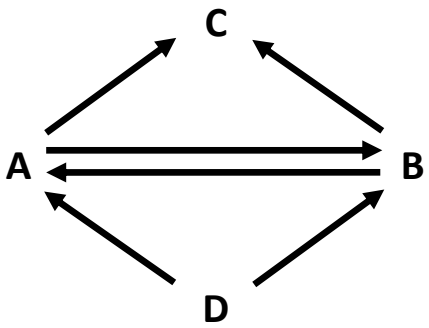


Figura 3: Weighted Direct Citation para la integración de las tres medidas de citación según Persson (2010)

De acuerdo a este esquema, y considerando las revistas como la unidad de análisis (tal y como ocurre en nuestras propuestas), tendríamos que:

- La *citación directa* (AB / BA) es un tipo de enlace o relación asimétrica entre dos revistas cualquiera. De esta forma **A** establece una relación directa con **B** que se expresaría mediante un valor o frecuencia que representa el total de veces que se produce esa conexión y simboliza su fuerza. Por otra parte, la relación opuesta establecida entre **B** y **A** es totalmente independiente de la anterior, pudiendo ser igual, distinta o, simplemente, no existir.
- La *co-citación* (DAB) es un tipo de relación simétrica que simboliza el número de veces que un par de revistas aparece citado conjuntamente. Esto quiere decir que, en la lista de referencias de la revista **D**, la asociación que se produce como resultado de la co-citación existente entre **A** y **B** es completamente igual a la que se produce entre **B** y **A**.
- El *coupling* o emparejamiento bibliográfico (ABC), representa igualmente un tipo de relación simétrica donde dos revistas (en nuestro caso **A** y **B)** citan a una tercera (**C**) de forma simultánea. En otras palabras, el coupling trata de identificar las referencias compartidas entre pares de revistas y, al igual que con la co-citación, el emparejamiento bibliográfico entre **A** y **B** es idéntico al de **B** y **A**.

Otro elemento común a las cuatro propuestas metodológicas es la exclusión de los enlaces de auto-citación al considerarse que este tipo de relaciones no resultan para nada relevantes en los procesos de clasificación. Igualmente, no se estableció ningún

tipo de filtro documental y las citas de cualquier tipo de documento incluido en una revista fueron tenidas en cuenta para el cálculo de las relaciones y su posterior agregación por revista.

Por último, otra característica transversal a las cuatro propuestas de clasificación abordadas tiene que ver con el sistema de etiquetado utilizado para clasificar las revistas. Así, al menos parte del proceso de etiquetado se realizó reutilizando el juego de etiquetas procedentes de las categorías temáticas del sistema de clasificación actual de SJR, con la excepción de etiquetas que hacen referencia a categorías difusas y excesivamente genéricas, como en el caso de las Misceláneas y Multidisciplinar, que fueron descartadas como descriptores temáticos. Utilizando estas etiquetas se procuró mantener la estructura central del esquema de clasificación actual de la plataforma SJR. Al mismo tiempo, se introdujeron numerosas modificaciones y actualizaciones por medio de un sistema complementario de etiquetas que, en el caso de las metodologías de clasificación basadas en el *clustering*, se configuró en base a la extracción de palabras significativas de los títulos de las revistas incluidas en cada clúster.

7. Resultados

En este apartado se exponen conjuntamente los resultados más relevantes generados a raíz de los diferentes experimentos sobre clasificación que conforman esta tesis. Para ello hemos tratado de elaborar una serie de indicadores y de gráficos aplicables a todas las soluciones de clasificación propuestas. De esta forma, tendremos una presentación resumida y organizada, por un lado, de los resultados obtenidos por cada solución así como su rendimiento y efectividad para la clasificación de revistas de forma individual. Al mismo tiempo, la aplicación de los mismos indicadores a todas las soluciones permitirá obtener una perspectiva global o visión de conjunto sobre el comportamiento y funcionamiento de las cuatro propuestas con respecto a la clasificación de revistas, así como su complementariedad, debilidades y fortalezas en comparación con el resto de métodos propuestos e implementados.

Los indicadores utilizados para la comparación de las diferentes propuestas han sido incluidos dentro de la tabla 4 y se describen a continuación:

1. *Total de revistas incluidas en el sistema*: Este indicador hace referencia al conjunto total de revistas SJR del que se parte originalmente y sobre el que se ejecutarán los diferentes procedimientos que integran cada método.
2. *Número de revistas clasificadas*: Representa el número de revistas clasificadas obtenidas una vez que se ejecutaron todos los procedimientos metodológicos que conformaban cada uno de los experimentos diseñados.
3. *Número de categorías*: Indica el número final de categorías útiles incluidas en el sistema. En los experimentos basados en algoritmos de *clustering*, la reutilización

de las etiquetas de las categorías de la plataforma SJR junto con la posibilidad de permitir la asignación múltiple de las revistas, tiene como consecuencia inmediata una diferencia entre el número de clústeres generados automáticamente y el número etiquetas de categorías utilizadas para su etiquetado.

4. *Media de revistas por categorías*: Expresa la ratio entre el número total de revistas del sistema y el número total de categorías existentes.
5. *Media de categorías por revistas*: Como resultado de la asignación múltiple de las revistas este indicador hace referencia al número medio de categorías que corresponde a las revistas del sistema.
6. *Porcentaje de solapamiento*: En conexión con el anterior indicador pero expresado en porcentajes, trata de recoger el solapamiento que se produce como consecuencia de la asignación múltiple de las revistas a las categorías. Para calcularlo es necesario conocer el número real de revistas incluidas en el sistema junto con el número total de asociaciones que se establecen entre esas revistas y las distintas categorías del sistema. En otras palabras, el indicador trata de reflejar el total de veces que *n* revistas aparecen asociadas a *n* categorías, incluyendo aquellas revistas repetidas como resultado de ser multi-asignadas. Su cálculo se basa en la siguiente fórmula: **B - A / A * 100**

   Donde ***B*** sería el total de revistas asignadas a las categorías del sistema (incluyendo las repeticiones derivadas de la asignación múltiple) y ***A*** representaría el número real de revistas incluidas en el mismo.
7. *Total de revistas que cambian su clasificación*: Indica el número de revistas que, en relación con la clasificación original de SJR, han sufrido modificaciones en alguna de las categorías asignadas tras someterse al nuevo proceso de clasificación. Estos

cambios sólo recogen alteraciones o cambios en la adscripción de categorías, no la pérdida o el añadido, que serán considerados aparte.

8. *Número de revistas que añaden categorías*: En relación con la clasificación original de SJR, apunta al número de revistas asignadas a una o más de una categoría nueva como consecuencia del proceso de clasificación ejecutado.
9. *Número de revistas que pierden categorías*: Tomando como referencia la clasificación original de SJR, expresa el número de revistas que, como resultado del proceso de clasificación, pierden su relación con una o varias categorías a las que anteriormente aparecían adscritas.

	SJR	Propuesta 1	Propuesta 2		Propuesta 3	Propuesta 4
	Clasificación Original	**Análisis de Referencias**	**Detección de Comunidades: Louvain**	**Detección de Comunidades: VOS**	***Clustering*: Ward**	**Detección de comunidades (VOS) y Visualización**
Total de revistas incluidas en el sistema	18.891	17.158	18.891	18.891	18.891	18.891
Número de revistas clasificadas	18.891	14.166	17.287	17.729	13.716	17.729
Número de categorías	308	198	272	267	298	267
Media de revistas por categorías	61,33	71,55	63,56	66,40	46,03	66,40
Media de categorías por revistas	1,61	2,06	1,48	1,50	1,42	1,50
Porcentaje de solapamiento	60,73	106,18	47,58	49,89	42,26	49,89
Total de revistas que cambian su clasificación	-	2.872 (20,27%)	5.784 (33,46%)	5.874 (33,13%)	1.988 (14,49%)	5.874 (33,13%)
Número de revistas que añaden categorías	-	7.249 (51,17%)	3.820 (22,10%)	4.192 (23,64%)	2.426 (17,69%)	4.192 (23,64%)
Número de revistas que pierden categorías	-	2.488 (17,56%)	4.540 (26,26%)	4.603 (25,96%)	3.951 (28,81%)	4.603 (25,96%)

Tabla 4: Indicadores aplicados a los diferentes sistemas de clasificación utilizados

Analizando detenidamente los resultados expuestos en la tabla 4 podemos observar que conforme al conjunto de indicadores aplicados, las propuestas con un rendimiento general más equilibrado son las dos basadas en los algoritmos de *clustering* de VOS y Louvain, también conocidos como algoritmos de d*etección de comunidades*. Así, por ejemplo, ambas propuestas son las que mantienen un número más alto de revistas clasificadas tras acometer los diferentes procedimientos metodológicos estipulados, con un total de 17.729 revistas (93,8%) en el caso de VOS y 17.287 (91,5%) en el de Louvain. En lo referente a número de revistas por categorías, aunque ambas propuestas incluyen un número menor de revistas finalmente clasificadas que la clasificación original de SJR, también es cierto que cuentan con un número de categorías más reducido. Aun así, mantienen una proporción cercana a la clasificación original de SJR y, además, ponen de manifiesto el valor añadido de estar caracterizadas por un solapamiento bastante menor que el que ocurre en la clasificación original de SJR.

La propuesta basada en el análisis de referencias, permitió reducir drásticamente el número de categorías en uso en el sistema, pasando de 308 en la clasificación original de SJR a tan sólo 198, lo que lo convierte en el sistema de clasificación con menor número de categorías en comparación con el resto. Esto influye a su vez en aspectos como el aumento de la media de revistas por categorías, lo que se traduce en mayores concentraciones de revistas en determinadas categorías del sistema. La proporción de revistas clasificadas con respecto al total de revistas al inicio del experimento no es tan alta como en el caso de los algoritmos para la detección de comunidades, pero asciende a un porcentaje bastante aceptable del 82.6% del total. Este último resultado

se ve fuertemente influenciado por la elección de ciertos parámetros condicionantes del estudio, como los umbrales de citación y publicación mínimos requeridos para cada revista incluida en el conjunto final considerado.

Por su parte, la propuesta basada en el *clustering* aglomerativo jerárquico de Ward acumula el peor resultado en cuanto a número total de revistas clasificadas, con tan sólo 13.716 revistas, lo que equivale a un 72,6% del conjunto inicial. Por el contrario, destaca positivamente su reducido porcentaje de solapamiento, el mejor en comparación con el resto, asegurando así menores concentraciones de revistas por categoría, como puede deducirse a partir del número medio de revistas por categoría que presenta el sistema. Este fenómeno puede estar también bastante influenciado por el hecho de ser la propuesta que cuenta con un mayor número de categorías en uso si excluimos del análisis la clasificación original de SJR. Concretamente, el número final de categorías para esta propuesta asciende a un total de 298.

En cuanto a los indicadores que reflejan algún tipo de alteración en la clasificación de las revistas, bien por medio de un intercambio de categoría, por la adición o por la pérdida de categorías, de nuevo las soluciones basadas en Louvain y VOS obtuvieron los valores más altos en cuanto al número de revistas que cambian sus categorías con respecto a SJR, con algo más del 33% del total. El valor más alto en cuanto al indicador de revistas que añaden categorías fue de lejos el alcanzado por la solución diseñada en base al análisis de referencia, con más de un 51% de revistas incrementando en una o varias categorías el número de categorías inicialmente asignadas en SJR. Por último, el indicador relacionado con la pérdida de categorías reveló que la solución con un mayor

porcentaje de revistas perdiendo una o más de una categoría en comparación con SJR fue la generada de acuerdo al *clustering* de Ward, con un total de 28,8% de revistas.

La figura 4, que representa las distribuciones completas de los diferentes conjuntos de revistas de cada sistema a lo largo de sus diferentes categorías, complementa los resultados de la tabla anterior y ayuda a obtener una visión todavía más precisa acerca del funcionamiento de cada una de las propuestas metodológicas puestas en práctica. Con objeto de facilitar su visualización se han representado las distribuciones basadas en el SJR original y en el análisis de referencias en el eje principal situado a la izquierda, mientras que las distribuciones referentes a Louvain, VOS y Ward, se han situado en el eje secundario de la derecha. Para evitar errores en su interpretación y en la evaluación de cada una de las distribuciones representadas, debe tenerse muy en cuenta que cada eje presenta una escala distinta.

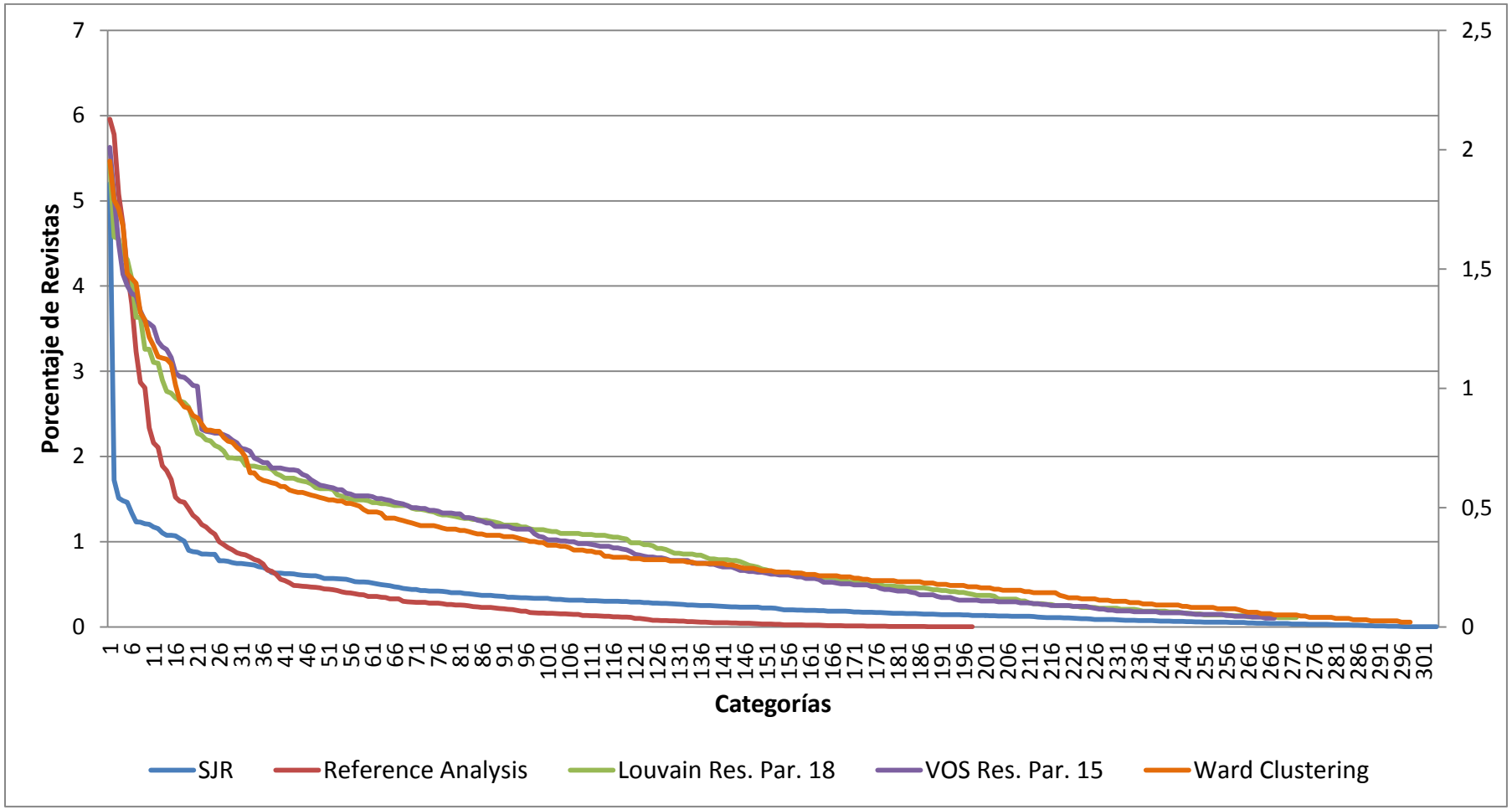


Figura 4: Distribuciones de revistas por categorías en los diferentes sistemas de clasificación analizados

A simple vista, la figura 4 se divide en dos partes bien diferenciadas como consecuencia de haber fijado dos escalas distintas para la visualización de las cinco distribuciones representadas, pero también como resultado de la propia similitud entre algunas de esas distribuciones. Así, en la parte inferior aparecen dos distribuciones con formas más o menos parecidas que representan las soluciones basadas en el SJR original y el análisis de referencias. Tal como puede apreciarse, se trata de soluciones con grandes concentraciones de revistas en las primeras categorías. En el caso de la distribución de SJR, esa concentración de revistas acontece prácticamente en la categoría de 'Medicine (miscellaneous)', mientras que en el análisis de referencias las concentraciones más altas de revistas aparecen aproximadamente durante las primeras 35 categorías de la distribución. A partir de ese momento, las distribuciones se equiparan entre sí, aunque la de SJR presenta un descenso incluso más suave y prolongado.

Por encima de esas dos distribuciones aparecen las tres distribuciones restantes, es decir, Louvain, VOS y Ward formando otro grupo bien diferenciado. Estas tres distribuciones presentan formas realmente similares, a veces fundiéndose casi en una sola línea y otras dibujando pequeñas fluctuaciones en ciertos segmentos de la distribución. A diferencia de las curvas trazadas en las distribuciones de SJR y análisis de referencias, estas tres soluciones aportan curvas más planas, con concentraciones mucho más moderadas de revistas por categorías, lo que se traduce, finalmente, en distribuciones más equilibradas y menos sesgadas con concentraciones de revistas más moderadas en las categorías que se sitúan en la parte inicial, es decir, a la izquierda de la curva de distribución.

Las cinco curvas de distribución representadas responden a un modelo denominado ley de potencias o *power law*. Este tipo de distribución es habitual en determinados tipos de rankings y a diferencia de la típica distribución gaussiana o normal presenta una cola muy prolongada en la parte derecha de la distribución. Así, atendiendo al tipo de y a la forma particular dibujada por la curva en cada una de las distribuciones, el sistema que aparentemente presenta una curva de más uniforme y menos sesgada es el sistema derivado de Ward. En gran medida, esto se debe a que de todos los sistemas representados, el originado por Ward es al mismo tiempo el que menor número de revistas engloba y el que mayor número de categorías en uso presenta. Esta particularidad, en principio, apunta a una influencia bastante considerable en lo que respecta a la forma final de la curva de distribución.

8. Discusión y Conclusiones

En esta sección se presentan de forma conjunta la discusión y las principales conclusiones que se derivan de los resultados generados en las diferentes propuestas metodológicas presentadas como artículos científicos en esta tesis. Las principales líneas de investigación de la tesis han sido ya introducidas dentro del capítulo de Objetivos Específicos. Para cada una de las preguntas de investigación formuladas en relación con los diferentes objetivos específicos proporcionamos ahora respuestas significativas y ajustadas, indicando también los artículos en los que se responde a dichas cuestiones. Tras la discusión pertinente y una vez contestadas todas las preguntas de investigación, procedemos a extraer y exponer las conclusiones más destacadas a tener en cuenta, sobre todo, para el desarrollo de futuros trabajos de investigación. El resto de la discusión y otras conclusiones más específicas se exponen detalladamente, junto con el resto de apartados correspondientes, en las distintas publicaciones agrupadas en la Parte II de esta tesis.

8.1. Análisis de referencias bibliográficas como técnica de clasificación de revistas

¿Se puede actualizar el esquema de clasificación de SJR y, al mismo tiempo, refinar o ajustar la adscripción de las revistas a las categorías de dicho esquema utilizando una metodología basada en el análisis iterativo de referencias bibliográficas de las revistas?

Sí. El análisis de referencias bibliográficas se mostró como un método apropiado para la actualización del esquema de clasificación y, en cierto modo, como un método

factible para mejorar la adscripción o asignación de las revistas a las categorías del sistema. Sin embargo, ciertos parámetros configurados en nuestros experimentos, como por ejemplo, los umbrales de citación mínimos establecidos, el número de iteraciones lanzadas, la granularidad deseada o, simplemente, la finalidad del propio sistema, pueden condicionar e influenciar en gran medida el resultado final de la clasificación generada.

En cuanto al esquema de clasificación, el procedimiento de análisis iterativo de referencias bibliográficas citadas planteado en el *artículo 1* provocó sensibles modificaciones tanto a nivel de áreas como de categorías. Su consecuencia más inmediata es una reducción del número de áreas y categorías temáticas con respecto al esquema de clasificación original de SJR. Así, el esquema de clasificación de SJR pasó de 27 áreas y 308 categorías a un total de 24 áreas y 198 categorías temáticas. En gran parte, este fenómeno se debió a las diferentes iteraciones del proceso y a la aplicación de umbrales en los porcentajes de referencias que definen las categorías de cada revista. El resultado visible de la aplicación de estos parámetros fue la reducción drástica del número de categorías ocasionado por la gran influencia y el mayor poder de atracción de algunas categorías dentro de la red de citación. Esta influencia o atracción puede deberse, entre otros factores, a los patrones y hábitos de citación específicos de ciertas disciplinas, un cierto sesgo en la cobertura temática de la base de datos, o un grado de consolidación más alto en determinadas disciplinas científicas.

La ejecución de múltiples iteraciones da lugar, por lo tanto, a un proceso continuado de acumulación de revistas en aquellas categorías caracterizadas por acumular una mayor varianza de la distribución. Estas categorías aumentan el número de referencias recibidas de las revistas en cada iteración, en favor del conjunto de categorías que acumulan una menor varianza en la distribución del número de referencias. Así, existe una correlación negativa entre el número de iteraciones y el número de categorías, de forma que un mayor número de iteraciones acaba generando un número de categorías cada vez más reducido. Por otra parte, la reducción de categorías favorece la aparición de concentraciones de revistas cada vez más altas en este grupo minoritario de categorías.

Parece obvio que el método puede funcionar como propuesta para la actualización del esquema de clasificación, en especial, en la reducción del número de áreas y categorías temáticas con menor peso e influencia en el esquema. No obstante, ciertas matizaciones son necesarias a la hora de evaluar positivamente su eficiencia en cuanto al refinamiento en la adscripción de las revistas a las categorías. La sucesiva reducción de categorías acontecidas a lo largo de las diferentes iteraciones favorecen, como ya hemos mencionado antes, la concentración de revistas en pocas categorías y, por lo tanto, este hecho no puede interpretarse como un refinamiento propiamente dicho (sí como un reajuste) en la clasificación de las revistas. Sin embargo, el funcionamiento y los resultados arrojados por el método parecen bastante apropiados para la actualización del sistema y el reajuste de las revistas a niveles de agregación más grandes, como es el caso de las áreas temáticas, lo que resulta también interesante

para el desarrollo de análisis bibliométricos con una mayor granularidad y un nivel de análisis más amplio. Otra opción razonable para tratar de refinar la adscripción de revistas a las categorías temáticas de SJR podría consistir en la ejecución no iterada del procedimiento. A la vista de los resultados de los experimentos, la ejecución del análisis de referencias una sola vez ayudaría a minimizar la aparición de las altas concentraciones de revistas. No obstante, esta decisión dependerá en gran medida de los objetivos definidos y de la finalidad de la clasificación generada.

8.2. Algoritmos de *clustering* sobre redes basadas en citación de revistas

¿Se puede clasificar la totalidad de las revistas de SJR mejorando su adscripción a las categorías y, al mismo tiempo, actualizar el esquema de clasificación de SJR a partir de otros métodos alternativos al análisis de referencias como, por ejemplo el clustering, utilizando combinaciones de medidas basadas en la citación? En cualquier caso, ¿existen marcadas diferencias entre las clasificaciones generadas por los diferentes algoritmos utilizados y en base a las dos formas adoptadas para la combinación de las medidas de citación?

Sí. No obstante, numerosos y variados matices se antojan necesarios para evitar posibles malinterpretaciones en esta afirmación. La primera parte de la pregunta de investigación formulada en el párrafo anterior tiene como base o punto de partida los resultados obtenidos en la primera propuesta metodológica descrita en el *artículo 1* y fundamentada en el uso del análisis de referencias. En dicha propuesta, del conjunto inicial de 17.158 revistas incluidas en el estudio, más de 2.700 quedaron sin clasificar

como consecuencia de los diferentes parámetros establecidos por el método. Es por esa razón por lo que nos planteamos si este inconveniente podría ser resuelto utilizando otros métodos alternativos, como por ejemplo, diferentes procedimientos de *clustering*.

Generalmente, cualquier algoritmo de *clustering* está capacitado por sí mismo para clasificar al completo el conjunto de entidades u objetos que se someten al procedimiento. Inicialmente, los tres algoritmos de *clustering* ejecutados tanto en los experimentos del *artículo 2* (algoritmos de detección de comunidades de Louvain y VOS) como en los experimentos recogidos en el *artículo 3* (algoritmo jerárquico aglomerativo de Ward) permitieron clasificar el conjunto total de revistas incluidas en el experimento. No obstante, tal como se ha podido comprobar en la tabla 4 situada en la sección de Resultados, los tres experimentos basados en *clustering* acabaron también excluyendo un elevado número de revistas del conjunto final clasificado. Este fenómeno, sin embargo, no debe atribuirse al funcionamiento de los algoritmos, sino más bien a la aplicación de otras consideraciones y criterios metodológicos diseñados para cada experimento, como por ejemplo, la fijación de un tamaño mínimo de clúster, la selección de la unidad de medida a utilizar y, en nuestro caso concreto, la propia combinación de las medidas empleadas en los experimentos.

Así, por ejemplo, en el *artículo 2*, los algoritmos de detección de comunidades de Louvain y VOS fueron capaces de clasificar todo el conjunto de revistas inicialmente dado. De acuerdo con la opinión de los creadores de VOS (Waltman, Van Eck, &

Noyons, 2010) pudimos verificar que introduciendo un valor adecuado y suficientemente grande en el parámetro de resolución del algoritmo, era posible identificar incluso aquellos clústeres de pequeño tamaño. Diversos experimentos previos demostraron que los clústeres que representaban agregaciones pequeñas de revistas, resultaban difíciles de etiquetar y de delimitar temáticamente.

Teniendo en cuenta nuestras pretensiones de mejorar la clasificación de SJR, muchos de los clústeres generados por estos algoritmos fueron descartados por no alcanzar el tamaño mínimo deseado y previamente establecido para representar de una forma precisa y consistente los campos o disciplinas de la investigación recogida en la literatura científica de las revistas de SJR. Concretamente, todos aquellos clústeres compuestos por un número inferior a 10 revistas fueron aislados del procedimiento de clasificación, lo que ascendió a un total de 578 clústeres en VOS y 784 clústeres en Louvain. La gran mayoría de estos clústeres estaban representados por unidades simples de una sola revista, los cuales se conocen como *singletons*.

De igual forma, el experimento basado en el método de *clustering* de Ward recogido en el *artículo 3*, se caracteriza por una pérdida significativa de revistas a lo largo del procedimiento metodológico de clasificación diseñado. En esta ocasión, las revistas aisladas del conjunto final clasificado tampoco responden a una consecuencia derivada del funcionamiento del algoritmo, sino principalmente, a la nueva forma de combinar las medidas basadas en la citación, la cual fuerza la aparición de las tres medidas al mismo tiempo por cada par de revistas incluidas en la matriz final (*hard combination*).

Esto provocó que del total de 18.891 revistas de partida, tan sólo 15.266 acabaran formando parte de la matriz final en la primera fase de *clustering* ejecutada. En esa primera fase de *clustering*, se obtuvieron además dos clústeres de gran tamaño que no se ajustaban a los objetivos de clasificación que perseguíamos. Este fenómeno suele ser habitual y característico en el funcionamiento del método de *clustering* de Ward. Para solucionarlo, se diseñó una segunda fase de *clustering* que se ejecutaría repitiendo todos y cada uno de los pasos desarrollados en la fase inicial, incluyendo la combinación de medidas entre las revistas de las dos sub-matrices de revistas extraídas de la matriz original y correspondientes a los dos grandes clústeres comentados. Ello condujo a la generación de nuevos grupos de revistas de menor tamaño y más específicos y, paralelamente, a una nueva reducción del conjunto total de revistas clasificadas, que quedó finalmente establecido en un total de 13.716.

En lo que se refiere a la actualización del esquema de clasificación de SJR, el sistema de etiquetado propuesto en los experimentos desarrollados en los *artículos 2* y *3* se erigió como un procedimiento exitoso y conveniente para este fin. El etiquetado de los clústeres derivados de cualquier procedimiento de *clustering* resulta uno de los aspectos más complejos y difíciles de abordar. Existen diversas propuestas basadas tanto en la utilización de texto (Janssens, Zhang, Moor, & Glänzel, 2009b; Zhang, Janssens, et al., 2010), el análisis de las asociaciones o conexiones (citación y derivados) entre los elementos de los diferentes clústeres (documentos, revistas…) (Glänzel & Thijs, 2011), la reutilización de las etiquetas de áreas o categorías temáticas

previamente definidas en un sistema de clasificación existente (Thijs, Zhang, & Glänzel, 2013) o el uso de varios procedimientos de forma paralela y combinada (Zhang, Liu, Janssens, Liang, & Glänzel, 2010) para resolver esta cuestión.

En los *artículos 2* y *3* utilizamos un método mixto basado en: (i) la reutilización de etiquetas de las categorías originales de SJR conforme a unos umbrales de citación ponderados con *tf-idf* y (ii) la extracción de términos significativos y relevantes de los títulos de las revistas que componen cada clúster. Este procedimiento permitió configurar esquemas de clasificación equilibrados y renovados que conservan una columna vertebral constituida por aquellas categorías más influyentes y estables dentro del sistema junto con otra serie de categorías nuevas obtenidas como resultado de la introducción del componente textual en el proceso de etiquetado. Ejemplos concretos de estas nuevas categorías pueden encontrarse en los experimentos desarrollados en estos dos artículos.

Por otra parte, nuestro proceso de etiquetado facilitó también la multi-asignación de las revistas a las categorías temáticas, aspecto importante por dos motivos fundamentales:

1. En primer lugar, si se tiene en cuenta la amplitud de la cobertura temática de la gran mayoría de revistas científicas, la asignación a múltiples categorías temáticas a la vez parece un enfoque mucho más realista y fiable que la asignación única.

2. De nuevo conviene dejar claro que el proceso de multi-asignación de las revistas es consecuencia directa del propio procedimiento de etiquetado utilizado en las propuestas metodológicas reflejadas en los *artículos 1*, *2* y *3*. Esta cuestión adquiere aún más importancia cuando se habla de los diversos algoritmos de *clustering* utilizados tanto en el *artículo 2* (Louvain, VOS) como en el *artículo 3* (Ward), puesto que todos ellos se denominan algoritmos *hard clustering*, lo que significa que proceden asignando cada revista a una única categoría.

Con respecto a las diferencias entre las clasificaciones, el análisis del número y el tamaño de los clústeres generados por los dos algoritmos de detección de comunidades a lo largo de las distintas soluciones obtenidas sirvió para revelar una diferencia esencial entre ambos métodos. En cada solución se introdujo un parámetro de resolución nuevo pero similar tanto para el método de Louvain como para el de VOS. Sin embargo, los resultados desvelaron que aun utilizando la misma parametrización el funcionamiento de ambos algoritmos era distinto. Concretamente, el método de Louvain produjo más clústeres de menor tamaño en cada una de las soluciones ejecutadas, lo que sugiere una granularidad más fina.

No obstante, al comparar sus resultados a la globalidad, los dos algoritmos mostraron la existencia de grandes similitudes entre las clasificaciones obtenidas por medio de Louvain y VOS. Podemos empezar, en primer lugar, señalando el gran parecido de ambas clasificaciones con respecto al número de categorías coincidentes en el ranking de revistas por categorías ofrecidas por ambas soluciones. En total, 14 de las primeras

20 categorías recogidas en este ranking son coincidentes en ambas clasificaciones (ver la tabla 2 del *artículo 2* en la Parte II de esta tesis). También conviene mencionar los indicadores referentes al 'Total de revistas que cambian su clasificación', 'Número de revistas que añaden categorías' y 'Número de revistas que pierden categorías' con respecto a la clasificación inicial de SJR, puesto que ofrecen porcentajes realmente parecidos en ambas clasificaciones. Esto, por lo tanto, denota un funcionamiento y un rendimiento bastante parecidos en el caso de los dos algoritmos de detección de comunidades utilizados en el *artículo 2*.

El mismo conjunto de indicadores más otros como el 'Número de revistas clasificadas', el 'Número de categorías', el 'Número medio de revistas por categorías', el 'Número medio de categorías por revistas' y el 'Porcentaje de solapamiento', son aplicados a las clasificaciones de SJR, Louvain, VOS y Ward en el *artículo 3* de esta tesis. La principal finalidad no era otra que obtener una visión general de la bondad y efectividad de las diferentes clasificaciones creadas. Si bien todas las clasificaciones mostraron algún tipo de mejora con respecto a la clasificación inicial de SJR, la clasificación obtenida por medio de la aplicación de Ward fue la que mejores resultados aportó en cuanto al solapamiento, en especial, como consecuencia de ser la clasificación que presentó los valores más altos en cuanto al número de revistas que perdieron categorías y, al mismo tiempo, las cifras más bajas con respecto a las revistas que añadieron nuevas categorías. Igualmente, la clasificación en base a Ward es la que presenta el número más pequeño de revistas que cambian de categoría, lo que puede interpretarse como

una clasificación que permanece bastante estable en comparación con la original de SJR.

Por otra parte, la aplicación del criterio que fuerza a la aparición de las tres medidas derivadas de la citación por cada par de revistas en el proceso de combinación tuvo como resultado un descenso del número de revistas finalmente clasificadas. Esto repercutió directamente en el valor del indicador que expresa la proporción final de revistas por categoría, que resulta sensiblemente más bajo en el sistema de clasificación basado en Ward (46,03) que en los sistemas de Louvain (63,56) o VOS (66,40). La clasificación obtenida conforme a Ward es también la que ofrece un número mayor de categorías nuevas generadas, un total de 139, en comparación con las restantes. Este hecho confirma la eficiencia y gran aportación del método no sólo de cara a la mejora de la adscripción de revistas, conseguida mediante el descenso del solapamiento y de las concentraciones de revistas por categorías, sino también de cara a la actualización del esquema utilizado para clasificar las revistas. Por el contrario, los procedimientos basados en Louvain y VOS aportaron un total de 83 y 77 nuevas categorías respectivamente. Recordemos también que el sistema de clasificación generado mediante el análisis de referencias y propuesto en el *artículo 1* destaca por la reducción del número de categorías temáticas y no por la creación de nuevas.

8.3. Evaluación de los resultados de la clasificación

¿Resulta la evaluación por comparación un mecanismo eficiente y suficiente para poder contrastar y validar los resultados obtenidos en las diferentes propuestas de clasificación elaboradas en este trabajo de investigación?

No. La evaluación por comparación resulta eficiente para obtener una visión general de la nueva clasificación obtenida a efectos de números de grupos o categorías temáticas, distribución de las revistas por categoría, coincidencias de revistas en grupos concretos, posibles solapamientos y diferencias en ciertas categorías, etc. No obstante, consideraciones más específicas como la bondad del sistema o la precisión de la adscripción de las revistas necesitan de otros sistemas complementarios, más costosos, completos y fiables, como la evaluación por medio de un panel de expertos.

En las tres primeras propuestas de clasificación expuestas en la Parte II de esta tesis se llevaron a cabo diferentes procedimientos para poder evaluar y contrastar las clasificaciones obtenidas como resultado de la aplicación de nuestras metodologías. Así, en los *artículos 1*, *2* y *3* las diferentes soluciones de clasificación generadas fueron comparadas con otros sistemas de clasificación para poder, cuanto menos, tener una visión general sobre el funcionamiento de los procedimientos metodológicos diseñados y ejecutados. La tabla 5 situada a continuación, resume las diferentes técnicas utilizadas para la generación de las nuevas clasificaciones en cada una de las cuatro propuestas junto con los sistemas de clasificación utilizados para realizar el

análisis comparativo y, finalmente, los indicadores o parámetros utilizados para establecer y centrar dicho análisis.

	Técnica	**Sistemas**	**Indicadores**
Propuesta 1	Análisis de Referencias	• SJR • Clasificación basada en Análisis de Referencias	• Categorized journals • Number of areas • Number of categories • Mean of categories per journal
Propuesta 2	Detección de Comunidades: • Louvain • VOS	• WoS • SJR • Clasificación basada en Louvain • Clasificación basada en VOS	• Total set of journals • Number of classified journals • Number of categories • Mean number of journals per category • Mean number of categories per journal • Overlapping percentage • Journals changing their classification • Journals adding categories • Journals losing categories
Propuesta 3	*Clustering* jerárquico aglomerativo de Ward	• SJR • Clasificación basada en Louvain • Clasificación basada en VOS • Clasificación basada en Ward	• Number of classified journals • Number of categories • Mean number of journals per category • Mean number of categories per journal • Overlapping percentage • Journals changing their classification • Journals adding categories • Journals losing categories
Propuesta 4	Visualización con VOS Viewer	• Clasificación basada en VOS	• Relaciones creadas entre las revistas (*intra-clúster*) y entre los grupos para corroborar los resultados del *clustering*

Tabla 5: Técnicas de clasificación empleadas en las diferentes propuestas y sistemas e indicadores utilizados para su análisis comparativo

La evaluación efectuada mediante la comparación de los distintos sistemas de clasificación utilizados ayudó a extraer algunas ideas generales, aunque de gran utilidad, para alcanzar unas primeras valoraciones sobre los resultados obtenidos. Así, en términos generales podemos enumerar:

- La existencia de una gran similitud entre las curvas definidas para la representación de la distribución de revistas por categorías, especialmente, entre las tres soluciones algorítmicas por un lado, y las soluciones basadas en una clasificación previa o existente, como en el caso de SJR y el análisis de referencias.
- De entre estos dos grupos, destaca la coherencia e incluso un cierto paralelismo entre los dos algoritmos de detección de comunidades de Luovain y VOS, no sólo en cuanto a la distribución de revistas por categorías, sino también en relación a la alta coincidencia en los rankings de categorías de ambas distribuciones y, por supuesto, en el conjunto de indicadores restantes aplicados de ambos sistemas, especialmente: 'Media de revistas por categorías', 'Media de categorías por revistas', 'Total de revistas que cambian su clasificación', 'Número de revistas que añaden categorías' y 'Número de revistas que pierden categorías'.
- En general, los sistemas de clasificación más equilibrados de acuerdo al conjunto total de indicadores para la evaluación son los de Louvain y VOS, mientras que las propuestas basadas en el análisis de referencias y en el *clustering* de Ward se mostraron como las más influyentes sobre el esquema de clasificación de SJR. La primera de ellas, redujo sensiblemente las categorías menos influyentes del sistema, mientras que la propuesta en base a Ward, fue la que más categorías nuevas introdujo. Esto, en gran medida fue también favorecido por procedimientos

paralelamente definidos en cada método, como el lanzamiento de varias iteraciones en el análisis de referencias o el modelo de etiquetado utilizado en la propuesta algorítmica basada en Ward.

A pesar del indiscutible valor de este tipo de análisis, pensamos que se necesitan otros métodos alternativos y complementarios a la propia comparación entre los sistemas de clasificación para contrastar los resultados de forma más profunda y ajustada. Para tal fin, se desarrolló una propuesta específica para el análisis y evaluación de los resultados del sistema de clasificación generado, a partir del algoritmo de *clustering* de VOS. El software VOSViewer (Eck & Waltman, 2010), que integra algoritmos para la implementación tanto del *clustering* por el método de VOS como para la posterior visualización de sus resultados, se convirtió en una herramienta de gran ayuda para el desarrollo y la consecución de esta tarea.

¿Resulta la visualización de información una herramienta eficiente y suficiente para poder contrastar y validar los resultados obtenidos en las diferentes propuestas de clasificación puestas en marcha en este trabajo de investigación?

Sí es eficiente, pero no del todo suficiente, tal como se puede comprobar en la última propuesta desarrollada como parte esta tesis y recogida en el *trabajo 4*, que sirvió para analizar y comprobar la efectividad de las técnicas de visualización de la información de cara a la validación y contrastación de los resultados de la clasificación. Dicho

trabajo, introduce una serie de grafos generados con el software de visualización VOSViewer que, como ya hemos mencionado arriba, permite representar gráficamente los resultados derivados de la aplicación del algoritmo de *clustering* de detección de comunidades de VOS. Los grafos o mapas generados por esta herramienta se asemejan a un tipo de MDS mejorado y permiten analizar las relaciones entre las revistas de forma que, a mayor cercanía entre los nodos o esferas (revistas) dibujadas, mayor similaridad o relación, mientras que a mayor tamaño de los nodos, mayor influencia o importancia dentro de la red. A partir de los mapas generados por el programa VOSViewer y analizando las relaciones entre las revistas y y sus agrupaciones, su interacción y su distribución en el espacio pudimos luego analizar, comparar y contrastar los resultados del *clustering*. Estos mapas deben interpretarse como representaciones lógicas y organizadas de la red de revistas SJR de acuerdo a la estructura de clústeres previamente creada.

En primer lugar, con ánimo de facilitar la visualización y la interpretación de los mapas, se realizó un análisis genérico de la estructura del mapa generado con ánimo de localizar agregaciones de revistas equiparables a las áreas temáticas de SJR. La mayor parte de estas agregaciones resultaron en grupos más o menos consistentes y bien definidos de revistas que presentaban bastante similitud con ciertas áreas de SJR, como Matemáticas, Física, Agricultura y Biología, Neurociencias, Psicología, Medicina, Ciencias Sociales, o Artes y Humanidades entre otras. El mapa completo, con forma parecida a un croissant, revela un gran parecido con trabajos anteriores de visualización de revistas, como el realizado por Leydesdorff y sus colaboradores (Loet

Leydesdorff, Moya-Anegón, & Guerrero-Bote, *in press*; Loet Leydesdorff, Rafols, & Chen, 2013). El mapa general presenta una alta consistencia no sólo en cuanto a las revistas que componen las diferentes agregaciones identificadas, sino también en la interrelación, el orden y la disposición final en el que aparecen dichas agregaciones.

Igualmente, un análisis pormenorizado y detallado de las revistas de áreas específicas como Matemáticas o Ciencias de la Información y la Documentación concentradas en los clústeres generados por el algoritmo de VOS, permitió comprobar la coherencia de las agregaciones de revistas establecidas y de sus relaciones tanto intra-clúster como inter-clúster. Así pues, la coherencia presente en la distribución de las revistas y de sus agregaciones, que presentan cierta correspondencia con las áreas temáticas de SJR anteriormente detectadas, confirmaron la eficiencia de la visualización de información como una técnica apropiada para contrastar y confirmar la validez del sistema de clasificación previamente generado por medio del *clustering*.

Este tipo de análisis más profundo puede llevarse a cabo mediante diferentes opciones del programa VOSViewer, especialmente, haciendo zoom en distintas partes del mapa, lo que permite visualizar más detenidamente las revistas que se localizan en una determinada zona de la representación gráfica. Una vez localizadas las revistas punteras de un área o categoría específica, a continuación, pueden realizarse búsquedas en los diferentes clústeres proporcionados por el programa para localizar aquellas otras revistas con las que se encuentran interrelacionadas. Esto permitirá

comprobar la consistencia y solidez de los clústeres creados en base a una relación temática evidente entre las revistas que los integran.

A pesar de presentarse como una herramienta de gran ayuda para corroborar los resultados de la clasificación, al igual que ocurre con la evaluación por comparación, consideramos que otros métodos alternativos además de las herramientas de visualización son necesarios para poder contrastar la fiabilidad de los grupos generados y la adscripción final de las revistas a las categorías. Como mencionamos anteriormente, la evaluación efectuada por medio de expertos parece una de las técnicas más efectivas, si bien es cierto que los costes económicos, de personal y de tiempo derivados de su implementación, dificultan enormemente su puesta en práctica.

Una vez analizados los resultados obtenidos, respondidas las preguntas de investigación planteadas y argumentados las ventajas, inconvenientes, fortalezas y debilidades de las diferentes propuestas metodológicas presentadas, estamos en condiciones de extraer las conclusiones más relevantes de este trabajo de investigación en su globalidad. Otras conclusiones más específicas pueden ser consultadas en los diferentes artículos que lo componen y que se localizan en la Parte II de esta tesis. Las conclusiones generales más importantes son:

- Las diferentes propuestas de clasificación presentadas aparecen como soluciones viables para mejorar la clasificación de SJR, tanto en su faceta de esquema de

organización y clasificación de la literatura, como en lo que respecta a la mejora y refinamiento en la asignación o adscripción de revistas en las categorías. No obstante, cada una de estas propuestas presenta también ciertas limitaciones, especialmente, la exclusión de ciertas revistas en el proceso de clasificación. Esto puede ser corregido con simples reajustes en los procedimientos metodológicos definidos, o bien, mediante el uso de una propuesta sistemática e integradora formada por los diferentes métodos utilizados en este trabajo. La solución final más apropiada apunta hacia una combinación de los diversos procedimientos utilizados siguiendo un protocolo de actuación previamente diseñado conforme a unas pautas bien definidas.

- Respecto a las revistas apartadas de los procesos de clasificación señaladas en el punto anterior, conviene señalar que la causa principal de este comportamiento no reside específicamente en el funcionamiento de las diversas técnicas automáticas empleadas (algoritmos de *clustering* o análisis de referencias), sino más bien en la definición de otros parámetros y condiciones que conforman parte de los procedimientos metodológicos diseñados, como por ejemplo, la combinación de medidas de citación (*clustering*), el umbral de referencias bibliográficas o artículos mínimo por revista (análisis de referencias), o el tamaño mínimo de clúster deseado para cada experimento llevado a cabo (*clustering*).
- La mezcla de medidas basadas en la citación parece influir positivamente en la clasificación final de las revistas, dando lugar a la aparición de agregaciones de revistas temáticamente consistentes. En el caso de los experimentos con algoritmos de *clustering*, esta combinación ayudó a maximizar el efecto clúster y a incrementar la consistencia de los grupos de revistas generados. Además, parece

bastante obvio que para los análisis estadísticos que incluyen la comparación entre casos, un mayor número de variables implica una mayor fortaleza en las relaciones establecidas. En los experimentos de *clustering*, donde se ha utilizado la triple combinación de medidas, la *soft combination* permitió obtener mayor exhaustividad en cuanto al número de revistas clasificadas, mientras que la *hard combination* aumentó la precisión temática de los grupos de revistas y el efecto clúster en favor del descenso del número final de revistas clasificadas.

- El sistema de etiquetado utilizado en el *clustering* permitió la mezcla de etiquetas de categorías temáticas de SJR y texto, aportando así estabilidad a las categorías más consolidadas del sistema y la integración de nuevas categorías ('Nanoscience and Nanotechnology', 'Social Works'…) derivadas del análisis de los términos significativos extraídos de los títulos de las revistas. Este sistema de etiquetado facilitó también la multi-asignación de revistas y, por consiguiente, el solapamiento de las categorías, comportamiento que, en el caso de los algoritmos de *clustering*, no es achacable a su funcionamiento, puesto que los tres métodos utilizados sólo permiten la asignación única de las revistas a las categorías.
- Los métodos de validación de las clasificaciones empleados resultaron válidos, sobre todo, para mostrar mejoras en lo relativo al solapamiento, concentraciones de revistas en las categorías, relaciones entre revistas y grupos de revistas, etc. No obstante, sería deseable el uso de técnicas más específicas para la comprobación del refinamiento y la mejora en la adscripción de las revistas a las categorías, como por ejemplo, una evaluación mediante un panel de expertos. El problema de este método de evaluación, como sucede en otros muchos procedimientos que se desarrollan en distintas facetas de la vida real, es la gran cantidad de recursos que

se necesitan para poder llevarlo a cabo. Una alternativa razonable a la evaluación de expertos puede ser la aplicación de medidas estadísticas específicas para comprobar la bondad y validez de los clústeres obtenidos, como por ejemplo, Rand Index, Silhouette, Modularity, Entropy, etc.

- En relación con muchas de estas decisiones, podemos también llegar a la conclusión de que la aplicación de procedimientos completamente automáticos de clasificación parece, por lo general, limitado por diferentes aspectos de la propia metodología que implican la intervención humana. Unas veces esta intervención supone únicamente la fijación de valores de ciertos parámetros, como por ejemplo, umbrales, puntos de corte o números de grupos temáticos que formarán parte del sistema de clasificación. Otras veces, por el contrario, la intervención implicó decisiones más complejas, como la reclasificación manual de revistas conforme a criterios heurísticos y empíricos, o la repetición o iteración de ciertos procedimientos con objeto de mejorar la clasificación final.

Antes de finalizar, conviene señalar una última cuestión de gran relevancia para el diseño, desarrollo e implementación de cualquier clasificación y que no es otra que el pragmatismo necesario y casi inherente asociado a esta actividad. Por lo general, cualquier proceso de clasificación está orientado a la consecución de unos propósitos y objetivos establecidos dentro de un marco o contexto específico. Parece complicado, por no decir casi imposible, elaborar un sistema de clasificación aislado de dicho contexto y, al mismo tiempo, no ceñido a estos objetivos previamente marcados. El fin último, por lo tanto, debe ser la efectividad y eficacia del sistema para facilitar la adscripción de las entidades u objetos clasificados en base a un equilibrio entre

exhaustividad y precisión. Esta cuestión, por lo tanto, parece poco factible sin la adopción de un enfoque pragmático (Glänzel & Schubert, 2003; Hampel, 2002; Jacob, 2004).

9. Perspectivas Futuras

El desarrollo de este trabajo de investigación ha servido para establecer diferentes propuestas metodológicas orientadas a mejorar la clasificación de la plataforma SJR y la adscripción temática de las revistas incluidas en ella. A pesar de alcanzar notables avances tanto en lo concerniente a la actualización del esquema de clasificación como en el ajuste y el refinamiento de la clasificación, un cierto margen de mejora de los procedimientos desarrollados es aún posible.

1. La primera de estas mejoras debería centrarse en el diseño, desarrollo e implementación de métodos alternativos para poder llevar a cabo la clasificación de las revistas excluidas de los diferentes procesos de clasificación expuestos en los *artículos 1*, *2* y *3*. Una solución razonable podría estar fundamentada en la creación de una nueva propuesta conjunta e integrada por los diferentes métodos utilizados en este trabajo. En función de la experiencia adquirida con la puesta en marcha de los distintos experimentos abordados en él, un posible protocolo a seguir sería el siguiente:
    a. En primer lugar, lanzar uno de los algoritmos de *clustering* para la detección de comunidades sobre la red de revistas integrando las tres medidas basadas en la citación (citación directa, co-citación y coupling) para obtener una primera clasificación fiable y consistente donde se incluya la mayor parte de las revistas abarcadas inicialmente.
    b. Las revistas excluidas en este primer proceso pueden ser clasificadas posteriormente a través de un procedimiento de análisis de sus referencias bibliográficas simple, es decir, no iterado.

c. En el caso de encontrar revistas que todavía quedasen fuera de ambos procesos de clasificación, podrían finalmente efectuarse procedimientos de clasificación basados en técnicas como:

   I. la relación de afinidad o familiaridad originalmente establecida con otras revistas, que puede resultar útil si se desea una solución efectiva con pocos recursos y requerimientos. De acuerdo a este sencillo método, las revistas excluidas de los procesos de clasificación se integrarían directamente en los grupos de revistas clasificadas atendiendo al vínculo inicial establecido por la compartición de una categoría temática concreta. En este procedimiento es inevitable partir de la siguiente premisa: las revistas que inicialmente compartían una categoría temática deben considerarse revistas "hermanas", lo que se traduce en la existencia de una afinidad temática entre ellas.

   II. Procedimientos manuales basados en el análisis detallado de información significativa para la descripción y delimitación del ámbito temático de las revistas, como por ejemplo, sus objetivos, su alcance, la audiencia a la que se destina o su adscripción temática en otros sistemas de información.

   III. Otros métodos más complejos pueden ser el análisis de redes y la visualización de información o bien, el uso de técnicas estadísticas previamente utilizadas con propósitos clasificatorios y con un rendimiento previamente contrastado en la tarea de clasificación de literatura, como el análisis factorial y de componentes principales.

2. Además de la posible aplicación de nuevas técnicas y métodos de clasificación, también resultaría interesante la adopción de nuevas medidas de relación o asociación entre las revistas, como por ejemplo, a partir del texto extraído de los documentos (análisis de co-words, frecuencias de aparición de términos, etc.), así como el diseño de propuestas híbridas basadas en texto y medidas citación de forma conjunta. Ya otra opción más compleja sería el diseño y desarrollo de una propuesta de clasificación viable orientada a los propios artículos. Esta clasificación podría establecerse luego como la base para una clasificación de las revistas fuentes utilizando criterios como los porcentajes de adscripción temática de los distintos artículos que las integran. Los últimos avances en la optimización de los algoritmos de *clustering* hacen pensar en la factibilidad e idoneidad de esta opción.
3. Respecto a la contrastación y validación de resultados, creemos necesario llevar a cabo una evaluación más profunda y detallada de las diversas clasificaciones obtenidas. Para dicho propósito, uno de los métodos más fiables es someter la clasificación final a la revisión de un panel de expertos con capacidad suficiente para valorar y contrastar los resultados obtenidos. No obstante, la gran cantidad de recursos necesarios para el desarrollo de este tipo de evaluación nos lleva a pensar en métodos de evaluación alternativos, como por ejemplo, el uso de índices y medidas estadísticas que sirvan como soporte para cuantificar la fortaleza, bondad y consistencia de los grupos obtenidos, es decir, la eficiencia de los sistemas de clasificación generados. Algunas de estas medidas, cuyo buen funcionamiento y eficacia han sido ya verificados en trabajos de clasificación de literatura previos (Janssens et al., 2009a; Liu et al., 2010) son, por ejemplo, *Normalized Mutual Information (NMI)*, *Rand Index*, *Silhouette*, *Modularity* o *Entropy*, entre otras.

10. Referencias Bibliográficas

Aksnes, D. W., Olsen, T. B., & Seglen, P. O. (2000). Validation of Bibliometric Indicators in the Field of Microbiology: A Norwegian Case Study. *Scientometrics*, *49*(1), 7–22. doi:10.1023/A:1005653006993

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011a). Towards a Multilingual, Comprehensive and Open Scientific Journal Ontology. In *E.C.M. Noyons, P. Ngulube, & J. Leta (Eds.), Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics* (pp. 66–77).

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011b). Towards a Multilingual, Comprehensive and Open Scientific Journal Ontology. In *E.C.M. Noyons, P. Ngulube, & J. Leta (Eds.), Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics* (pp. 66–77).

Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of scientific journals. *Scientometrics*, *44*(3), 323–345. doi:10.1007/BF02458483

Batagelj, V., & Mrvar, A. (1997). Pajek – Program for Large Network Analysis. Retrieved from http://pajek.imfm.si/doku.php

Bichteler, J., & Parsons, R. G. (1974). Document retrieval by means of an automatic classification algorithm for citations. *Information Storage and Retrieval*, *10*(7-8), 267–278. doi:10.1016/0020-0271(74)90022-9

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *October*(10), P10008. doi:10.1088/1742-5468/2008/10/P10008

Borko, H. (1962). The construction of an empirically based mathematically derived classification system. In *Proceedings of the May 1-3, 1962, spring joint computer conference on - AIEE-IRE '62 (Spring)* (pp. 279–289). New York, USA: ACM Press. doi:10.1145/1460833.1460865

Borko, H. (1964). Measuring the reliability of subject classification by men and machines. *American Documentation*, *15*(4), 268–273. doi:10.1002/asi.5090150405

Borko, H., & Bernick, M. (1963). Automatic Document Classification. *Journal of the ACM*, *10*(2), 151–162. doi:10.1145/321160.321165

Borko, H., & Bernick, M. (1964). Automatic Document Classification Part II . Additional Experiments. *Journal of the ACM*, *11*(2), 138–151. doi:10.1145/321217.321219

Börner, K., Chen, C., & Boyack, K. W. (2005a). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255. doi:10.1002/aris.1440370106

Börner, K., Chen, C., & Boyack, K. W. (2005b). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255. doi:10.1002/aris.1440370106

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., … Boyack, K. W. (2012). Design and update of a classification system: the UCSD map of science. *PloS ONE*, *7*(7), e39464. doi:10.1371/journal.pone.0039464

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. doi:10.1002/asi.21419

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Börner, K. (2011). Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PloS One*, *6*(3), e18029. doi:10.1371/journal.pone.0018029

Bruin, R. E., & Moed, H. F. (1993). Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications. *Scientometrics*, *26*(1), 65–80. doi:10.1007/BF02016793

Campos, A. (2004). Laplace: Ensayo filosófico sobre las probabilidades. *Revista Colombiana de Estadistica*, *27*(2), 153–177.

Cantos-Mateos, G., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Zulueta, M. A. (2012). Stem cell research: bibliometric analysis of main research areas through KeyWords Plus. *Aslib Proceedings*, *64*(6), 561–590. doi:10.1108/00012531211281698

Cañedo Andalia, R. (2001). Ciencia y tecnologia en la sociedad. Perspectiva hitórico-conceptual. *ACIMED*, *9*(1), 72–76.

Carpenter, M. P., & Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science*, *24*(6), 425–436. doi:10.1002/asi.4630240604

Chan, L. M. (1981). *Cataloging and Classification: An Introduction* (p. 397). New York: McGraw-Hill.

Cormack, R. M. (1971). A Review of Classification. *Journal of the Royal Statistical Society. Series A (General)*, *134*(3), 321–367.

Costas, R., & Bordons, M. (2008). Desarrollo de un filtro temático para la delimitación bibliométrica de un área interdisciplinar: el caso de Ciencias del Mar. *Revista Española de Documentación Científica*, *31*(2), 261–272.

Croft, W. B. (1980). A model of cluster searching based on classification. *Information Systems*, *5*(3), 189–195. doi:10.1016/0306-4379(80)90010-1

Cronin, B., & Atkins, H. B. (Eds.). (2000). *The Web of Knowledge: A Festschrift in honor of Eugene Garfield (ASIS&T Monograph Series)*. Medford, New Jersey: American Society for Information Science.

Crowson, A. R. (1970). *Classification and Biology* (p. 350). London: Heinemann Educational.

Doyle, L. B. (1965). Is Automatic Classification a Reasonable Application of Statistical Analysis of Text? *Journal of the ACM*, *12*(4), 473–489. doi:10.1145/321296.321298

Eck, N. J. Van, & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. doi:10.1007/s11192-009-0146-3

Elsevier. (2004). Scopus. Retrieved May 30, 2013, from http://www.scopus.com/home.url

Fingerman, S. (2006). Web of Science and Scopus: Current features and capabilities. *Issues in Science and Technology Librarianship*, *48*. doi:10.5062/F4G44N7B

Garland, K. (1983). An experiment in automatic hierarchical document classification. *Information Processing & Management*, *19*(3), 113–120. doi:10.1016/0306-4573(83)90064-X

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Glänzel, W., Schubert, A., & Czerwon, H.-J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics*, *44*(3), 427–439. doi:10.1007/BF02458488

Glänzel, W., Schubert, A., Schoepflin, U., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in journals covered by the SSCI database using reference analysis. *Scientometrics*, *46*(3), 431–441. doi:10.1007/BF02459602

Glänzel, W., Schubert, A., Schoepflin, U., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in journals covered by the SSCI database using reference analysis. *Scientometrics*, *46*(3), 431–441. doi:10.1007/BF02459602

Glänzel, W., & Thijs, B. (2011). Using “core documents” for detecting and labelling new emerging topics. *Scientometrics*, *91*(2), 399–416. doi:10.1007/s11192-011-0591-7

Gómez, I., Bordons, M., Fernández, M. T., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics*, *35*(2), 223–235. doi:10.1007/BF02018480

Gómez-Núñez, A. J., Vargas-Quesada, B., Moya-Anegón, F., & Glänzel, W. (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, *89*(3), 741–758. doi:10.1007/s11192-011-0485-8

Hampel, F. (2002). Some thoughts about classification. In K. Jajuga, A. Sokolowski, & H.-H. Bock (Eds.), *Classification, Clustering, and Data Analysis: Recent Advances and Applications* (p. 492). Berlin: Springer.

He, Q. (1999). *A review of clustering algorithms as applied to IR*. Tech. Rep. UIUCLIS-1999/6+IRG. Univ. Illinois at Urbana-Champaign.

Hjorland, B. (1992). The concept of “subject” in Information Science. *Journal of Documentation*, *48*(2), 172–200. doi:10.1108/eb026895

Hjorland, B. (1998). The classification of Psychology: a case study in the classification of a knowledge field. *Knowledge Organization*, *25*(4), 162–201.

Hjorland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content . . . and relevance. *Journal of the American Society for Information Science and Technology*, *52*(9), 774–778. doi:10.1002/asi.1131

Hjorland, B. (2003). Fundamentals of knowledge organization. In J. A. Frías & C. Travieso (Eds.), *Tendencias de investigación en organización del conocimiento = Trends in knowledge organization research* (pp. 83–116). Salamanca: Universidad de Salamanca.

Hjorland, B. (2008a). Core classification theory: a reply to Szostak. *Journal of Documentation*, *64*(3), 333–342. doi:10.1108/00220410810867560

Hjorland, B. (2008b). What is Knowledge Organization (KO)? *Knowledge Organization*, *35*(2/3), 86–101.

Hjorland, B. (2011). Knowledge Organization = Information Organization ? *Advances in Knowledge Organization*, *13*(January), 1–8.

Hjorland, B. (2012). Is classification necessary after Google? *Journal of Documentation*, *68*(3), 299–317. doi:10.1108/00220411211225557

Hjorland, B., & Pedersen, K. N. (2005). A substantive theory of classification for information retrieval. *Journal of Documentation*, *61*(5), 582–597. doi:10.1108/00220410510625804

Hodge, G. (2000). *Systems of knowledge organization for digital libraries: beyond traditional authority files*. Washington: The Digital Library Federation, Council on Library and Information Resources.

Iyer, H. (2012). *Classificatory structures : concepts, relations and representation*. Wurzburg: Ergon.

Jacob, E. K. (1991). Classification and categorization: Drawing the line. In *2nd ASIS SIG/CR Classification Research Workshop* (pp. 63–80). doi:10.7152/acro.v2i1.12548

Jacob, E. K. (2004). Classification and categorization: A difference that makes a difference. *Library Trends*, *52*(3), 515–540.

Jacsó, P. (2005a). As we may search – Comparison of major features of the Web of Science , Scopus , and Google Scholar citation-based and citation-enhanced databases. *Current Science*, *89*(9), 1537–1547.

Jacsó, P. (2005b). Comparison and Analysis of the Citedness Scores in Web of Science and Google Scholar. *Lecture Notes in Computer Science*, *3815*, 360–369. doi:10.1007/11599517_41

Jacsó, P. (2009). Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. *Online Information Review*, *33*(2), 376–385.

Jacsó, P. (2013). The need for end-user customization of the journal-sets of the subject categories in the Scimago Journal Ranking database for more appropriate league lists – a case study for the Library & Information Science field. *El Profesional de La Información (EPI)*, *22*(5), 459–473.

Janssens, F., Zhang, L., Moor, B. De, & Glänzel, W. (2009a). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683–702. doi:10.1016/j.ipm.2009.06.003

Janssens, F., Zhang, L., Moor, B. De, & Glänzel, W. (2009b). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683–702. doi:10.1016/j.ipm.2009.06.003

Kar, G., & White, L. J. (1978). A distance measure for automatic document classification by sequential analysis. *Information Processing & Management*, *14*(2), 57–69. doi:10.1016/0306-4573(78)90063-8

Kwok, K. L. (1975). The use of title and cited titles as document representation for automatic classification. *Information Processing & Management*, *11*(8-12), 201–206. doi:10.1016/0306-4573(75)90017-5

Kwok, K. L. (1984). A document-document similarity measure based on cited titles and probability theory, and its application to relevance feedback retrieval. In *SIGIR '84 Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval* (Vol. 36, pp. 221–231). British Computer Society Swinton. doi:10.1002/asi.4630360510

Kwok, K. L. (1985). A probabilistic theory of indexing and similarity measure based on cited and citing documents. *Journal of the American Society for Information Science*, *36*(5), 342–351. doi:10.1002/asi.4630360510

Lewes, G. H. (1893). *The Physical Basis of Mind* (p. 493). London: Kegan Paul, Trench , Trubner.

Lewison, G. (1996). The definition of biomedical research subfields with title keywords and application to the analysis of research outputs. *Research Evaluation*, *6*(1), 25–36. doi:10.1093/rev/6.1.25

Lewison, G. (1999). The definition and calibration of biomedical subfields. *Scientometrics*, *46*(3), 529–537. doi:10.1007/BF02459609

Lewison, G., & Paraje, G. (2004). The classification of biomedical journals by research level. *Scientometrics*, *60*(2), 145–157. doi:10.1023/B:SCIE.0000027677.79173.b8

Leydesdorff, L. (2002). Dynamic and evolutionary updates of classificatory schemes in scientific journal structures. *Journal of the American Society for Information Science and Technology*, *53*(12), 987–994. doi:10.1002/asi.10144

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in. *Journal of Documentation*, *60*(4), 371–427. doi:10.1108/00220410410548144

Leydesdorff, L., & Cozzens, S. E. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the SCI. *Scientometrics*, *26*(1), 135–156. doi:10.1007/BF02016797

Leydesdorff, L., Moya-Anegón, F., & Guerrero-Bote, V. P. (n.d.). Journal maps, interactive overlays, and the measurement of interdisciplinarity on the basis of Scopus data (1996–2012). *Journal of the Association for Information Science and Technology*. doi:10.1002/asi.23243

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362. doi:10.1002/asi.20967

Leydesdorff, L., Rafols, I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal citations. *Journal of the American Society for Information Science and Technology*, *64*(12), 2573–2586. doi:10.1002/asi.22946

Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology*, *61*(6), 1105–1119. doi:10.1002/asi.21312

López Cerezo, J. A., & Luján López, J. L. (1997). Los estudios de Ciencia-Tecnología-Sociedad (CTS) y la historia de la ciencia. In F. J. Rodríguez Alcázar & R. M. Medina Doménech (Eds.), *Ciencia, tecnología y sociedad: contribuciones para una cultura de la paz* (pp. 49–184). Granada: Universidad de Granada.

López-Illescas, C., Noyons, E. C. M., Visser, M. S., De Moya-Anegón, F., & Moed, H. F. (2009). Expansion of scientific journal categories using reference analysis: How can it be done and does it make a difference? *Scientometrics*, *79*(3), 473–490. doi:10.1007/s11192-007-1975-6

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, *1*(4), 309–317. doi:10.1147/rd.14.0309

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, *8*(3), 404–417. doi:10.1145/321075.321084

Mateos Sánchez, M., & Garcia-Figuerola Paniagua, C. (2009). *Aplicacion de técnicas de clustering en la recuperacion de informacion web* (p. 182). Gijón: Trea.

Meunier, J. G., Forest, D., & Biskri, I. (2005). Classification and Categorization in Computer Assisted Reading and Analysis of Texts. In H. Cohen & C. Lefebvre (Eds.), *Handbook on Categorization in Cognitive Science* (pp. 955–978). Amsterdam ; Boston: Elsivier.

Montoya Suárez, O. (2007). Aplicación del análisis factorial a la investigación de mercados. Caso de estudio. *Scientia et Technica*, *XIII*(35), 281–286.

Moravcsik, M. J. (1986). The classification of science and the science of classification. *Scientometrics*, *10*(3-4), 179–197. doi:10.1007/BF02026040

Moreno Rodríguez, R. M. (1997). Los estudios de Ciencia-Tecnología-Sociedad (CTS) y la historia de la ciencia. In F. J. Rodríguez Alcázar & R. M. Medina Doménech (Eds.), *Ciencia, tecnología y sociedad: contribuciones para una cultura de la paz* (pp. 149–184). Granada: Universidad de Granada.

Moya-Anegón, F. (dir), Chinchilla-Rodriguez, Z. (coord. ., Corera-Álvarez, E., González-Molina, A., López-Illescas, C., & Vargas-Quesada, B. (2013). *Indicadores*

bibliométricos de la actividad científica española: 2010 (p. 137). Madrid: FECYT. Retrieved from http://icono.fecyt.es/informesypublicaciones/Documents/indicadores bibliometricos_web.pdf

Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., & Herrero-Solana, V. (2007). Coverage analysis of Scopus: A journal metric approach. *Scientometrics*, *73*(1), 53–78. doi:10.1007/s11192-007-1681-4

Moya-anegón, F., Vargas-Quesada, B., Chinchilla-rodríguez, Z., Corera-álvarez, E., Munoz-fernández, F. J., & Herrero-solana, V. (2007). Visualizing the Marrow of Science, *58*(14), 2167–2179. doi:10.1002/asi

Munoz-Ecija, T., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Gómez-Nuñez, A. J., & Moya-Anegón, F. de. (2013). Nanoscience and Nanotechnology in Scopus: Journal Identification and Visualization. In *14th International Society of Scientometrics and Informetrics Conference, Vienna (Austria), 15th-19th July 2013* (p. 3).

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. London: Routledge & Kegan Paul.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity* (p. 338). New Jersey.

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. doi:10.1103/PhysRevE.69.026113

Olivera Betrán, J. (2011). Aproximación a una clasificación y categorización de las revistas científicas españolas de Ciencias de la Actividad Física y el Deporte. *Apunts Educación Física Y Deportes*, (105), 4–11. doi:10.5672/apunts.2014-0983.es.(2011/3).105.00

Persson, O. (2010a). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, *4*(3), 415–422. doi:10.1016/j.joi.2010.03.006

Persson, O. (2010b). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, *4*(3), 415–422. doi:10.1016/j.joi.2010.03.006

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, *12*(5), 297–312. doi:10.1016/0306-4573(76)90048-0

Price, N., & Schiminovich, S. (1968). A clustering experiment: First step towards a computer-generated classification scheme. *Information Storage and Retrieval*, *4*(3), 271–280. doi:10.1016/0020-0271(68)90006-5

Qu, B., Cong, G., Li, C., Sun, A., & Chen, H. (2012). An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology*, *63*(5), 889–903. doi:10.1002/asi.22611

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823–1835. doi:10.1002/asi.21086

Rak, R., Kurgan, L., & Reformat, M. (2007). Multilabel associative classification categorization of MEDLINE aticles into MeSH keywords. *IEEE Engineering in Medicine and Biology Magazine*, *26*(2), 47–55. doi:10.1109/MEMB.2007.335581

Roitblat, H. L., Kershaw, A., & Oot, P. (2010). Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, *61*(1), 70–80. doi:10.1002/asi.21233

Rossner, M., Van Epps, H., & Hill, E. (2007). Show me the data. *The Journal of Cell Biology*, *179*(6), 1091–2. doi:10.1083/jcb.200711140

San Segundo Manuel, R. (1996). *Sistemas de organización del conocimiento: la organización del conocimiento en las bibliotecas españolas*. Madrid: Universidad Carlos III. Retrieved from http://e-archivo.uc3m.es/bitstream/10016/4256/2/sansegundo_sistemas_1996.pdf

Schiminovich, S. (1971). Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm. *Information Storage and Retrieval*, *6*(6), 417–435. doi:10.1016/0020-0271(71)90008-8

SCImago. (2007). SCImago Journal & Country Rank (SJR). Retrieved April 15, 2011, from http://www.scimagojr.com/

Slavic, A. (2007). On the nature and typology of documentary classifications and their use in a networked environment. *El Profesional de La Informacion*, *16*(6), 580–589. doi:10.3145/epi.2007.nov.05

Small, H., & Griffith, B. C. (1974). The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Social Studies of Science*, *4*(1), 17–40. doi:10.1177/030631277400400102

Spasser, M. A. (1997). Mapping the terrain of pharmacy: Co-classification analysis of theInternational Pharmaceutical Abstracts database. *Scientometrics*, *39*(1), 77–97. doi:10.1007/BF02457431

Thijs, B., Zhang, L., & Glänzel, W. (2013). Bibliographic Coupling and Hierarchical Clustering for the validation and improvement of subject-classification schemes. In *14th International Conference on Scientometrics and Informetrics (15–19 July,*

2013), Vienna (Austria) (pp. 237–250). Viena: International Society of Scientometrics and Informetrics. Retrieved from http://www.mtakszi.hu/kszi_aktak/

Thomson Reuters. (2009). Web of Science. Retrieved September 01, 2013, from http://thomsonreuters.com/thomson-reuters-web-of-science/

Todorov, R. (1989a). Representing a scientific field: A bibliometric approach. *Scientometrics*, *15*(5-6), 593–605. doi:10.1007/BF02017072

Todorov, R. (1989b). Representing a scientific field: A bibliometric approach. *Scientometrics*, *15*(5-6), 593–605. doi:10.1007/BF02017072

Torres-Salinas, D., Bordons, M., Giménez-Toledo, E., Delgado-López-Cózar, E., Jiménez-Contreras, E., & Sanz-Casado, E. (2010). *Clasificación integrada de revistas científicas (CIRC)*: propuesta de categorización de las revistas en ciencias sociales y humanas. *El Profesional de La Informacion*, *19*(6), 675–684. doi:10.3145/epi.2010.nov.15

Van Cott, H. P., & Zavala, A. (1968). Extracting the basic structure of scientific literature. *American Documentation*, *19*(3), 247–262. doi:10.1002/asi.5090190307

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. doi:10.1007/s11192-009-0146-3

Wagner, C. S. (2005). Six case studies of international collaboration in science. *Scientometrics*, *62*(1), 3–26. doi:10.1007/s11192-005-0001-0

Waltman, L., Eck, N. J. Van, & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. doi:10.1016/j.joi.2010.07.002

Waltman, L., & van Eck, N. J. (2012). A New Methodology for Constructing a Publication-Level Classification System of Science, *63*(12), 2378–2392. doi:10.1002/asi

Weinberg, A. M. (1962). Criteria for scientific choice. *Minerva*, *1*(2), 158–171.

Zhang, L., Janssens, F., Liang, L., & Glänzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics*, *82*(3), 687–706. doi:10.1007/s11192-010-0180-1

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, *4*(2), 185–193. doi:10.1016/j.joi.2009.11.005

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, *42*(6), 1513–1531. doi:10.1016/j.ipm.2006.03.016

PARTE II: PUBLICACIONES CIENTÍFICAS

11. Listado de Publicaciones Científicas

I. Gómez-Núñez AJ, Vargas-Quesada B, Moya-Anegón F & Glänzel W (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics* 89 (3), 741-758.

II. Gómez-Núñez AJ, Batagelj V, Vargas-Quesada B, Moya-Anegón F & Chinchilla-Rodríguez Z (2014). Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics* 8 (2), 369-383.

III. Gómez-Núñez AJ, Vargas-Quesada B & Moya-Anegón F (*in press*). Updating the SCImago Journal & Country Rank classification: a new approach using Ward's clustering and alternative combination of citation measures. Aceptado para su publicación el 16 de junio de 2014 en la revista *Journal of the Association for Information Science and Technology*.

IV. Gómez-Núñez AJ, Vargas-Quesada B, Muñoz-Écija T & Moya-Anegón F (2013). Visualización y análisis de la estructura de la base de datos Scopus. En: Fernanda Ribeiro & Elisa Cerveira (coord.). *I Congresso ISKO Espanha e Portugal / XI Congresso ISKO Espanha. Porto, 7 a 9 de Novembro 2013*. ISBN 978-989-8648-10-5, pp. 264-275.

Pre-print of: ***Scientometrics*****, 2011, vol. 89 (3), p. 741-758.**

http://dx.doi.org/10.1007/s11192-011-0485-8

IMPROVING SCIMAGO JOURNAL & COUNTRY RANK (SJR)

SUBJECT CLASSIFICATION THROUGH REFERENCE ANALYSIS

Gómez-Núñez, Antonio Jesus [(1-5)]; Vargas-Quesada, Benjamín [(2-5)]; Moya-Anegón, Félix [(3-5)]; Glänzel, Wolfgang [(4)]

(1) CSIC – Institute of Public Goods and Policies (IPP), Madrid (Spain). E-mail: antoniojesus.gomez@cchs.csic.es

(2) University of Granada – Faculty of Communication and Documentation

(3) CSIC – Institute of Public Goods and Policies (IPP), Madrid (Spain)

(4) Centre for R&D Monitoring (ECOOM) and Dept. MSI, K.U. Leuven, Leuven (Belgium)

(5) SCImago Research Group Associated Unit, Granada (Spain)

Abstract: In order to re-categorize the SCImago Journal & Country Rank (SJR) journals based on Scopus, as well as improve the SJR subject classification scheme, an iterative process built upon reference analysis of citing journals was designed. The first step entailed construction of a matrix containing citing journals and cited categories obtained through the aggregation of cited journals. Assuming that the most representative categories in each journal would be represented by the highest citation values regarding categories, the matrix vectors were reduced using a threshold to discern and discard the weakest relations. The process was refined on the basis of different parameters of a heuristic nature, including 1) the development of *several tests applying different thresholds*, 2) the designation of a *cutoff*, 3) the *number of iterations* to execute, and 4) a *manual review* operation of a certain amount of multi-categorized journals. Despite certain shortcomings related with journal classification, the method showed a solid performance in grouping journals at a level higher than categories — that is, aggregating journals into subject areas. It also enabled us to redesign the SJR classification scheme, providing for a more cohesive one that covers a good proportion of re-categorized journals.

Keywords: Reference Analysis, Journal Classification, Subject Categorization, Multidisciplinary Databases, SCImago Journal & Country Rank

Introduction

Problems related to the classification of scientific knowledge have been widely discussed by scholars and researchers from different disciplines throughout history. In the limelight of the Library and Information Science field stand contributions by figures such as Dewey, Otlet, Ranganathan or Hjørland, for instance. According to Glänzel (Glänzel & Schubert 2003) "classification of science into a disciplinary structure is at least as old as science itself". However, the facet of human knowledge that Chen described as a "complex and dynamic network" (Chen 2008) likewise complicates the development of any reliable and representative disciplinary classification scheme that might allow to effectively delimit different subjects or disciplines configuring this highly complex network.

Knowledge takes place as a result of the curiosity and interest of human beings focused on explaining the surrounding environment and the phenomena taking place. To this end, it is necessary to conduct research-related processes that may be considered inherent to human being and essential to reach knowledge. Nowadays, research is influenced by factors such as its strong relationship with society, the ultra-specialization of areas and disciplines of knowledge, the competitiveness exercised by increasing practitioners, groups and research institutions, or the dynamism resulting from new trends and fashions. All this gives rise to a considerable growth of literature, as well as a constant restructuring and redefinition of the areas and disciplines of scientific knowledge, which ultimately interferes with our ability to design and implement classification systems that represent scientific knowledge.

Databases, regarded as great repositories responsible for storing the results of scientific research, require the use of efficient classification schemes. This is a vital need not only when searching and retrieving information, but also for preparing cohesive and reliable bibliometric analyses. Isabel Gomez (Gómez et al. 1996), in the context of growing interdisciplinary research, underlined the meteoric changes involving disciplines and journals (titles changing, journals merging, etc.), or the establishment of classification systems targeted to the specific interests of each particular database as common problems regarding the organization of recorded scientific knowledge. Although a number of models can be adopted to classify the contents of the various scientific databases, two major multidisciplinary databases —the ISI Web of Science (WOS) (Thomson Reuters 2010) and Scopus (Elsevier 2002)– both opted to use a similar model of classification that relies upon one hierarchical scheme encompassing a number of areas (first level) and subject categories (second level). All the source journals collected by these databases are placed in one or more area and category on the basis of criteria such as title, scope or citation patterns. Thus, in contrast to databases with a more specialized coverage, such as Medline or INSPEC, where papers are directly assigned to categories, under the WOS or Scopus classification model, journals are classified into categories, while the papers covered by them are assigned to source categories through indirect assignation.

Both WOS and Scopus have become key tools for the development of bibliometric surveys which, in the face of science evaluation, aid decision-making on the part of scientific administrators and politicians concerned with funding and the efficient assignment of resources. As far back as 1963, Weinberg (Weinberg 1963) claimed that the extensive growth of science required more resources in a society of limited resources, meaning it was necessary to choose among different areas or fields of science (*scientific choice*) and between the various institutions receiving government assistance (*institutional choice*). He therefore put forth a number of useful criteria for prioritizing when selecting, divided into *internally generated within the scientific field* itself, and *externally generated out of the field*, which included aspects such as technological, scientific and social merit. It is clear that bibliometric and scientometric analysis as developed from data covered by major scientific databases must be considered essential instruments in the evaluation process and in selecting the best ones within a system mainly based on merit. However, in order to ensure that surveys have high credibility and precision, it is necessary to define and delimit in a reliable manner each one of the subject fields and subfields of knowledge generated through research. For this reason the design of a flexible and adaptable classification model for categorizing scientific literature is held to be an essential matter.

Review

This work introduces a proposal to improve the categorization of Scopus database journals included at the SCImago Journal and Country Rank (SJR) portal (SCImago Lab 2007) using journal reference analysis, one of the many techniques applied in the vast arena of scientific literature for the classification, categorization and delimitation of subject fields. Narin (Narin 1976) was a pioneer in proposing that papers be classified by allocating them to the category of journals to which they belonged. He held that citation recount was useful not only for bibliometric purposes, but also for the classification of publications. In earlier work (Narin, Carpenter, & Nancy 1972), using references and citation analysis, different graphic models were developed to represent the relations established between a set of journals and the disciplines they pertained to. In papers published with Pinski (Narin, Pinski, & Gee 1976) (Pinski & Narin 1976), he used the analysis of bibliographic references to aggregate journals into different groups and subject categories.

Glänzel employed reference analysis to develop an item-by-item classification model applicable at the level of items (papers) rather than at the source level. Firstly, he analyzed paper reference lists of the Science Citation Index (SCI) *multidisciplinary* and *general journals* (Glänzel, Schubert, & Czerwon 1999). He then applied a methodology similar to that used with papers published in journals covered by the *Social Science Citation Index* (SSCI) (Glänzel and others 1999). Finally, in a further contribution, Glänzel (Glänzel & Schubert 2003) devised a new classification scheme applicable to all areas of scientific knowledge (science, social sciences, and arts & humanities) with scientometric evaluation purposes. The three-step building process included, at step 3, the classification of papers appearing in journals with

ambiguous or poorly defined categories (i.e. multidisciplinary), on the basis of reference analysis.

Searching for a way to upgrade and restructure classification of journals, Leydesdorff (Leydesdorff 2002) developed a proposal to define shifts in the classification schemes of databases due to the inclusion of new journals or modifications (merges or title changes) affecting them. This proposal focused on transactions and relationships among journals involving citation. Its main goal was to define a posteriori the changes in categories represented by different sets of journals, giving rise to a dynamic and evolutionary update of the classification schemes used in databases.

More recently, in order to define and delimit the ISI categories of *Oncology* and *Cardiac & Cardiovascular System*, Lopez-Illescas (López-Illescas and others 2009) put forth an approach combining_the use of WOS specialized journal categories together with reference analysis. Under this approach, it is assumed that scientific journal articles are well categorized within a gjven subfield of the source journal. Therefore, a subfield could be properly delimited by a group of papers from specialized journals in a particular subfield (*subfield's specialist journals*) and another group of papers belonging to non-specialist journals (*additional journals*) that cite journal papers from a previously established citation threshold.

Improving and updating the categorization of Scopus database journals included in the SCImago Journal and Country Rank (SIR) website calls for some reallocation and delimitation of subject areas and categories in order to restructure the scientific knowledge encompassed by SJR journals. It is thus intended to represent a consistent and congruent new disciplinary structure founded upon a set of well defined subject categories. Once the new classification scheme is defined, it is necessary to re-categorize journals, assigning the subject categories considered under the new scheme. This process largely entails reference analysis. In the case of journals with an insufficient number of bibliographic references, e.g. social science or arts & humanities journals, it will be necessary to tackle other methodological procedures in order to categorize them.

The final goal of our proposal is therefore to redefine the subject areas and categories of SJR journals through reference analysis. Narin (Narin 1976) established the importance of citations among papers to define the structure of scientific literature. At a macro level, he found citation analysis useful for representing and relating areas and subject categories by mapping journals. By further exploring this idea, we intend to more soundly define categories or disciplines representing the knowledge covered by the scientific literature of SJR journals.

Methodology

The SJR two-level hierarchical classification scheme, consisting of 27 areas and 308 subject categories, was used in this study. It had been previously defined by SCImago group members on an empirical basis, taking into account characteristic and discriminative journal features such as title or scope, and expert opinions. The starting point was the editors' journal categorization based on their scope statements. Many authors hold a priori classification schemes developed by experts (relying on their scholarship, knowledge and experience in specific fields) to be useful not only for information retrieval, but also for bibliometric and scientometric purposes. Glänzel (Glänzel & Schubert 2003) judged this proposal as sensible and pragmatic. Earlier on, Schubert (Schubert & Braun 1996) affirmed that in reference standardization processes, essential for the development of scientometric indicators and their subsequent comparison, "comparative assessments based on prior classification schemes are usually easier to comprehend and accept".

We then submitted a query to retrieve Scopus data from SJR in order to derive a neighbor list containing citing-journals, cited-journals and values representing the relationships established among them. The data set covered a 6-year period, from 2003 to 2008, with references going back as far as 1996 (to 2008), and including a total of 17158 journals. For this process, journal self-citation values were discarded. By using this list, an asymmetric journal-category citation matrix was constructed whose values display the amount of citing-journal references linking to SJR categories, reached via aggregation of the cited-journal categories. Therefore, improved final categorization of journals was achieved on the basis of SJR categorization (previously assigned) of cited-journals. The relationship values established among journals and categories were later transformed into percentages. Finally, categories labeled as Miscellaneous and Multidisciplinary were removed from the analysis.

Observation of journal-category vectors derived from the journal-category matrix, revealed the existence of a large amount of residual values in each one. These values reflect weaker relations established between the journals and certain categories (Figure 1). We assumed that, for each journal vector, the most representative categories were reflected by higher percentage values. So as to avoid the weakest links representing the categories with less influence on journal topic, a threshold was established. This allowed us to transform original vectors by keeping only values or aggregate values (cumulative sum) equal to or higher than the threshold defined, while values below it were isolated (Figure 2). Thus, the method works by stressing the generality of journals in order to define their definitive categories.

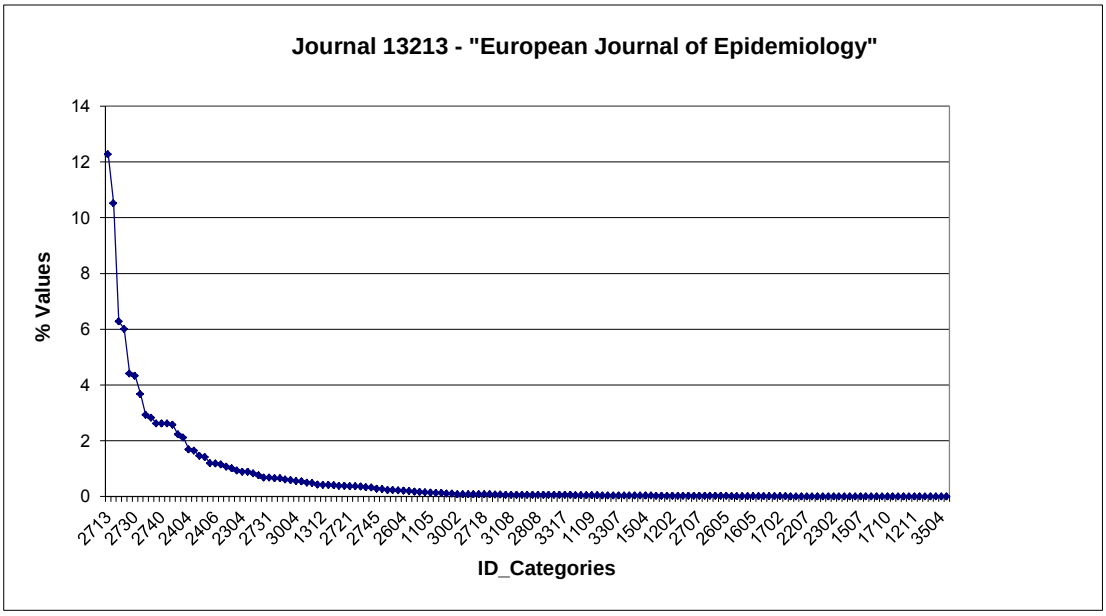


Figure 1: Distribution of Categories of Journal 13213 after Iteration 1

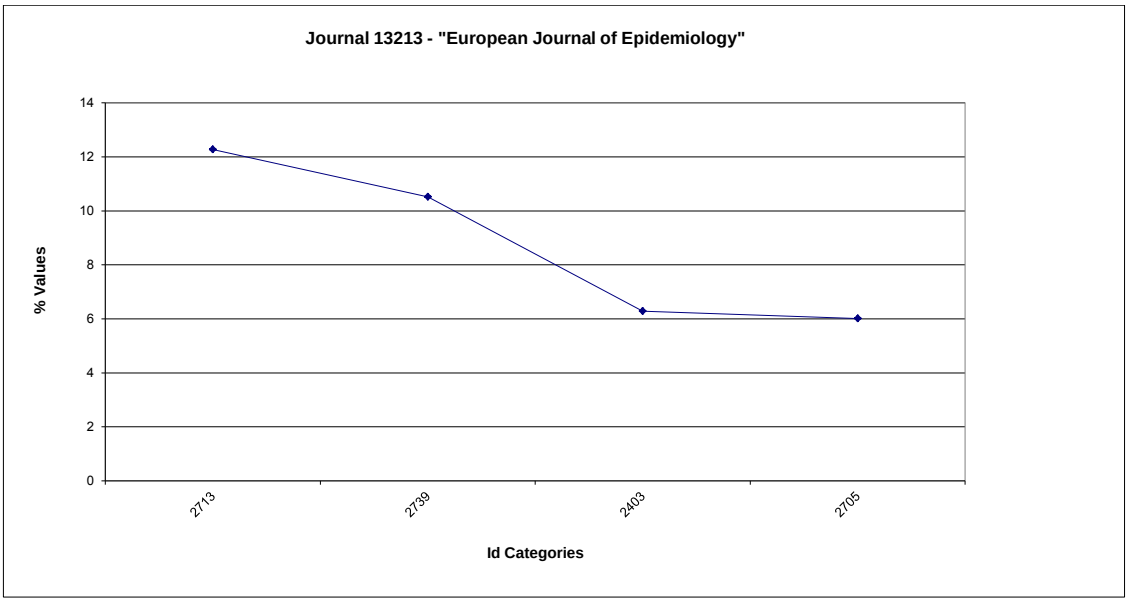


Figure 2: Distribution of Categories of Journal 13213 after Iteration 1 and Threshold 31%

As the next step, we ran an iterative categorization process during *n times*, so that classification of the journals constituting the citation network would be enhanced and gain in pertinence. That is, as the process advanced, relations among journals and categories would be seen to change; while journal-journal links remained identical, cited journal categories were altered through iterations by means of a feedback process. It was therefore necessary to clarify some important questions for the method‘s performance, such as: 1) *number of*

iterations to execute; 2) *cutoff* or *stopping point* in order to reach a well-delimited and consistent subject categories scheme; and 3) *the threshold value* to apply to vectors.

On the one hand, through an heuristic approach based on observation of changes produced at distribution of categories per iteration —i.e., the number of categories keeping cited by journals after each replication (Figure 3)— we concluded that a total of 12 iterations was enough to obtain an overall view of process performance. On the other hand, it was noted that cutoff depends largely upon the threshold value established. To select the more appropriate threshold, we resorted to several empirical tests using values from 25% to 60%. These tests evidenced that best threshold was 31%. Thereby, once vectors were optimized, the matrix was reduced to approximately a 1/3 share of their values, retaining only the strongest relations between journals and categories. Then, by analyzing certain indicators achieved after adopting 31% threshold, iteration 2 was determined to be the best point to halt the process. At that point, changes resulting in journal categorization were the most balanced according to relevant indicators such as *Mean categories per journal* or *Number of cited categories per iteration*. Table 1, collecting these indicators, makes manifest that the steepest drop in the distribution of categories per iteration was between 1 and 3. Not only did this reinforced iteration 2 serve as an excellent cutoff, but also, journals with a high number of assigned categories could thereby be avoided. Distribution of journals with *n categories* can be seen in Figure 4.

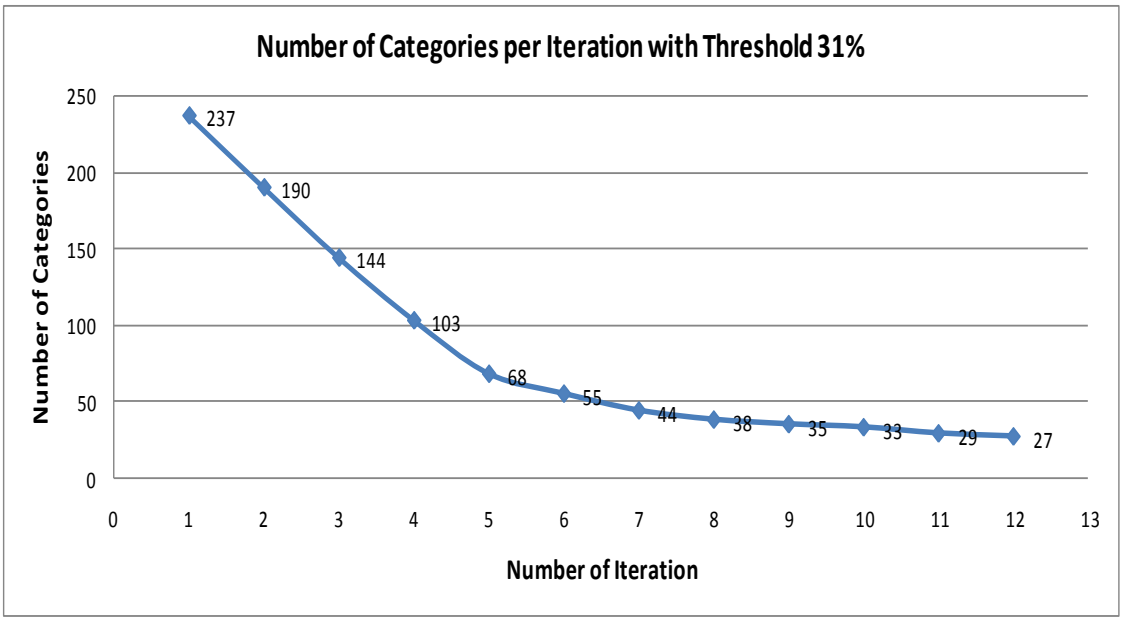


Figure 3: Distribution of Categories over Iterations using a Threshold of 31%

	Journals Categorized	Journals Not Categorized	Number of Records per Table	Mean of Categories per Journal	Num. of Cited Categories	Slope	Slope Percentage	Num. of Non-cited Categories
iteration 1	15584	1574	34046	2.19	237			66
iteration 2	**15595**	**1563**	**32317**	**2.07**	**190**	**47**	**22.4**	**113**
iteration 3	15595	1563	25606	1.64	144	46	21.9	159
iteration 4	15595	1563	20277	1.30	103	41	19.5	200
iteration 5	15595	1563	17514	1.12	68	35	16.7	235
iteration 6	15595	1563	16560	1.06	55	13	6.2	248
iteration 7	15595	1563	16209	1.04	44	11	5.2	259
iteration 8	15595	1563	16089	1.03	38	6	2.9	265
iteration 9	15595	1563	15982	1.03	35	3	1.4	268
iteration 10	15595	1563	15872	1.02	33	2	1.0	270
iteration 11	15595	1563	15815	1.01	29	4	1.9	274
iteration 12	15595	1563	15769	1.01	27	2	1.0	276

Table 1: Indicators obtained for a Threshold of 31%

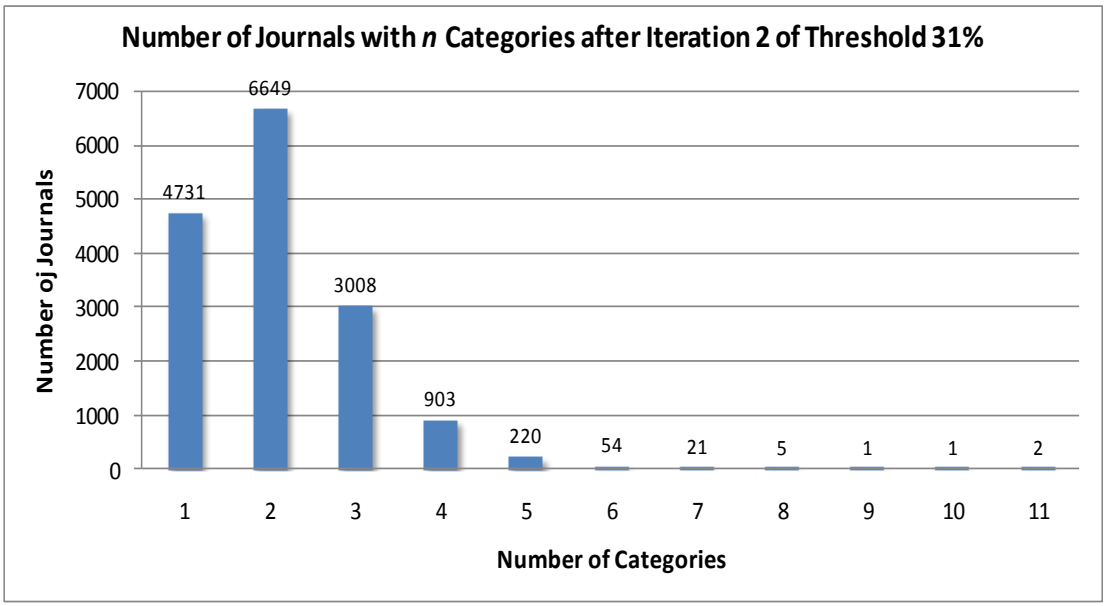


Figure 4: Distribution of Journals with *n* Categories after Iteration 2 and Threshold 31%

Having resolved the threshold, cutoff and number iterations to be applied, we considered that, in order to establish a stronger journal categorization and a more robust scheme, it was necessary that the categorized journals would satisfy the condition of having at least 30 items and 30 references pointing to database sources. Of course, some journals did not reach these values, and so the number of journals categorized decreased around 1500; but at the same time this ensured a good ratio of journals categorized and a well delimited set of subject categories.

The last stage of the method involved the application of a manual review process of journals with more than 4 assigned categories so as to obtain a finer categorization and to prevent an extensive number of multi-categorized journals. This was done following a set of well defined rules, making it possible to readjust the final categorization of around 300 journals and allowing us to discard categories with low rates, to determine a maximum of 5 categories per journal, and to aggregate categories into homogenous sets of the same subject area (*miscellaneous categories*), or into heterogeneous ones of different areas (*multidisciplinary category*). By doing so, 8 new categories (*7 miscellaneous* and *1 multidisciplinary*) were obtained and added to the final set of categories forming part of the renewed classification scheme. Additionally, the multidisciplinary category was included in one new *multidisciplinary area*.

Results and Discussion

Table 2 collects some indicators related to different moments in the development and evolution of this categorization process, from SJR initial categorization to the final classification scheme obtained after applying thresholds, iterations and reviews. Comparison of the original SJR classification scheme and new one revealed various perceptible changes implying new aggregations of journals into different categories than before, and, most importantly, the disappearance of some categories that are not used by journals —in other words, categories that are not being linked by the current journal references. To a lesser degree, this affects areas as well, whose removal is fully tied to a total disuse of categories included into them. The total number of changes in journal categorization was over 12000, involving the addition, loss or new ranking (based on percentages) of categories assigned to journals.

	SJR	Threshold 31% Iteration 1	Threshold 31% Iteration 2	Threshold 31% Iteration 2 Papers & Refs. to DB Sources >= 30	Threshold 31% Iteration 2 Review of Journals >4 Categories Assigned
Categorized Journals	17158	15584	15595	14166	14166
Number of Areas	27	25	23	23	24
Number of Categories	308	237	190	186	198
Mean of Categories per Journal	1.54	2.19	2.07	2.11	2.06

Table 2: Differences among SJR original categorization and new one

The final categorization scheme (giving rise to a total of 14166 categorized journals) including subject areas, subject categories covered by them, journals citing each category, and their corresponding percentages over total journals categorized, can be found in the *Appendix* section of this paper. To calculate percentages, the overlap due to multi-categorization of journals was studied. At a higher level, Table 3 captures the final distribution of journals per area under the new SJR categorization scheme, collecting the number of journals covered by areas, percentages of this ratio, and the number of categories included in every given area together with their respective percentages.

AREA	Journals Covered per Areas	Percentage of Journals	Categories per Area	Percentage of Categories
Multidisciplinary	28	0.10	1	0.51
Agricultural and Biological Sciences	1543	5.28	11	5.56
Arts and Humanities	586	2.01	8	4.04
Biochemistry, Genetics and Molecular Biology	5213	17.85	13	6.57
Business, Management and Accounting	792	2.71	9	4.55
Chemical Engineering	232	0.79	6	3.03
Chemistry	938	3.21	7	3.54
Computer Science	629	2.15	12	6.06

Decision Sciences	110	0.38	1	0.51
Earth and Planetary Sciences	1056	3.62	11	5.56
Economics, Econometrics and Finance	840	2.88	2	1.01
Energy	172	0.59	5	2.53
Engineering	1968	6.74	13	6.57
Environmental Science	1169	4.00	8	4.04
Immunology and Microbirology	2023	6.93	5	2.53
Materials Science	619	2.12	8	4.04
Mathematics	739	2.53	9	4.55
Medicine	6931	23.73	35	17.68
Neuroscience	107	0.37	2	1.01
Pharmacology, Toxicology and Pharmaceutics	278	0.95	3	1.52
Physics and Astronomy	847	2.90	8	4.04
Psychology	278	0.95	5	2.53
Social Sciences	2037	6.97	14	7.07
Health Professions	72	0.25	2	1.01
Total	**29207**	**100**	**198**	**100**

Table 3: Final Categorization Scheme at Area Level

A simple glance at Table 3 suffices to discover a group of dense areas covering a high number of journals. In more specific terms, this means that a set of 21,940 journals, that is, approximately 75% of the total taking into account the overlap, cite categories covered by just 8 of the 24 areas constituting the final scheme. *Medicine* (23.73%) and *Biochemistry, Genetics and Molecular Biology* (17.85%) stand out quite clearly. As a general rule, the denser the area appears, the more categories it includes, although there are some exceptions, for instance, in *Computer Science*; *Immunology and Microbiology*; or *Economics, Econometrics and Finance*.

At the level of categories (see Appendix), we encountered one small group of very populous categories covering thousands of journals, a medium-size group of categories including hundreds of journals, and a great one formed by categories embracing fewer than 100 journals. To explore this finding, a ranking of categories based on the number of journals citing each category was constructed. These values were then transformed into percentages, and the cumulative percentages for this distribution were finally added as well. Similar to what

happened with the areas, these findings (partially given in Table 4) evidenced a large aggregation of journals in several categories of the new classification scheme. It was moreover seen that only 15 of the 198 categories of the new classification scheme proved sufficient to categorize nearly 50% of the 14,416 journals conforming the final set (again, there was overlap in calculating percentages). Both the aggregations and the decreasing number of areas and categories most likely occurred because the method implies a flow of journals moving from certain categories to others as iterations proceed, as well as the final isolation of many categories.

Rank	CATEGORIES	Journals Citing Category	Percentage	Cumulative Percentage
1	Immunology	1740	5.96	5.96
2	Cell Biology	1688	5.78	11.74
3	Biochemistry	1483	5.08	16.81
4	Psychiatry and Mental Health	1380	4.72	21.54
5	Cardiology and Cardiovascular Medicine	1221	4.18	25.72
6	Public Health, Environmental and Occupational Health	1106	3.79	29.51
7	Sociology and Political Science	944	3.23	32.74
8	Economics and Econometrics	838	2.87	35.61
9	Electrical and Electronic Engineering	820	2.81	38.42
10	Oncology	683	2.34	40.75
11	Condensed Matter Physics	631	2.16	42.91
12	Ecology	616	2.11	45.02
13	Physical and Theoretical Chemistry	552	1.89	46.91
14	Cancer Research	536	1.84	48.75
15	Physiology	505	1.73	50.48

Table 4: Top 15 Ranked Journals per Category with New Categorization

A number of factors play some role in this phenomenon. Firstly, an implicit feature of the approach is that it keeps the most outstanding categories and discards the less representative ones per each journal. Thus, the method focuses on generality rather than specificity in its

attempt to delineate and define a journal subject. Figures 1 and 2 (above) serve to illustrate this aspect of performance.

A second reason is the drawing power of certain categories, particularly from the area of pure sciences. The data provided in Table 4 reveal that, on the whole, categories ranked in foremost positions (*Immunology*; *Cell Biology*; *Biochemistry*; *Psyichiatry and Mental Health*; *Cardiology and Cardiovascular Medicine*, etc.) are encompassed in pure science areas such as *Immunology and Microbiology*; *Biochemistry, Genetics and Molecular Biology*; or *Medicine*. Only the categories of *Sociology and Political Science*; and *Economics and Econometrics,* more connected to the area of the social sciences area, are an exception within the top 15 categories list. Therefore, since the method developed is based on journal reference analysis, we infer the existence of a substantial share of journals citing database pure science sources, despite the subject area or category where they are actually included. The disuse of different categories is a common issue for those categories with a low rate of journals citing them. Thus, categorized journals are finally attracted to more powerful categories as a consequence of the Matthew Effect. This phenomenon becomes more acute as more iterations are run.

Nevertheless, another possibility concerns the relatively new disciplines with a non-cohesive background. Normally these disciplines cite intellectual bases (Chen 2006) pertaining to other fields with very close boundaries, or which can find a "fertile ground in a neighboring field" (Small 1999) evoking inter-disciplinarily symptoms. Some examples of this are the categories *Gender Studies*; *Human Factors and Ergonomy*; *Nature and Landscape Conservation*, or a few from the area of *Nursing*.

Of course, all these explanations can be extrapolated to the application of the different thresholds used in the development and design of our method. One additional disadvantage is that, the higher the threshold, the higher the ratio of multi-categorized journals proved to be; and conversely, the lower the threshold, the lower the number of categories falling into the final categorization scheme.

Conclusions

The proposal featured permitted us to categorize 14,416 Scopus journals from an initial set of 17,158 as well as to restructure and redefine the SJR classification scheme at two levels of aggregation. Admittedly, while the method provided a consistent SJR classification scheme, we are mindful that it can not be considered as a definitive classification solution, since it does not provide a comprehensive and definitive placement of the journals assessed. For the time being, this approach should be supplemented with additional techniques, based either on citation or on text, in order to classify the whole set of covered journals.

A good performance of the method is closely linked to a good set-up of the main parameters, namely, total number of iterations to use, threshold to apply and cutoff fixed. Heuristic processes and empirical tests were determining factors for configuring it. The designation of 12 iterations was enough to make manifest that more iterations meant bigger aggregations of journals into a small set of categories. This fact may be useful in the case that one keeps running iterations until grouping journals into vast, basic areas of scientific knowledge. Regarding thresholds and cutoff, we noted they were very closely related. From the whole set of tests executed, the most balanced mix of these parameters, in terms of *number of categories cited by journals*, *mean of categories per journal*, and *number of multi-categorized journals*, took place at iteration 2 of the threshold 31%. Of course, the results of this combination were not the same in every test.

The method inevitably entails missed categories due to a large aggregation of journals into a reduced number of categories. Thus, an ever-increasing share of journals is seen to use an ever-decreasing share of categories. In other words, a small set of categories would suffice to categorize a vast set of journals, and we believe the method could offer better results by categorizing journals to a high level of aggregation, such as subject area. The category aggregation problem could be minimized by modifying the method, for instance using only the first iteration and discarding the remaining ones.

The citation flows between categories evidenced a clear attraction exerted by sources covering pure science. This happened among categories of different subject areas and also among categories of the same area. Some causes behind this might be related to database coverage, citation behavior, or the degree of consolidation of each particular discipline.

It is also interesting to highlight another positive aspect of the method, concerning the decreasing number of journals categorized under the Multidisciplinary category. Journals assigned to this category shifted to narrower categories later, mostly to Miscellaneous, which covers different categories inside the same subject area. In our study, the number of Multidisciplinary journals went from the 65 journals of the SJR original categorization to 28 journals under the new scheme.

Before closing, we underline that upcoming studies should provide a good framework to implement alternative techniques and to improve our method, so that a complete assignation of categories for each journal gathered and analyzed could be carried out. Forthcoming research efforts will thus be directed toward cluster analysis, examining the citation dimension through coupling and cross-citation. Later on, other possibilities may be explored, such as text dimension, using keywords or text parts extracted from journal articles.

Acknowledgment

The authors thank Jean Sanders for editing the text.

Bibliography

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology,* 57 (3), 359-377.

Chen, C. M. (2008). Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. *Journal of the American Society for Information Science and Technology,* 59 (14), 2296-2304.

Elsevier (2002). Scopus. http://www.scopus.com/home.url. Accessed 20-6-2011

Glänzel, W. and Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics,* 56 (3), 357-367.

Glänzel, W., Schubert, A., and Czerwon, H. J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics,* 44 (3), 427-439.

Glänzel, W., Schubert, A., Schoepflin, U., and Czerwon, H. J. (1999). An item-by-item subject classification of papers published in journals covered by the SSCI database using reference analysis. *Scientometrics,* 46 (3), 431-441.

Gómez, I., Bordons, M., Fernández, M. T., and Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics,* 35 (2), 223-235.

Leydesdorff, L. (2002). Dynamic and evolutionary updates of classificatory schemes in scientific journal structures. *Journal of the American Society for Information Science and Technology,* 53 (12), 987-994.

López-Illescas, C., Noyons, E. C. M., Visser, M. V., Moya-Anegón, F., and Moed, H. F. (2009). Expansion of scientific journal categories using reference analysis: How can it be done and does it make a difference? *Scientometrics,* 79 (3), 473-490.

Narin, F. *Evaluative bibliometrics : the use of publication and citation analysis in the evaluation of scientific activity*, Cherry Hill, N.J.: Computer Horizons, Inc., 1976.

Narin, F., Carpenter, M., and Nancy, C. (1972). Interrelationships of scientific journals. *Journal of the American Society for Information Science,* 23 (5), 323-331.

Narin, F., Pinski, G., and Gee, H. H. (1976). Structure of the Biomedical Literature. *Journal of the American Society for Information Science,* 27 (1), 25-45.

Pinski, G. and Narin, F. (1976). Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics. *Information Processing and Management,* 12 (5), 297-312.

Schubert, A. and Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics,* 36 (3), 311-324.

Scimago Lab (2007). Scimago Journal & Country Rank. http://www.scimagojr.com/. Accessed 20-6-2011

Small, H. (1999). A passage through science: crossing disciplinary boundaries. *Library Trends,* 48 (1), 72-108.

Thomson Reuters (2010). ISI Web of Science. Thomson Reuters. http://apps.isiknowledge.com/WOS_GeneralSearch_input.do?preferencesSaved=&product=WOS&SID=Z1CB8C2BE1G8nk3f9gM&search_mode=GeneralSearch. Accessed 20-6-2011

Weinberg, A. (1963). Criteria for scientific choice. *Minerva,* 1 (2), 159-171.

Appendix

AREAS	CATEGORIES	Journal Citing Category	Percentage
Multidisciplinary	Multidisciplinary	28	0.096
Agricultural and Bioogical Sciences	Agricultural and Biological Sciences (miscellaneous)	2	0.007
	Agronomy and Crop Science	82	0.281
	Animal Science and Zoology	247	0.846
	Aquatic Science	272	0.931
	Ecology, Evolution, Behavior and Systematics	67	0.229
	Food Science	255	0.873
	Forestry	45	0.154
	Horticulture	1	0.003
	Insect Science	63	0.216

	Plant Science	427	1.462
	Soil Science	82	0.281
Arts and Humanities	History	250	0.856
	Language and Linguistics	159	0.544
	Classics	1	0.003
	Literature and Literary Theory	45	0.154
	Music	16	0.055
	Philosophy	87	0.298
	Religious Studies	19	0.065
	Visual Arts and Performing Arts	9	0.031
Biochemistry, Genetics and Molecular Biology	Biochemistry, Genetics and Molecular Biology (miscellaneous)	34	0.116
	Aging	15	0.051
	Biochemistry	1483	5.078
	Biophysics	3	0.010
	Biotechnology	69	0.236
	Cancer Research	536	1.835
	Cell Biology	1688	5.779
	Developmental Biology	13	0.045
	Endocrinology	292	1.000
	Genetics	431	1.476
	Molecular Biology	142	0.486
	Physiology	505	1.729
	Structural Biology	2	0.007
Business, Management and Accounting	Business, Management and Accounting (miscellaneous)	3	0.010
	Accounting	10	0.034
	Business and International Management	82	0.281
	Management Information Systems	9	0.031

	Management of Technology and Innovation	383	1.311
	Marketing	65	0.223
	Organizational Behavior and Human Resource Management	5	0.017
	Strategy and Management	227	0.777
	Tourism, Leisure and Hospitality Management	8	0.027
Chemical Engineering	Catalysis	61	0.209
	Chemical Health and Safety	1	0.003
	Colloid and Surface Chemistry	14	0.048
	Filtration and Separation	4	0.014
	Fluid Flow and Transfer Processes	105	0.360
	Process Chemistry and Technology	47	0.161
Chemistry	Chemistry (miscellaneous)	3	0.010
	Analytical Chemistry	121	0.414
	Electrochemistry	34	0.116
	Inorganic Chemistry	21	0.072
	Organic Chemistry	195	0.668
	Physical and Theoretical Chemistry	552	1.890
	Spectroscopy	12	0.041
Computer Science	Computer Science (miscellaneous)	7	0.024
	Artificial Intelligence	114	0.390
	Computational Theory and Mathematics	77	0.264
	Computer Graphics and Computer-Aided Design	39	0.134
	Computer Networks and Communications	3	0.010
	Computer Science Applications	8	0.027
	Computer Vision and Pattern Recognition	22	0.075
	Hardware and Architecture	43	0.147
	Human-Computer Interaction	7	0.024

	Information Systems	88	0.301
	Signal Processing	5	0.017
	Software	216	0.740
Decision Sciences	Management Science and Operations Research	110	0.377
Earth and Planetary Sciences	Earth and Planetary Sciences (miscellaneous)	4	0.014
	Atmospheric Science	191	0.654
	Computers in Earth Sciences	26	0.089
	Earth-Surface Processes	77	0.264
	Geochemistry and Petrology	351	1.202
	Geology	75	0.257
	Geophysics	111	0.380
	Geotechnical Engineering and Engineering Geology	73	0.250
	Oceanography	44	0.151
	Paleontology	54	0.185
	Space and Planetary Science	50	0.171
Economics, Econometrics and Finance	Economics and Econometrics	838	2.869
	Finance	2	0.007
Energy	Energy (miscellaneous)	1	0.003
	Energy Engineering and Power Technology	85	0.291
	Fuel Technology	17	0.058
	Nuclear Energy and Engineering	23	0.079
	Renewable Energy, Sustainability and the Environment	46	0.158
Engineering	Engineering (miscellaneous)	30	0.103
	Aerospace Engineering	15	0.051
	Biomedical Engineering	43	0.147
	Civil and Structural Engineering	136	0.466
	Computational Mechanics	140	0.479

	Control and Systems Engineering	232	0.794
	Electrical and Electronic Engineering	820	2.808
	Industrial and Manufacturing Engineering	57	0.195
	Mechanical Engineering	445	1.524
	Mechanics of Materials	1	0.003
	Ocean Engineering	17	0.058
	Safety, Risk, Reliability and Quality	13	0.045
	Building and Construction	19	0.065
Environmental Science	Environmental Science (miscellaneous)	6	0.021
	Ecology	616	2.109
	Environmental Chemistry	343	1.174
	Environmental Engineering	85	0.291
	Health, Toxicology and Mutagenesis	13	0.045
	Management, Monitoring, Policy and Law	8	0.027
	Waste Management and Disposal	1	0.003
	Water Science and Technology	97	0.332
Immunology and Microbiology	Applied Microbiology and Biotechnology	2	0.007
	Immunology	1740	5.957
	Microbiology	240	0.822
	Parasitology	40	0.137
	Virology	1	0.003
Materials Science	Materials Science (miscellaneous)	1	0.003
	Biomaterials	22	0.075
	Ceramics and Composites	48	0.164
	Electronic, Optical and Magnetic Materials	131	0.449
	Materials Chemistry	139	0.476
	Metals and Alloys	128	0.438

	Polymers and Plastics	130	0.445
	Surfaces, Coatings and Films	20	0.068
Mathematics	Algebra and Number Theory	134	0.459
	Analysis	15	0.051
	Applied Mathematics	370	1.267
	Computational Mathematics	20	0.068
	Discrete Mathematics and Combinatorics	18	0.062
	Logic	11	0.038
	Mathematical Physics	21	0.072
	Statistics and Probability	101	0.346
	Theoretical Computer Science	49	0.168
Medicine	Medicine (miscellaneous)	35	0.120
	Anesthesiology and Pain Medicine	76	0.260
	Cardiology and Cardiovascular Medicine	1221	4.181
	Critical Care and Intensive Care Medicine	8	0.027
	Complementary and Alternative Medicine	2	0.007
	Dermatology	85	0.291
	Emergency Medicine	14	0.048
	Endocrinology, Diabetes and Metabolism	67	0.229
	Epidemiology	30	0.103
	Gastroenterology	117	0.401
	Genetics (clinical)	3	0.010
	Geriatrics and Gerontology	48	0.164
	Health Informatics	4	0.014
	Health Policy	33	0.113
	Hematology	23	0.079
	Microbiology (medical)	1	0.003

	Nephrology	47	0.161
	Neurology (clinical)	406	1.390
	Obstetrics and Gynecology	137	0.469
	Oncology	683	2.338
	Ophthalmology	86	0.294
	Orthopedics and Sports Medicine	265	0.907
	Otorhinolaryngology	118	0.404
	Pathology and Forensic Medicine	102	0.349
	Pediatrics, Perinatology and Child Health	152	0.520
	Pharmacology (medical)	7	0.024
	Psychiatry and Mental Health	1380	4.725
	Public Health, Environmental and Occupational Health	1106	3.787
	Pulmonary and Respiratory Medicine	97	0.332
	Radiology, Nuclear Medicine and Imaging	164	0.562
	Rehabilitation	40	0.137
	Rheumatology	37	0.127
	Surgery	282	0.966
	Transplantation	1	0.003
	Urology	54	0.185
Neuroscience	Behavioral Neuroscience	10	0.034
	Cognitive Neuroscience	97	0.332
Pharmacology, Toxicology and Pharmaceutics	Pharmaceutical Science	38	0.130
	Pharmacology	180	0.616
	Toxicology	60	0.205
Physics and Astronomy	Acoustics and Ultrasonics	35	0.120
	Astronomy and Astrophysics	1	0.003
	Condensed Matter Physics	631	2.160

	Instrumentation	2	0.007
	Nuclear and High Energy Physics	80	0.274
	Atomic and Molecular Physics, and Optics	76	0.260
	Radiation	7	0.024
	Statistical and Nonlinear Physics	15	0.051
Psychology	Applied Psychology	4	0.014
	Clinical Psychology	4	0.014
	Developmental and Educational Psychology	125	0.428
	Experimental and Cognitive Psychology	143	0.490
	Social Psychology	2	0.007
Social Sciences	Archeology	67	0.229
	Development	2	0.007
	Education	328	1.123
	Geography, Planning and Development	318	1.089
	Health (social science)	8	0.027
	Law	105	0.360
	Library and Information Sciences	105	0.360
	Sociology and Political Science	944	3.232
	Transportation	37	0.127
	Anthropology	64	0.219
	Communication	39	0.134
	Cultural Studies	5	0.017
	Demography	5	0.017
	Urban Studies	10	0.034
Health Professions	Radiological and Ultrasound Technology	70	0.240
	Speech and Hearing	2	0.007

Pre-print of: *Journal of Informetrics*, 2014, vol. 8(2), p. 369-383.
http://dx.doi.org/10.1016/j.joi.2014.01.011

OPTIMIZING SCIMAGO JOURNAL & COUNTRY RANK CLASSIFICATION BY COMMUNITY DETECTION

Gómez-Núñez, Antonio J.[a]; Batagelj, Vladimir[b]; Vargas-Quesada, Benjamín[c,e]; Moya-Anegón, Félix[d,e]; Chinchilla-Rodríguez, Zaida[d,e]

(a) CSIC, SCImago Research Group Associated Unit. Faculty of Communication and Documentation, Campus de Cartuja s/n, 18071 Granada, Spain anxusgo@gmail.com ✉
(b) University of Ljubljana, Faculty of Mathematics and Physics. Jadranska 19, 1000 Ljubljana, Slovenia vladimir.batagelj@fmf.uni-lj.si
(c) University of Granada, Department of Information and Communication. Faculty of Communication and Documentation, Campus de Cartuja s/n, 18071 Granada, Spain benjamin@ugr.es
(d) CSIC, Institute of Public Goods and Policies. Albasanz 26-28, 28037 Madrid, Spain felix.demoya@csic.es, zaida.chinchilla@csic.es
(e) SCImago Research Group

Abstract
Subject classification arises as an important topic for bibliometrics and scientometrics as to develop reliable and consistent tools and outputs. For this matter, a well delimited underlying subject classification scheme reflecting science fields becomes essential. Within the broad ensemble of classification techniques clustering analysis is one of the most successful.

Two clustering algorithms based on modularity, namely, VOS and Louvain methods, are presented in order to update and optimise journal classification of SCImago Journal & Country Rank (SJR) platform. We used network analysis and visualization software *Pajek* to run both algorithms on a network of more than 18,000 SJR journals combining three citation-based measures, that is, direct citation, co-citation and bibliographic coupling. The set of clusters obtained was termed through category labels assigned to SJR journals and significant words from journal titles.

Despite of both algorithms exhibiting slight performance differences, the results showed a similar behaviour in grouping journals and, consequently, they seem to be appropriate solutions for classification purposes. The two new generated algorithm-based classifications were compared to other bibliometric classification systems such as the original SJR one and WoS Subject Categories in order to validate their consistency, adequacy and accuracy. Although there are notable differences among the four classification systems analysed, we found a certain coherence and homogeneity among them.

Keywords: Community detection; Clustering; SCImago Journal & Country Rank; Journal classification; Citation-based network.

1.- Introduction

Classification is a broadly covered topic in *Bibliometrics* and *Scientometrics* because of its significance in developing of final bibliometric and scientometric outputs, mainly based on scientific literature included in databases and repositories. Thus, the literature collected by these information and reference sources need to be organized through an appropriate and consistent classification scheme not only for information retrieval purposes, but also for designing reliable and solid tools as rankings, domain analysis or scientograms, which are of an outstanding value, for instance, in science policy design and science evaluation processes.

Normally, database subject classification schemes are constructed on the basis of a disciplinary structure which pretends to replicate the main fields and subfields of research and scientific knowledge recorded in the literature stored in databases. Then, classification of scientific literature can be made at journal or paper level. The most highly reputed scientific databases at present, namely, Web of Science (Thomson Reuters, 2009) and Scopus (Elsevier, 2004), have a very similar two-level hierarchical subject classification schemes consisting of subject areas at a high and wider level and subject categories at low and more specific level. In both databases, journals are assigned to one or more categories and their papers are inheriting subject categories of journals which they belong to. In Web of Science (WoS) case, journal assignment is executed by ISI (currently, Thomson Reuters) staff taking into account several criteria as journal titles or citation patterns (Pudovkin & Garfield, 2002).

Delimitation of scientific fields required in developing disciplinary subject classification schemes can be done through many different approaches varying from empirical and pragmatic techniques to automated procedures based on statistics and computerized methods. Within the latter ones, clustering analysis is one of the most valuable and usual methods used for classification tasks in several and distinct scientific fields as *Library and Information Science*, *Psychology*, *Medicine* or *Biology* among others.

2.- Related Works

Many clustering algorithms and techniques have been developed in order to get optimal solutions for the classification problems befallen in scientific fields above mentioned. However, clustering methods have been widely used by researchers dealing with information visualization techniques in order to map the structure of scientific knowledge and research. For this reason, they needed a good underlying classification of fields and subfields to be mapped. A total of 20 representative approaches in mapping science fields and their relations working from Web of Knowledge and Scopus database literature were compared and condensed by Klavans and Boyack (2009).

Clustering and mapping procedures have been conducted on different levels of aggregation, or in other words, using different units of analyses. Thus, at journal level a large number of researchers have applied different cluster algorithms to journal-journal relation matrices or networks based on citations, co-citations or bibliographic coupling. Chang and Chen (2011)

applied the *minimum span clustering (MSC)* method to a citation square matrix of roughly 1,600 SSCI journals. Leydesdorff, Hammarfelt and Salah (2011) tried to merge a map of humanities based on Thomson Reuters' A&HCI database in a global map of science previously developed (Rafols, Porter, & Leydesdorff, 2010) and used the k-core algorithm for mapping 25 specific A&HCI subject categories. Archambault, Beauchesne and Caruso (2011) designed a scientific journal ontology aimed to simplify the output of bibliometric data and analysis. The new journal ontology was built on feedback from previous existing journal classification whose categories were considered as "seeds" for the initial journal assignment. Three automatic classification procedures using either text or citation data from papers published in around 34,000 journals and conference proceedings from Scopus and WoS were executed. However, the final solution was generated according to the iterative analysis of citation and references patterns between subject fields and journals. Leydesdorff and Rafols (2012) collaborative work produced a study where a 9,162 journal-journal citation matrix extracted from the 2009 volume of the SCI-Expanded was used to map interactive global journal maps. They compared several methods and, among them, different clustering algorithms to group journals into clusters. More recently, Börner et al. (2012) introduced a methodology to design and subsequently update a map of science and classification system solicited by the University of California, San Diego (UCSD). To build the map a combination of text and link journal-journal similarity matrices based on Scopus and WoS data were used. Then, journal clustering was executed on a filtered matrix derived from modified cosine similarities. Finally, the calculation of similarities among clusters as well as their positions and relationships enabled depicting the UCSD map.

Lately, there has existed a research trend working with clustering algorithms for analysis, validation, and improvement of classification schemes based on journals from various perspectives. ECOOM research group of KU Leuven has addressed this topic throughout several publications where different clustering algorithms as Ward clustering or Multi-level Aggregation Method (also known as Louvain method) were applied on journal cross-citation and hybrid (text/citation) matrices (Janssens, Zhang, Moor, & Glänzel, 2009; Zhang, Glänzel, & Liang, 2009; Zhang, Janssens, Liang, & Glänzel, 2010).

On the other side, by taking documents as unit of analysis Small (1999) developed a methodology to visualize and to obtain a hierarchical multidisciplinary map of science through a method combining fractional citation counting of cited papers, co-citation single-linkage clustering with limits on cluster size, and two-dimensional ordination according to a geometric triangulation process. Ahlgren and Colliander (2009) studied different document-document similarity approaches based on text, coupling and a combination of both as well as several methods to map and classify a set of 43 documents from the journal *Information Retrieval*. Complete-linkage clustering was applied to group articles and the final result of assignment was compared with an expert-based classification using adjusted Rand Index. Similarly, Boyack et al. (2011) employed a combination of graph layout and average-link clustering to different text-based similarity-measure matrices constructed through relevant information from titles, abstracts, and MeSH subject headings of 2.15 million of papers extracted from the Medline database. They compared and assessed nine similarity approaches through Jensen-Chanon divergence and concentration measures. Later on, Waltman and Van Eck (2012) faced an even

more complex challenge by designing a detailed methodology to create a publication-level classification system using a multilevel clustering algorithm on a direct citation (disregarding the direction) network constituted of almost 10 million publications. In their opinion, the methodology strength is sustained on transparency and simplicity as well as modest computing and memory requirements. Klavans, Small and Boyack (2013) introduced the reference pair proximities as a new variable to improve accuracy of co-citation clustering. To do so, they used a corpus of 270,521 Scopus full text documents from 2007 and compared the results of traditional co-citation clustering approach to their new co-citation clustering, which evidenced a significant accuracy improvement.

Generally, clustering procedures on networks and matrices involves complex and hard calculations. This fact is more relevant when large datasets are being manipulated since hardware and software requirements are generally high. Another important issue is related to visualization of clustered data which should be clear and comprehensible. Both software VOSViewer (Van Eck & Waltman, 2010) and Pajek (Batagelj & Mrvar, 1997; Nooy, Mrvar, & Batagelj, 2012) arise as good tools for network analysis and information visualization, especially when large networks have to be manipulated. Additionally, VOSViewer includes its own classification algorithm whereas Pajek integrates different clustering algorithms that can be run easily once dataset is adapted to appropriated format required by the software.

3.- Objectives

The main goal of this study is to optimise and update journal classification of SCImago Journal & Country Rank (SJR) platform (SCImago, 2007) via clustering techniques. Using the software Pajek, we ran two automatic classification algorithms as to detect and extract communities (subject clusters) from a SJR journal network combining three citation-based measures. The set of automatic-extracted communities is representing the subject disciplinary structure of science and research recorded in SJR journals. Finally, the new resulted cluster-based systems will be compared to other classification systems such as WoS Subject Categories and the original SJR Classification to validate their consistency and accuracy by analysing and discussing the strengths and weakness of the results.

4.- Material

Our data set, covering a total number of 18,891 journals for a two-year time window (2009-2010), was gathered from SCImago Journal & Country Rank (SJR) database. In this set, only cited references going back from 2010 to 2000 were contemplated. All references were counted at paper level and later aggregated to journal level.

5.- Methods

In order to clarify and favour a better understanding of the distinct procedures developed in performing our study we have divided this section in 7 stages covering and detailing the required steps to follow.

5.1.- SJR Journal Classification: The Starting Point

Scopus classification system, and by extension, SJR original classification, is an a-priori two-level hierarchical classification system originally designed according to an up-bottom approach.

Hence, at first level, the classification covers a total of 27 broad *subject areas* which, at once, comprise a set of 308 specific subject categories at second level. Then, journals recorded at database are ascribed to one or several subject categories. Areas and categories tags were determined on the basis of All Science Journal Classification (ASJC). Generally, each subject area includes a subject category taking the same tag followed by '*miscellaneous*' addition. Journal assignment to categories was made on the basis of items adscription. Then, SCImago Research Group conducted some improvements on classification based on journal scope analysis and a constant feedback from journal editors. From the point of view of improving journal classification, feedback from editors may be an interesting argument to take into account. Thus, Archambault et al. (2011) even claim to be interested in feedback from researchers and practitioners using their journal ontology as to persist in refining journal assignment. However, in spite of various attempts, a wider improvement of SJR journal classification is needed in order to remove inconsistencies inherited from Scopus by allowing to final users to customise the journal-sets of SJR subject categories as to generate tailored rankings (Jacsó, 2013). A previous work based on SJR journal reference analysis (Gómez-Núñez, Vargas-Quesada, Moya-Anegón, & Glänzel, 2011) was oriented to this end.

5.2.- Journal Citation-based Relatedness Measures: Calculation and Formatting

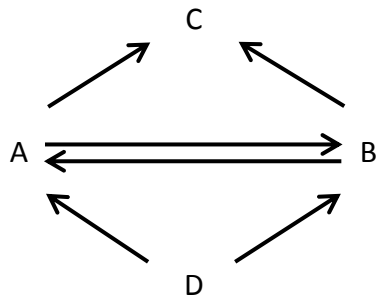
A plenty of publications dealing with classification and mapping of science and research have adopted text-based networks (Cantos-Mateos, Vargas-Quesada, Chinchilla-Rodríguez, & Zulueta, 2012; Liu, Hu, & Wang, 2011), citation-based networks (Leydesdorff & Rafols, 2012; Rafols & Leydesdorff, 2009) or combination of both (Glänzel, 2012; Janssens et al., 2009). Boyack y Klavans (2010) applied *Jensen-Shannon divergence* and *concentration* metrics as to prove the accuracy of clustering solutions emerging from different citation-based mapping methods. The results revealed the best performance in bibliographic coupling approach, followed closely by co-citation and direct citation further. Also, Waltman and Van Eck (2012) analysed advantages and disadvantages of three citation-based approaches. After that, they chose direct citation as relatedness measure in constructing a publication-level classification. Primarily, they based their decision on saving computer resources for processing the large data set of almost 10 million of publications that they copied with. However, they argued that direct citation are expected to provide strongest relatedness links between publication, contrary to co-citation and bibliographic coupling, which could be considered more indirect mechanisms. On the other hand, they noted that the use of direct citations can lead up to a loss of information because of citations to earlier publications and, similarly, citations from later publications are not being contemplated.

In this work, we are exploiting citation-based approaches on journal networks. This allows us to cover the three main types of citation links expressing a degree of relatedness between journals. In this way, we will be adding both strengths and weakness from each measure. Thereby, our approach could be considered a 'fair' and balanced one by offsetting all weakness coming from *direct citation*, *co-citation* and *bibliographic coupling* separately. When these important points were reflected, we constructed three journal networks, one for each citation-based measure. The three networks were calculated at the document level and then aggregated to journals. For co-citation and bibliographic coupling calculation, references co-occurring were counted only once per paper by following the binary counting described by

Rousseau and Zuccala (2004) and avoiding what Vargas-Quesada and Moya-Anegón (2007) named latent co-citation.

5.3.- Citation-based Measures Combination

Once the three citation-based networks were generated we combined them into a new one collecting pairwise journals and their relatedness strength expressed by the sum of direct citation, co-citation and bibliographic coupling links. By doing so, we got a final network based on raw data and containing what Persson (Persson, 2010) named *Weighted Direct Citation (WDC)* links. Below, we can display the diagram used by Persson in order to integrate these three citation-based measures and calculate the WDC. Nevertheless, we have introduced a small shift referring to both senses of the direct citation links.



Thus, we have used the next formula in citation based-measures combination:

c_ij = cu_ij + cc_ij + max(ci_ij,ci_ji)

Where cu_ij = coupling, cc_ij= co-citation, ci_ij = direct citacion from i to j and ci_ji = direct citation from j to i.

Also, by knowing that A, B, C and D are journals we can adapt this formula according to Person's diagram in this way:

c_ij = ABC + DAB + max(AB,BA)

5.4.- Network Normalization

At the following stage of out method, the final network resulted from aggregation of raw data links was normalized using *Geo similarity* formula as follows:

s_ij = c_ij / sqrt(c_i*c_j), c_i = sum{ j: j!=i: c_ij }

This similarity measure is close to Cosine one and performs dividing elements of the matrix by geometric mean of both diagonal elements (Batagelj & Mrvar, 2003). Thereby, raw data were corrected and relatedness values between pairwise journals were transformed to values ranging from 0 to 1. This avoids problems related to misleading representations and overestimation of some science fields characterized by strong citation habits or covering large-size journals with a high power of attraction.

5.5.- Clustering Procedures

The next step in our methodology was to run clustering algorithms included in Pajek software on the normalized network. Pajek integrates several clustering methods in order to decompose networks by extracting different partitions such as islands, k-neighbours or block modelling. However, after several initial tests, we targeted on communities detection algorithms, namely, *VOS Clustering* (Waltman, Van Eck, & Noyons, 2010) and *Louvain Method* (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Both methods are grounded in modularity clustering proposed by Newman and Girvan (2004). However, while Louvain method optimizes modularity, VOS Clustering focused on optimizing a quality function (Batagelj & Mrvar, 2011). For this experiment, we chose Louvain and VOS methods based on Multilevel Coarsening + Single Refinement. Moreover, we had to set up several options regarding *resolution parameter*, *random restarts*, *maximum number of levels in each iteration*, and *maximum number of repetitions in each level*. Here, we fixed just the same options for VOS and Louvain algorithms. Firstly, we introduced distinct values in *resolution parameter*, moving them from 10 to 20 in order to get different Pajek partitions depicting diverse solutions in decomposing network and producing different set of clusters or communities. Then, the remaining parameters were configured with default values.

By analysing certain relevant indicators for each parameterized clustering algorithm solution, basically, the *number of clusters generated* and the *number of journal per clusters*, we estimated that network decompositions providing between 250 and 300 groups would be interesting for our final journal classification objective. Here, some important issues were considered. Firstly, we took into account the 250 subject categories currently included in WoS database since this scientific information source is not only an international referent within bibliometric and scientometric fields but also for scientists and researchers in general. Presently, SJR is including 308 subject categories and, therefore, we thought that a final set ranging from 250-300 categories will provide a balanced and refined subject structure. This point was reinforced with the experience acquired in a previous work (Gómez-Núñez et al., 2011). There, we noticed a regular behaviour in grouping journals which reveals a strong concentration of them in a few leading categories from a final set of 198 SJR categories. These leading categories are characterized by a high attractiveness, especially, when iterative reference analysis method was used. As commented earlier, this behaviour may be derived from citation habits of some scientific fields with an intense and well-defined citation practice such as the *Medicine* and allied sciences or some social science subfields as *Economics* or *Education*.

Apart from indicators above mentioned we applied some others (see Results section) to VOS and Louvain partitions matching with different 10-20 resolution parameters and we proceed to compare the results of both of them. Every partition was executed in Pajek and, later, saved to files as to be processed using spreadsheets and statistical software. Concretely, we selected VOS partition referring to resolution parameter 15, while a resolution parameter 18 was appointed in the Louvain case. In this decision, we basically looked for similar partitions in terms of the final number of clusters generated by VOS and Louvain methods under the premise of making comparable the results and the classification solutions in both clustering

algorithms evaluated. Besides, we established a threshold to define the minimum cluster size to 10 journals, discarding all those clusters which were not complying with this requirement.

5.6.- Labelling

After executing automatic clustering techniques we had to label the different subject groups or communities depicted by both algorithms and recorded in the selected partitions. To this end, we designed a multi-phase approach to solve the various instances occurred. At this moment, it is well to explain that in this work we are proposing a journal multi-assignment. Nevertheless, journal multi-assignment was due to labelling process and not to clustering methods used which have conducted a journal single assignment per cluster.

5.6.1. Labelling through SJR category tags

In a first approach, we took into consideration the citation frequencies from journals to former SJR categories. Thus, we counted how many times journals forming part of a cluster were citing original categories from SJR. After that, frequencies were transformed into percentages and into weighted scores using tf-idf formula by Salton and Buckley (1988) which we adapted to our particular case so:

$w_{i,j} = catf_{i,j} \times Log\ (N / cluf_i)$

Where $w_{i,j}$ = total weighted score; $catf_{i,j}$ = raw frequency of category 'i' into cluster 'j'; N = total number of clusters; and $cluf_i$ = number of clusters containing category 'i'

After that, all the categories were ranked by tf-idf scores and only those categories amounting at least a 33% over the total set of references cited by journals forming distinct clusters were selected as to delineate the cluster subjects. By means of this procedure journals were allocated to one up to four categories. Although many research works have defended a single and exclusive assignment of journals to clusters or categories (Archambault et al., 2011; Thijs, Zhang, & Glänzel, 2013; Waltman & Van Eck, 2012), there are strong reasons to think in journal multi-assignment. Generally, most of scientific journals are not covering a unique topic. This can be checked, for instance, by having a look at journal scopes. In some cases, authors have interest for publishing in journals out of their expertise field in order to get a higher prestigious, visibility or even impact. Moreover, current science often follows an interdisciplinary and collaborative model with several fields involved in solving different problems, copying with new challenges or looking for a continuous advance and development of science and research. Finally, we are aware of journal multi-assignment carried out in original SJR journal classification and we have pretended to keep taking this approach but with the aim of improving it.

5.6.2. Labelling through significant words of journal titles

This labelling approach was adopted in two particular cases:

1) When using category tags we found two clusters with exactly the same categories assigned, and, then, representing two identical subject groups.
2) In the whole labelling procedure, *Miscellaneous* and *Multidisciplinary* categories were rejected. After removing these categories, percentages and tf-idf scores were re-calculated.

However, in some clusters the number of journals was lower than the number or links pointing to SJR categories. This was not satisfying the condition of at least one link to category per journal.

In the two instances above noted, we reconsidered the labelling approach for clusters by using a textual component, such as significant words extracted from journal titles. After counting them, frequencies of most repeated words were taken as to delineate the subject topic of clusters. To support the text-based labelling stage and to fine-tune in denoting clusters we used some *Voyeur Tools* platform, which provides a set of online text analysis tools forming part of *Hermeneuti.ca* collaborative project (Sinclair & Rockwell, 2009).

5.7.- Validating Classification Proposals.

In closing our method, a validation of classifications generated by algorithms was desirable. There are different approaches aimed to this end. Expert assessment could be the best one, but, generally, it is very time- and cost-consuming. We thought that a suitable and less resource-consuming method is a comparison with some other classification systems. Especially useful would be a comparison versus the original SJR classification since the journal data set is just the same which facilitates the process. Nevertheless, we also included a comparison with ISI Subject Categories which is the subject classification system of the referent database in bibliometric scope, namely, WoS (and consequently, JCR + Arts & Humanities). To make possible this comparison we prepared a combined list consisting of SCI+SSCI journals collected from JCR 2010 release. JCR do not include journals of Arts & Humanities areas so an extra list of A&HCI journals of 2012 release downloaded from Thomson Reuter's website was added. The final list of journals was integrated by 11,715 journals, pertaining 8,005 to SCI, 2,678 to SSCI and 1,758 to A&HCI respectively. Therefore, there is a certain level of overlapping because a total of 726 journals were covered by distinct indexes together. Finally, we used ISSN field as to generate matching between journals of SJR, WoS, VOS and Louvain classifications which rise to 9694 journals, that is, an 82.75% from the total set.

6.- Results

6.1. Analysis of results derived from algorithm solutions

In an attempt to optimise and update SJR journal classification we analysed and compared the results derived from VOS and Louvain clustering methods according to distinct indicators related to the proper performance of both algorithms, such as (1) *number of given clusters*, (2) *number of journals classified* after applying the threshold of 10 journals as the minimum cluster size, and (3) *mean number of journals per cluster*. Besides, we developed two indicators coming from cluster labelling process, just as the (4) *journal multi-assignment*, and the (5) *weighted average of categories assigned to journals*. Again, we would like to remark that journal multi-assignment was a consequence of our labelling procedure and not due to VOS and Louvain performance which carries out a journal single assignment as they are hard clustering techniques.

Table 1 captures the values of the indicators (1), (2) and (3). As we pointed in the previous section, we projected around 250-300 journal subject groups to trace a basic and cohesive disciplinary structure in order to classify scientific journals. Therefore, we retained this premise

during the parameterization of VOS and Louvain algorithms as well as in choosing final partitions giving suitable results and better adapting to our final classification aim.

	(1) number of given clusters				(2) number of journals classified				(3) mean number of journals per cluster			
Resolution Parameter	*VOS*	*Louvain*	*VOS Threshold 10*	*Louvain Threshold 10*	*VOS*	*Louvain*	*VOS Threshold 10*	*Louvain Threshold 10*	*VOS*	*Louvain*	*VOS Threshold 10*	*Louvain Threshold 10*
10	531	550	174	153	18,891	18,891	18,271	18,170	35.6	34.3	105.0	118.8
11	593	601	200	173	18,891	18,891	18,212	18,085	31.9	31.4	91.1	104.5
12	662	666	225	186	18,891	18,891	18,080	17,966	28.5	28.4	80.4	96.6
13	723	727	234	201	18,891	18,891	18,018	17,890	26.1	26.0	77.0	89.0
14	787	794	261	216	18,891	18,891	17,896	17,739	24.0	23.8	68.6	82.1
15	848	862	270	234	18,891	18,891	17,729	17,652	22.3	21.9	65.7	75.4
16	904	932	297	245	18,891	18,891	17,665	17,488	20.9	20.3	59.5	71.4
17	973	999	308	266	18,891	18,891	17,504	17,412	19.4	18.9	56.8	65.5
18	1,043	1,064	337	280	18,891	18,891	17,422	17,287	18.1	17.8	51.7	61.7
19	1,120	1,126	348	301	18,891	18,891	17,266	17,235	16.9	16.8	49.6	57.3
20	1,170	1,195	367	319	18,891	18,891	17,135	17,086	16.1	15.8	46.7	53.6

Table 1: Number of clusters, number of journals classified and mean number of journals per cluster according to the diverse resolution parameters of VOS and Louvain

A simple glance to the distinct figures exposed in this work denotes a considerable parallelism in the distributions depicted over the alternative resolution parameters of VOS and Louvain indicators. In a certain way, this might be expected as a normal event since both algorithms are grounded in modularity clustering method proposed by Newman and Girvan (2004). If we focus on the (1) *number of clusters* offered by two clustering methods it can be noticed that VOS algorithm needed a resolution parameter lower than Louvain to get a similar number of groups. Moreover, when a threshold of 10 journals as minimum cluster size was set, VOS algorithm presented a more balanced ratio between the clusters upholding this threshold and the clusters without doing it as compared to Louvain one. Thus, VOS partition with resolution parameter 15 returned just 270 clusters collecting ten or more journals from the total set of 848 clusters, what is equivalent to a ratio of 0.3184 or almost 32% of clusters satisfying the threshold. On the other hand, Louvain partition with resolution parameter 18 produced a total set of 1064 clusters with only 280 clusters reaching ten or more journals, what involves a ratio amounting to 0.2632, meaning a little bit more of 26% of clusters with more than ten journals. Figures 1 and 2 show the whole distribution of clusters according to the resolution parameter defined in VOS and Louvain algorithms.

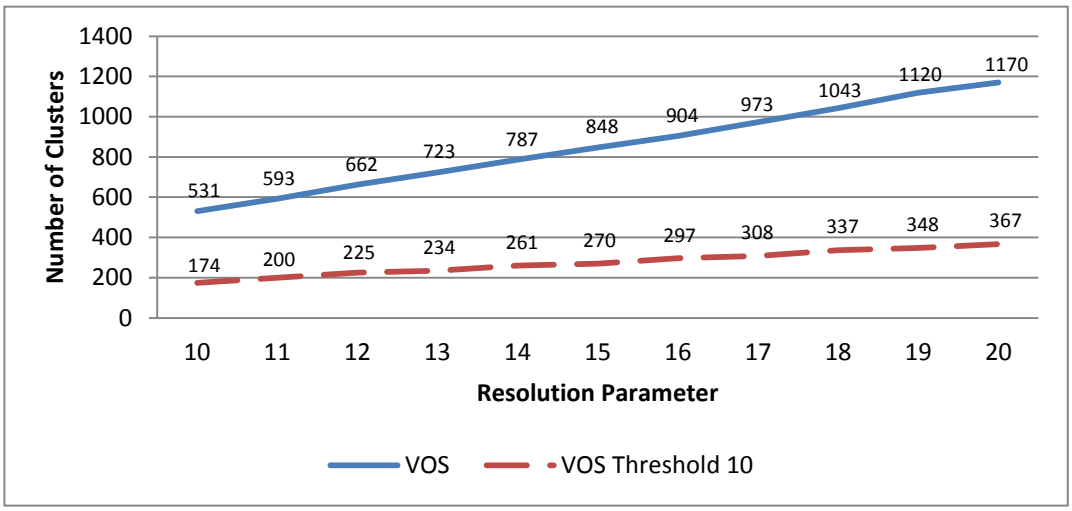


Figure 1: VOS cluster distribution over the different resolution parameters tuned

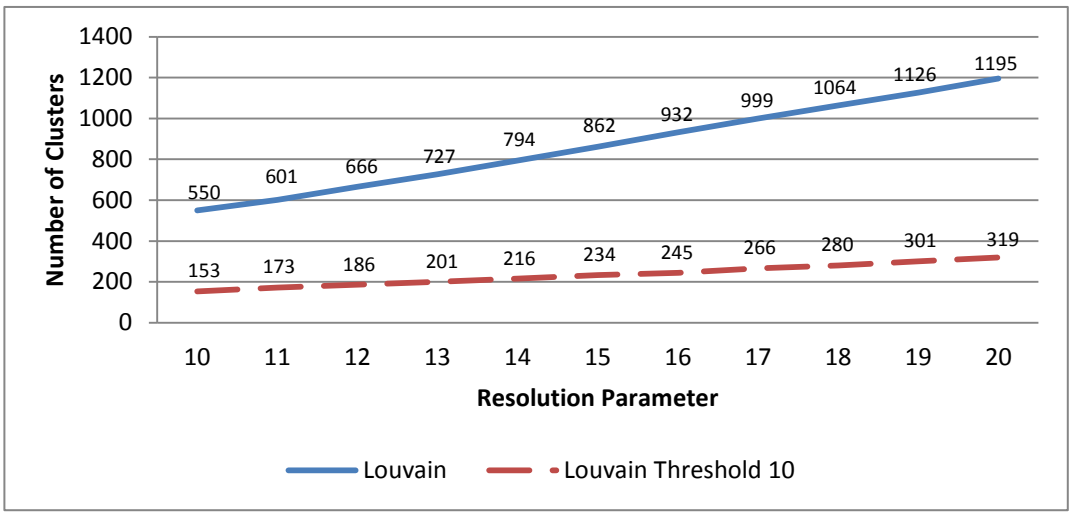


Figure 2: Louvain cluster distribution over the different resolution parameters tuned

In the own words of the authors of VOS clustering algorithm the resolution parameter included in their algorithm "helps to deal with the resolution limit problem of modularity-based clustering". They also claim that by introducing a sufficiently large value for the resolution parameter of their clustering technique all small clusters can always be determined, being the number of clusters generated larger when the value of resolution parameter is higher (Waltman et al., 2010). Then, the final number of clusters is directly proportional to the value of resolution parameter. Indeed, VOS and Louvain methods permitted to classify the 18,891 journals forming part of the initial network explored through the set of clusters provided. Nevertheless, a wide amount of this clusters were too small and were not able to form reliable and solid groups of journals. We have pointed out that only 31.84% of the total number of VOS clusters had a size higher than 10, while a mere 26.32% of clusters reached this threshold in Louvain method. This phenomenon could be due to the use of citation and their derivatives as measure units. Earlier on, we mentioned that some scientific fields portray a strong concentration and an outstanding attraction power of citations linking to publications including in them, normally, because of the own marked citation habits occurring inside these fields. So, the subject categories defining these fields are characterized by a great variance aggregation derived from the high quantity of citation received.

By observing indicators related to the (2) *number of journals classified* and the (3) *mean number of journals per cluster*, we detected a general behaviour which describes a better performance of VOS algorithm in classifying journals, that is, including journals in a particular cluster. In general, the mean of journals per cluster over the different resolution parameters returned by Louvain algorithm was higher. However, by examining the two partitions selected for our classification purpose, the mean number of journals per cluster was also a bit higher in favour of VOS resolution parameter 15. Figure 3 shows the whole distribution of journals classified in VOS and Louvain clusters over the distinct resolution parameters executed. In the same way, Figure 4 exposes the mean number of journals per cluster in two selected VOS and Louvain partitions.

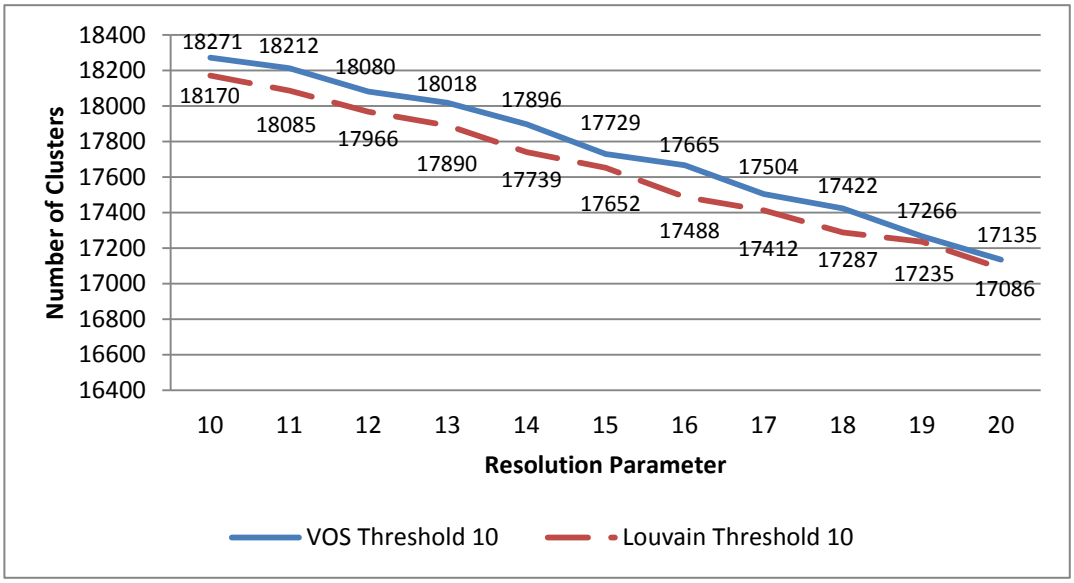


Figure 3: Distribution of classified journals over the different resolution parameters tuned in VOS & Louvain clustering algorithms

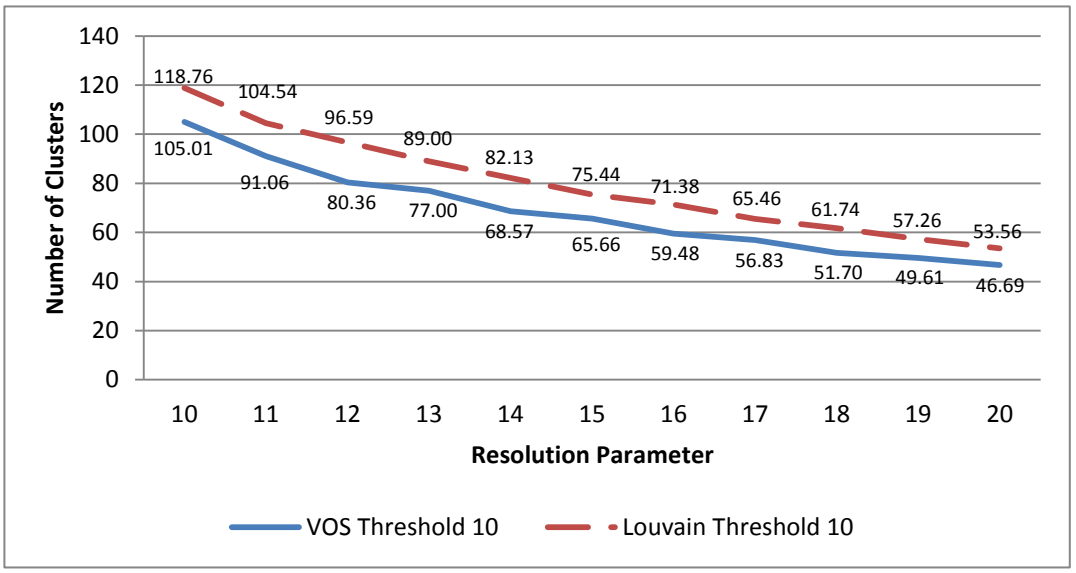


Figure 4: Distribution of mean of journals per cluster over the different resolution parameters tuned in VOS & Louvain clustering algorithms

Another example of the similitude of the results yielded by two clustering algorithms concerns to the (4) *journal multi-assignment indicator* which reflects the number of journals assigned to one or multiple categories at once. Figure 5 traces a very similar distribution of journals assigned by VOS and Louvain with more than 50% of them being ascribed to only one category and slight differences in journal multi-assignment. Admittedly, Louvain method had a worse result in assigning journals to four categories but, on the whole, Louvain assignment was a little bit better by concentrating more journals in only one category and fewer journals than VOS in two and three categories respectively. Also, this point can be supplemented and inferred by comparing the (5) *weighted average of categories assigned to journals* of VOS, which amounts to 1.50 categories per journal, and Louvain, which rises only to 1.48. Even so, the differences in VOS and Louvain algorithms multi-assignment are not very significant.

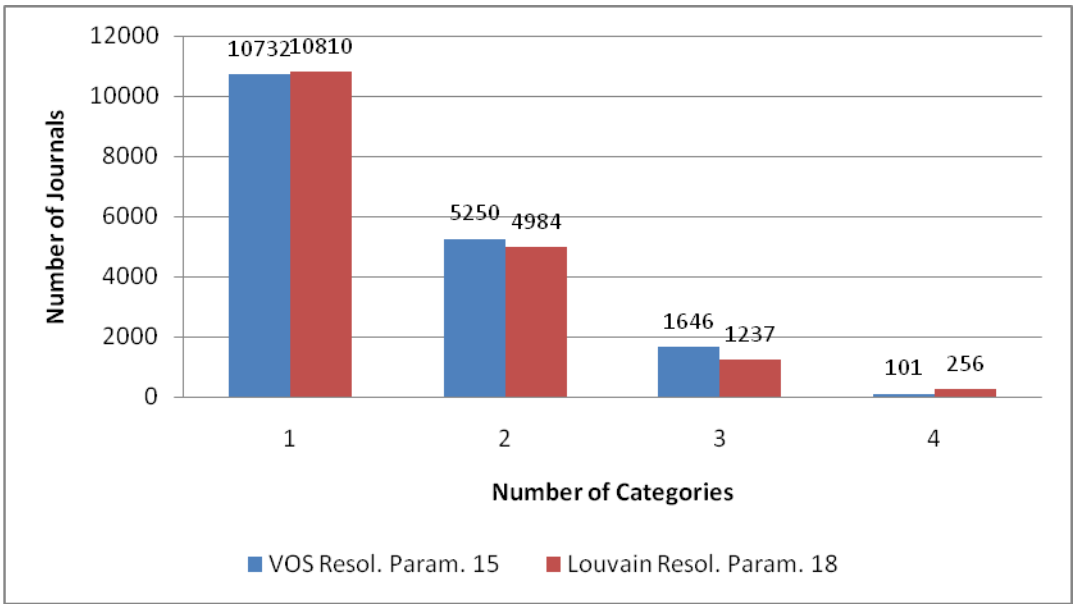


Figure 5: VOS & Louvain journal multi-assignment

6.1. Overall analysis and comparison among four different classification systems

Hitherto, we have highlighted only the analysis of VOS and Louvain clustering algorithms on the basis of statistical data and indicators. At this moment, we are detailing results related to differences and resemblances in journal final classification obtained after applying both algorithms. To do so, we are going to analyse and compare classifications originated by both clustering techniques together with original SJR classification and WoS (ISI Subject Categories).

	WoS					SJR					Louvain 18					VOS 15				
Total Set of Journals	11,715					18,891					18,891					18,891				
Number of Classified Journals	11,715					18,891					17,287					17,729				
Number of Categories	251					308					272					267				
Mean Number of Journals per Category	46.67					61.33					63.56					66.40				
Mean Number of Categories per Journal	1.54					1.61					1.48					1.50				
Overlapping Percentage	54.48%					60.73%					47.58%					49.89%				
Journals changing their Classification	-					-					7,159					7,606				
Journal Multi-Assignment	**Number of Categories**					**Number of Categories**					**Number of Categories**					**Number of Categories**				
	1	**2**	**3**	**4**	**+**	**1**	**2**	**3**	**4**	**+**	**1**	**2**	**3**	**4**	**+**	**1**	**2**	**3**	**4**	**+**
	6,990	3,432	986	261	46	12,025	3,893	1,863	751	359	10,806	4,986	1,237	256	0	10,730	5,251	1,646	101	0
	59.7%	29.3%	8.4%	2.2%	0.4%	63.7%	20.6%	9.9%	4.0%	1.9%	62.5%	28.8%	7.2%	1.5%	0%	60.5%	29.6%	9.3%	0.6%	0%

Table 2: Overall comparison among four classifications systems analysed. The Number of Classified Journals in Louvain and VOS systems results from application of minimum cluster size threshold ($t \geq 10$)

Table 2 captures overall data about four classifications compared. A detailed analysis of it enables to note some important observations. Regarding the *total set of journals* included, it is worth to mention that the number of journals covered by SJR overcomes the WoS set more than 1/3. Related to the *number of classified journals* we can see that after fixing threshold 10 in Louvain and VOS algorithms, the number of journals being classified descended to 17287 and 17729 respectively. This is not a result of performance of algorithms which were able to classify the whole set of 18891 original SJR journals. Journals left out the final set will have to be classified separately by a different solution. Reference analysis applied in a previous work (Gómez-Núñez et al., 2011) or 'sibling journals' which can be defined as those journals originally sharing former SJR categories and then extending their new cluster-based classification to journals under the threshold could be used.

The next point to address is the final *number of categories* forming part of the classification system. Here is convenient to clarify why the number of clusters expressed in Table 1 for VOS and Louvain methods are not in consensus with the number of categories (subject clusters) displayed in Table 2. Table 1 collects the number of clusters generated by algorithms without labelling. On its side, Table 2 is reflecting the number of clusters after our labelling process. Our approach made possible to have some clusters with different number and tags of categories assigned. For instance, cluster #82 in Louvain solution was labelled as 'Artificial Intelligence' + 'Information Systems' + 'Software' category tags, while cluster #90 was assigned to 'Artificial Intelligence' + 'Theoretical Computer Science' categories. The potential combinations of different number and tags of categories among the set of clusters is, therefore, the main reason to explain the difference in the number of categories included in Table 2. The final number of categories in VOS and Louvain decreased meaningfully in comparison to original SJR subject classification system and being closer to WoS system. This can be understood as a broad improvement, especially when data referring to overlap (Table 2) and distribution of journals over categories (Table 3) are observed. After indicating the *number of classified journals* and the final *number of categories* we can calculate the *mean number of journals per category*. VOS, Louvain and SJR systems overcome the value of 60 journals per category, being the highest value for VOS one with a total of 66.4. On his side, WoS system mean number only amounts to 46.67 journals per category although it is true that WoS journal coverage is much more reduced in comparison to the other three systems.

Other two interesting points of Table 2 are concerning *the mean number of categories per journal* and *overlapping percentage*. Both indicators are totally correlated and show the level of overlapping existing in four classifications compared. The main difference holds that mean number of categories per journal is expressed as per unit. The lowest level of overlapping was reached by Louvain system, followed closely by VOS. In both cases, overlapping levels are not going over the 50%. WoS and SJR systems surpass this level, being the worst overlapping figure the SJR one with a 60.73%. In this sense, again VOS and Louvain methods evidenced better solutions than SJR and WoS. Overlapping percentage was calculated by subtracting the number of records corresponding to (A) journals covered by the system (or in other words, the set of journals under consideration) from the number of records referring to final (B) journal multi-assignment (set of classified journals including multi-assigned journals), then multiplying by 100 and dividing the total by the (A) journals covered by the system [B-A/A*100].

The next row displayed in the table 2 is dealing with the number of journals *changing their classification* from SJR to Louvain and VOS system. Again VOS system get the highest figures by allocating a total of 7606 journals in new subject categories either by changing or by adding a new subject category to journals. This is equal to a 42.0% of the total set of journals classified. On the other hand, Louvain rise a 41.4% of journals changing their old classification. Such as the whole comparison process, both algorithm solutions yielded very similar results.

Finally, Table 2 is displaying the figures related to *journal multi-assignment* in four classification systems compared. Here, the best assignment of journals to one category was for SJR system with a 63.7% of the total set. The last place in the ranking was for WoS with a 59.7%. However, the four classification systems offered close percentages of journals assigned to one category. By taking into account our desire of allowing journal multi-assignment, the results obtained by Louvain and VOS can be judged as convenient because they concentrated the most of the journals in one and two categories. Louvain and VOS relative figures representing journal assignment executed on three and four categories are outperforming SJR and WoS systems by far. In addition, SJR and WoS systems made possible a journal assignment to more than four categories. Louvain and VOS solutions did not enable this kind of multi-assignment and, therefore, they provided a more balanced classification system.

A last important issue to analyse among the four classification systems is the proper *distribution of journals over the set of subject clusters or categories* generated. Table 3 is covering the top-20 categories regarding the number of journals included and expressed in raw data and percentage. Finally, we added a cumulative percentage in order to calculate the continuing aggregation of journals spread over categories. Now, the slightest distribution of journals over categories becomes WoS one. However, when classified journal set of WoS is compared with Louvain or VOS ones, then, these distributions are very similar among them. This is underscored through the percentage values of journals calculated in three classification systems. Admittedly SJR system has the largest set of classified journals it achieved the worst distribution of journals over categories as well. Furthermore, 'Medicine (miscellaneous)' category resulted especially remarkable by showing a high concentration of journals in it. More concretely, a 5.2% of the total set of SJR journals was included in this category. Of course, all indicators and calculations relating journals and categories were made considering journal overlap in four classification systems. A last thing to mind in Table 3 is related to the number of the same or really close categories which appear in 20-top ranking. A detailed analysis allowed uncovering that 7 of 20-top categories covered by the four classification systems were appearing in four systems together. This leads to think that despite the fact of having four different classification systems, there are certain coherence and homogeneity among them. Thus, while changes in number and position of categories may imply, in the case of algorithm systems a refinement of original SJR classification, a matching of a considerable number of categories may be a symptom of stability and consistency. The seven categories matching in WoS are: (1) 'HISTORY'; (2) 'ECONOMICS'; (3) 'MATHEMATICS'; (4) 'ENGINEERING, ELECTRICAL & ELECTRONIC'; (5) 'PSYCHIATRY'; (6) 'LANGUAGE & LINGUISTICS'; (7) 'EDUCATION & EDUCATIONAL RESEARCH'. The correspondence of these categories in SJR is: (1) 'History'; (2) 'Economics and Econometrics'; (3) 'Mathematics (miscellaneous)'; (4) 'Electrical and Electronic Engineering'; (5) 'Psychiatry and Mental Health'; (6) 'Language and Linguistics'; (7) 'Education'.

Finally, the set of categories in Louvain and VOS system was identical to SJR one, except for the category (3) 'Mathematics (miscellaneous)' which were labelled as 'Mathematics (general)'.

The final master tables covering the new classification of SJR journals proceeding from VOS and Louvain clustering methods can be accessed through the following links:

- VOS Classification: http://www.ugr.es/local/benjamin/vos15_classification.pdf
- Louvain Classification: http://www.ugr.es/local/benjamin/louvain18_classification.pdf

WoS				SJR				LOUVAIN RES. PAR. 18				VOS RES. PAR. 15			
Category	Num. of Journals	% Journals	Cumul. %	Category	Num. of Journals	% Journals	Cumul. %	Category	Num. of Journals	% Journals	Cumul. %	Category	Num. of Journals	% Journals	Cumul. %
HISTORY	331	1.829	1.829	Medicine (miscellaneous)	1,579	5.200	5.200	Sociology and Political Science	496	1.944	1.944	Electrical and Electronic Engineering	534	2.009	2.009
ECONOMICS	302	1.669	3.498	Education	524	1.726	6.926	Geology	417	1.634	3.579	Sociology and Political Science	480	1.806	3.816
BIOCHEMISTRY & MOLECULAR BIOLOGY	284	1.569	5.067	Sociology and Political Science	460	1.515	8.441	Literature and Literary Theory	415	1.627	5.205	Literature and Literary Theory	427	1.607	5.423
MATHEMATICS	276	1.525	6.592	Geography, Planning and Development	450	1.482	9.923	Geography, Planning and Development	393	1.540	6.746	Plant Science	393	1.479	6.901
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH	254	1.404	7.996	History	444	1.462	11.386	Electrical and Electronic Engineering	393	1.540	8.286	Geology	380	1.430	8.331
PHARMACOLOGY & PHARMACY	249	1.376	9.372	Electrical and Electronic Engineering	406	1.337	12.723	Psychiatry and Mental Health	372	1.458	9.744	Artificial Intelligence	371	1.396	9.728
ENGINEERING, ELECTRICAL & ELECTRONIC	247	1.365	10.737	Cultural Studies	375	1.235	13.958	Software	331	1.297	11.041	Education	369	1.389	11.116
NEUROSCIENCES	235	1.299	12.035	Social Sciences (miscellaneous)	374	1.232	15.190	Education	331	1.297	12.339	Software	352	1.325	12.441
MATHEMATICS, APPLIED	235	1.299	13.334	Economics and Econometrics	368	1.212	16.402	Hardware and Architecture	297	1.164	13.503	Psychiatry and Mental Health	341	1.283	13.724
PSYCHIATRY	233	1.288	14.621	Literature and Literary Theory	366	1.205	17.607	Religious Studies	297	1.164	14.667	Water Science and Technology	338	1.272	14.996
MATERIALS SCIENCE, MULTIDISCIPLINARY	219	1.210	15.831	Engineering (miscellaneous)	356	1.172	18.779	Applied Mathematics	283	1.109	15.776	Mathematics (general)	334	1.257	16.253
ENVIRONMENTAL SCIENCES	192	1.061	16.892	Psychology (miscellaneous)	351	1.156	19.935	Geochemistry and Petrology	282	1.105	16.882	Economics and Econometrics	318	1.197	17.449

LANGUAGE & LINGUISTICS	192	1.061	17.953	Public Health, Environmental and Occupational Health	335	1.103	21.039	Cultural Studies	264	1.035	17.916	Agronomy and Crop Science	312	1.174	18.623
SURGERY	186	1.028	18.981	Plant Science	327	1.077	22.116	Economics and Econometrics	252	0.988	18.904	Paleontology	309	1.163	19.786
CLINICAL NEUROLOGY	185	1.022	20.003	Language and Linguistics	327	1.077	23.193	History	250	0.980	19.884	History	300	1.129	20.915
PLANT SCIENCES	185	1.022	21.026	Psychiatry and Mental Health	325	1.070	24.263	Mechanical Engineering	245	0.960	20.844	Geography, Planning and Development	283	1.065	21.980
ONCOLOGY	181	1.000	22.026	Animal Science and Zoology	315	1.037	25.301	Civil and Structural Engineering	242	0.949	21.793	Mechanical Engineering	279	1.050	23.030
PHILOSOPHY	178	0.984	23.009	Mathematics (miscellaneous)	306	1.008	26.308	Rehabilitation	240	0.941	22.734	Developmental and Educational Psychology	278	1.046	24.076
EDUCATION & EDUCATIONAL RESEARCH	177	0.978	23.987	Cardiology and Cardiovascular Medicine	273	0.899	27.207	Mathematics (general)	235	0.921	23.655	Language and Linguistics	274	1.031	25.107
CELL BIOLOGY	174	0.961	24.949	Agricultural and Biological Sciences (miscellaneous)	269	0.886	28.093	Language and Linguistics	222	0.870	24.525	Cultural Studies	269	1.012	26.120

Table 3: Top-20 categories of the four classifications systems analysed

7.- Discussion and conclusions

A wide variety of research works have approached the problem of science classification for mapping, knowledge organisation, information retrieval or bibliometric and scientometric purposes. Up to date, some authors have commented the non-existence of a classification system which is considered an international standard in bibliometric fields (Gomez & Bordons, 1996; Archambault et al., 2011; Waltman & Van Eck, 2012) Different levels of aggregation, the distinct systems adopted for organising information as well as the degrees of specialisation or multidisciplinary of several scientific databases, are reasons enough to make difficult the construction of an international classification system for bibliometric ends. At this work, however, we have proposed a methodology to update and refine SJR journal classification system which can be applied to others through clustering and bibliometric techniques.

Another topic commonly addressed over the scientific literature on classification is the adequacy and possibility of developing automatic classification systems which avoids as far as possible the human intervention. First works on it were developed by authors as Luhn (Luhn, 1957) in Information Retrieval scope at the end of 1950s, but the interest kept holding over the 1960s (Garland, 1982) and further, especially, with the advance and development of scientific databases, bibliometric indicators, science mapping, etc., and spanning up to the present. Some of research works reviewed here have tried to avoid human intervention and they conclude it was not possible to do it completely. Waltman and Van Eck (2012) ensured that human involvement was minimized to the choice of certain values in parameters. Archambault et al. (2011) asserted that human intelligence and expertise originates more useful and flexible classification schemes as the same time as they can be considered inadequate and biased systems. They continue claiming that “From the outset, we decided that it would also be necessary to use expert judgment to finalize the work. In the end, it took substantially more work than initially expected, with alternating iterations using an algorithmic approach followed by manual fine-tuning”.

In accordance with above cited works and the own experience gained from our previous studies we thought that a classification system based on a fully automatic approach has been not possible to be conducted up to date. Furthermore, there are many choices which can be enriched by expertise and human learning. Some relevant stages emanating from automatic classification implementations such as labelling in clustering approaches are very complicated to conduct without human involvement. Decisions as labelling based on significant words or citation links, single or multiple assignments, definition of thresholds, etc., are really difficult to do. Moreover, human expertise and guidance can become very helpful during these tasks. In this work, we have avoided human intervention as much as possible, but now, we think that a mixed approach could be very realistic and convenient, particularly, after examining the final results. There is no doubt that clustering algorithms used here work fine in classifying journals. This is clearly evident when results of our tests are checked. However, once our algorithms have been run and the set of clusters have been labelled, we have found that some of them have been termed through adjacent and close categories. In some cases, these categories were coming from original tags of SJR system, and others, resulting from our text-based approach. For instance, in VOS system we obtained categories as ‘Anatomy’ or ‘Anatomy and Morphology’ covering 18 and 15 journals respectively. Also, in Louvain system we got

‘Women's Reproductive Health’ and ‘Women's and Children's Health’ categories including 10 and 28 journals respectively. Following our expertise and insight of SJR database we have considered that we can group categories covering very close knowledge domains, above all, after checking journals inside them. Then, we could obtain a VOS final category named ‘Anatomy and Morphology’ and consisting of 33 journals and a Louvain final category termed ‘Women's and Children's Health’ and embracing a total of 38 journals. These examples can be extended to approximately two tens of categories in both algorithm classifications.

After analysing and comparing clustering methods introduced in this work it should be emphasized the similitude in final results of VOS and Louvain clustering solutions in relation to facets studied, as evidence figures and tables shown throughout the text. However, the same value for resolution parameter produces a higher number of clusters in Louvain method which reveals a finer granularity than VOS one. According to the initial objectives pursued, this could be an important criterion to consider in selecting one or other algorithm. Anyway, by taking into consideration the several points analysed it is hard to decide which one of clustering algorithms analysed is suiting better to our journal classification aim. In our particular case, we consider that both VOS and Louvain clustering solutions provide a good performance in classifying SJR journals deriving from the extensive journal citation-based measures network. A particular analysis of journals assigned to clusters of specific and well-known knowledge field for authors (such as Library and Information Science) and, additionally, one or some cluster validation techniques based on expert opinions or statistical methods to validate the number and the goodness of clusters generated (Rand Index, Silhouette, Entropy, etc.) might be useful in selecting a final clustering solution.

In comparison with the original SJR journal classification, we have found an especially marked improvement regarding the distribution of journals over categories and the final number of categories available in the new solutions based on VOS and Louvain methods. The original SJR classification scheme includes 304 categories where a number of 29 have less than 10 journals assigned and the remaining categories are covering more than 10. This means that almost the totality of 18,891 journals is included in only 275 categories. Besides, journal multi-assignment is reduced and ‘Miscellaneous’ categories are removed so that, by extension, overlapping is minimized for both algorithm solutions provided. In addition, we have compared our algorithm classifications with WoS Subject Categories and we have found a certain consistency and consensus among the several classification systems both in the number of journals distributed over categories and in the number of categories appearing in top-20 categories together.

A final but not less important issue arises with regards to large and leading Multidisciplinary journals such as Science, Nature or PNAS which are not included in any cluster of size higher than 10. This might be due to their special features, with a citation pattern characterized by a vast quantity of citations emitted and received which differentiate them from the remaining journals. By looking at the whole set of clusters, including those below the threshold 10, we uncovered Science, Nature and PNAS are allocated in different singletons. Thus, it seems necessary to look for an alternative method in classifying Multidisciplinary journals. A good and reasonable choice may be to classify these journals on the basis of the papers published in them. Multidisciplinary label could be ascribed to analysed journals with papers covering a

broad spectrum of topics and overcoming a limit in journal multi-assignment previously defined.

This work corresponds to a succession of several studies (Gómez-Núñez, Vargas-Quesada, & Moya-Anegón, ('unpublished results'); Gómez-Núñez et al., 2011) concerned with optimising and boosting of SJR journal classification system and the related subsequent journal assignment. We can articulate the future research by testing new clustering algorithms and automatic techniques (factor analysis) as well as different units of analysis (papers) and measures (text-approaches). However, we do not pretend to proclaim none of these classification proposals as definitive or exclusive among them. For sure, we think it is necessary to keep working in combining several techniques and processing units of analysis in order to get a consensus from scientific community aimed to develop a new final SJR classification.

8.- Acknowledgments

The authors thank María Montserrat Posse Moure for editing the text and the reviewers for their helpful and relevant comments.

9.- Bibliographic References

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49–63. doi:10.1016/j.joi.2008.11.003

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a Multilingual , Comprehensive and Open Scientific Journal Ontology. In *E.C.M. Noyons, P. Ngulube, & J. Leta (Eds.), Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics* (pp. 66–77).

Batagelj, V., & Mrvar, A. (1997). Pajek – Program for Large Network Analysis. Retrieved from http://pajek.imfm.si/doku.php

Batagelj, V., & Mrvar, A. (2003). Density based approaches to network analysis Analysis of Reuters terror news network, 1–20. Retrieved from http://www.cs.cmu.edu/~dunja/LinkKDD2003/papers/Batagelj.pdf

Batagelj, V., & Mrvar, A. (2011). Pajek: Program for Analysis and Visualization of Large Networks: Reference Manual. *version 3.13*. Retrieved September 02, 2013, from http://pajek.imfm.si/lib/exe/fetch.php?media=dl:pajekman.pdf

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *October*(10), P10008. doi:10.1088/1742-5468/2008/10/P10008

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., … Boyack, K. W. (2012). Design and update of a classification system: the UCSD map of science. *PloS one*, *7*(7), e39464. doi:10.1371/journal.pone.0039464

Boyack, K. W., & Klavans, R. (2010). Co-Citation Analysis , Bibliographic Coupling , and Direct Citation : Which Citation Approach Represents the Research Front Most Accurately?, *61*(12), 2389–2404. doi:10.1002/asi

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Börner, K. (2011). Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PloS one*, *6*(3), e18029. doi:10.1371/journal.pone.0018029

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*(9), 1759–1767. doi:10.1002/asi.22896

Cantos-Mateos, G., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Zulueta, M. A. (2012). Stem cell research: bibliometric analysis of main research areas through KeyWords Plus. *Aslib Proceedings*, *64*(6), 561–590. doi:10.1108/00012531211281698

Chang, Y. F., & Chen, C. (2011). Classification and Visualization of the Social Science Network by the Minimum Span Clustering Method, *62*(12), 2404–2413. doi:10.1002/asi

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. doi:10.1007/s11192-009-0146-3

Elsevier. (2004). Scopus. Retrieved May 30, 2013, from http://www.scopus.com/home.url

Garland, K. (1982). An experiment in automatic hierarchical document classification. *Information Processing & Management*, *19*(3), 113–120.

Glänzel, W. (2012). Bibliometric methods for detecting and analysing emerging research topics. *El Profesional de la Informacion*, *21*(2), 194–201. doi:10.3145/epi.2012.mar.11

Gomez, I., & Bordons, M. (1996). Coping with the problem of subject classification diversity, *35*(2), 223–235.

Gómez-Núñez, A. J., Vargas-Quesada, B., & Moya-Anegón, F. ('unpublished results'). A new SJR journal classification through a combination of citation measures.

Gómez-Núñez, A. J., Vargas-Quesada, B., Moya-Anegón, F., & Glänzel, W. (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, *89*(3), 741–758. doi:10.1007/s11192-011-0485-8

Jacsó, P. (2013). The need for end-user customization of the journal-sets of the subject categories in the SCImago Journal Ranking database for more appropriate league lists. A case study for the Library & Information Science field. *El Profesional de la Informacion*, *22*(5), 459–473. Retrieved from http://dx.doi.org/10.3145/epi.2013.sep.12

Janssens, F., Zhang, L., Moor, B. De, & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683–702. doi:10.1016/j.ipm.2009.06.003

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476. doi:10.1002/asi.20991

Leydesdorff, L., Hammarfelt, B., & Salah, A. (2011). The Structure of the Arts & Humanities Citation Index : A Mapping on the Basis of Aggregated Citations Among 1 , 157 Journals, *62*(12), 2414–2426. doi:10.1002/asi

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from Web-of-Science data. *Journal of Informetrics*, *6*(2), 318–332. doi:10.1016/j.joi.2011.11.003

Liu, G.-Y., Hu, J.-M., & Wang, H.-L. (2011). A co-word analysis of digital library field in China. *Scientometrics*, *91*(1), 203–217. doi:10.1007/s11192-011-0586-4

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, *1*(4), 309–317. doi:10.1147/rd.14.0309

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. doi:10.1103/PhysRevE.69.026113

Nooy, W. de, Mrvar, A., & Batagelj, V. (2012). *Exploratory Social Network Analysis with Pajek, Revised and Expanded 2nd Edition. Cambridge University Press.* (Revised an.). Cambridge [etc.]: Cambridge University Press.

Persson, O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, *4*(3), 415–422. doi:10.1016/j.joi.2010.03.006

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals, *53*(13), 1113–1119.

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823–1835. doi:10.1002/asi.21086

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science Overlay Maps : A New Tool for Research Policy and Library Management. *Journal of the American Society for Information Science and Technology*, *61*(9), 1871–1887. doi:10.1002/asi

Rousseau, R., & Zuccala, A. (2004). A classification of author co-citations: Definitions and search strategies. *Journal of the American Society for Information Science and Technology*, *55*(6), 513–529. doi:10.1002/asi.10401

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Scimago. (2007). Scimago Journal & Country Rank (SJR). Retrieved April 15, 2011, from http://www.scimagojr.com/

Sinclair, S., & Rockwell, G. (2009). Voyeur Tools: See Through Your Texts. *Hermeneuti.ca – The Rhetoric of Text Analysis*. Retrieved September 01, 2013, from http://hermeneuti.ca/

Small, H. (1999). Visualizing Science by Citation Mapping, *50*(1973), 799–813.

Thijs, B., Zhang, L., & Glänzel, W. (2013). Bibliographic Coupling and Hierarchical Clustering for the validation and improvement of subject-classification schemes. In *14th International Conference on Scientometrics and Informetrics (15–19 July, 2013), Vienna (Austria)* (pp. 237–250). Viena: International Society of Scientometrics and Informetrics. Retrieved from http://www.mtakszi.hu/kszi_aktak/

Thomson Reuters. (2009). ISI Web of Knowledge. Retrieved September 01, 2013, from http://thomsonreuters.com/web-of-science/

Vargas-Quesada, B., & Moya-Anegón, F. (2007). *Visualizing the structure of science*. New York: Springer.

Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. doi:10.1016/j.joi.2010.07.002

Waltman, L., & Van Eck, N. J. (2012). A New Methodology for Constructing a Publication-Level Classification System of Science, *63*(12), 2378–2392. doi:10.1002/asi

Zhang, L., Glänzel, W., & Liang, L. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, *81*(3), 821–838. doi:10.1007/s11192-008-2245-y

Zhang, L., Janssens, F., Liang, L., & Glänzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics*, *82*(3), 687–706. doi:10.1007/s11192-010-0180-1

Pre-print of: ***Journal of the Association for Information Science and Technology*****, 2015 [early view]. http://dx.doi.org/10.1002/asi.23370**

UPDATING SCIMAGO JOURNAL & COUNTRY RANK CLASSIFICATION: A NEW APPROACH FROM WARD'S CLUSTERING AND ALTERNATIVE COMBINATION OF CITATION MEASURES

Antonio J. Gómez-Núñez[1 *], Benjamín Vargas-Quesada[2], Félix Moya-Anegón[3]

[1] CSIC – SCImago Research Group Associated Unit. Faculty of Communication and Documentation, Campus de Cartuja s/n, 18071, Granada, SPAIN.
Phone: +34 958240923 / * Corresponding Author: anxusgo@gmail.com✉

[2] University of Granada, Department of Information and Communication. Faculty of Communication and Documentation, Campus de Cartuja s/n, 18071, Granada, SPAIN.
Phone: +34 958240923 / Mail: benjamin@ugr.es

[3] CSIC – Institute of Public Goods and Policies. Albasanz 26-28, 28037 Madrid, SPAIN.
Mail: felix.demoya@csic.es

Abstract

This study introduces a new proposal to refine classification of Scimago Journal & Country Rank (SJR) platform by using clustering techniques and an alternative combination of citation measures from an initial 18,891 SJR journal network. Thus, a journal-journal matrix including fractionalized values of direct citation, co-citation and coupling at once was symmetrized by *cosine similarity* and later transformed into distances before performing clustering.

The results provided a new cluster-based subject structure comprising 290 clusters which emerge from executing Ward's clustering in two phases and a mixed labeling procedure based on *tf-idf* scores of original SJR category tags and significant words extracted from journal titles. A total of 13,716 SJR journals were classified through this new cluster-based scheme generated.

Although over 5,000 journals were omitted in the classification process, the method produced a consistent classification with a balanced structure of coherent and well-defined clusters, a moderated multi-assignment of journals and a softer concentration of journals over clusters than in SJR original categories. New subject disciplines as 'Nanoscience & Nanotechnology' or 'Social Work' were also detected evidencing a good performance of our approach both by refining journal classification and updating the subject classification structure.

Keywords

Journal classification ; Scimago Journal & Country Rank ; Ward hierarchical clustering ; Citation-based networks ; Bibliometrics

Introduction

Classification schemes for bibliometric application increased in the wake of the remarkable growth and development of scientific disciplines and research topics. This issue is closely related to the proliferation of scientific information described by Price (1963) in his book entitled *Little Science, Big Science*. According to Yagi, Badash, and Beaver (1996) Price's interest in measuring the growth of science was fostered by the emergence of the first large multidisciplinary citation database, namely, the Science Citation Index (SCI) at the beginning of 1960s. Since then, databases started to arise as the main service for recording bibliographic data of scholarly literature mainly published in scientific journal articles, conference proceedings, books, etc.

However, all information covered by databases has to be maintained and organized in an adequate way in order to facilitate tasks linked to information seeking and retrieval, but also to assist information specialists in quantifying various aspects arising from the analysis of science systems and research performance. Bibliometric studies devoted to these issues have helped science professionals in making decisions on funding, in science policy and research management. In order to be able to offer consistent and reliable results, correct delineation and definition of subject disciplines is indispensible. Gómez, Bordons, Fernández, and Méndez. (1996) pointed to the lack of unified classification criteria among the different databases as well as to the absence of standardization of subject classifications in those as one of the most crucial problems in obtaining appropriate subject delineation for conducting sound, reliable and reproducible bibliometric and scientometric studies.

Indeed, the fact of organizing and representing the structure of scientific knowledge in an appropriate system of disciplines and subject categories is not only interesting from the viewpoint of research evaluation but from other perspectives as well. Thus, several authors dealing with computerized visualization methods attempted to map the structure of scientific knowledge using information extracted from the large multidisciplinary abstract and citation indexes, above all, Thomson Reuters' Web of Knowledge (Small & Garfield 1985; Bassecoulard & Zitt, 1999; Boyack, Klavans & Börner, 2005; Moya-Anegón et al., 2004; Moya-Anegón et al., 2007; Janssens, Zhang, De Moor & Glänzel, 2009; Leydesdorff & Rafols, 2009; Zhang, Janssens, Liang & Glänzel, 2010) and Elsevier's Scopus (Klavans & Boyack, 2007; Klavans & Boyack, 2009; Leydessdorff, Moya-Anegón & Guerrero-Bote, 2010). These two databases are considered the most prestigious and reliable multidisciplinary bibliographic information sources for the scientific community up to the present. Both databases assign documents through the journals, in which they have been published, using a similar two-level disciplinary classification scheme consisting of subject areas and subject categories.

At this point, it should be remembered that science and research are ever-changing phenomena. New topics are emerging for different intra-scientific, social, economic and further reasons while other topics might gradually diminish and disappear or just change their thematic focus (cf. Glänzel & Thijs, 2012), respectively. An obvious consequence of these processes is that subject classification schemes should mirror and reflect evolution. Due to this necessity to adapt to the shifts of science and research, classification needs to be regularly updated and improved. This could result in some changes in the structure of the subject

classification scheme, for instance, in introducing new categories, removing old and "extinct" ones, or merging or splitting up some of them. Leydesdorff (2002) considered this problem and propose to refresh subject classification schemes by analyzing structural links among journals covered in Science Citation Index, what he called "dynamic and evolutionary updates".

A vast number of methods and techniques have been developed to categorize and organize information stored in both specialized and multidisciplinary bibliographic databases. Among the different proposals we found solutions based on expert assessment, intellectual, heuristic and pragmatic approaches as well as various statistical and computerized methods and techniques based on social-network analysis, reference analysis, subject filtering, specialized queries, and so on. All these proposals have been applied to different units of analysis, that is, papers, journals or even subject categories themselves and, of course, using different units of measures based on lexical components (keywords or terms extracted from title, abstract, authors' addresses, or the full text of the document), citation links, or a combination of both.

Related works

The exponential growth of information aforementioned involved simultaneous implications as launching of databases to cope with this flood of information. This was also boosted by the breathtaking development of computer science, information technology and electronic communication, which allowed the rise of digital formats and expedited the diffusion of information around the world. This technological development motivated a considerable number of researchers to start exploring the possibility of automatic classification of information recorded in databases, notably, for information retrieval purposes. For this goal different computerized techniques were used. Among these, clustering was one of the most popular methods. As far back 70s, Cormack (1971) published an exhaustive and helpful bibliography study of subject classification topics focusing on definitions of different similarity measures and the theoretical framework of alternative clustering techniques.

Classification for Information Retrieval Purposes

In information retrieval, automatic classification was mainly used to facilitate retrieval and to improve its efficiency. In the first tests, the set of analyzed items were rather small. However, the advances and development of computer science (increasing CPU power and speed, storage capacity and improving algorithms) allowed to process large data sets. Price and Schiminovich (1968) ventured into automatic classification scheme generation through a clustering experiment based on bibliographic coupling of 240 theoretical documents in high-energy physics. Later on, Schiminovich (1971) published a new study in which an algorithm for the automatic detection of related document clusters based on bibliographic links was determined. He proposed his *bibliographic pattern discovery algorithm*, as a tool for automatic document classification and retrieval, offering good balance of recall and precision. By using a modified release of Schiminovich's algorithm, Bichteler and Parsons (1974) proposed an automatic classification through citations extracted from a *triggering file* of documents associated via bibliographic information. Croft (1980) developed a clustering method to retrieve relevant document clusters matching with queries. Queries were thus grouped into different clusters of similar documents using probability estimates. Van Rijsbergen carried out single-link clustering

on different samples in order to establish a hierarchical cluster classification to improve efficiency of document retrieval (van Rijsbergen, 1974; van Rijsbergen & Croft, 1975). An interesting and broad review of research literature on clustering approaches used in document retrieval was published by Willett (1988), who basically targeted on hierarchical agglomerative clustering methods.

Bibliometric/Scientometric classification schemes

In bibliometrics/scientometrics the application of citation-based methods in building measures for clustering techniques and classification algorithms was a very common research point in the use of SCI data. Small and Griffith (1974) applied a single-link clustering method using co-citation strength on a set of 1832 SCI highly cited documents to detect and map scientific specialties. Garfield (1975) reported that SCI categories were, in part, "algorithmically identified by the simplest clustering techniques". He tested clustering on the basis of citation links among SCI documents as to set an automatic classification system that enabled updating and modifying classification scheme periodically (Garfield, Malin & Small, 1974). Small and other scientists tried to cluster SCI database using co-citation links as the underlying measure. In a first study, Small and Sweeney (1985) exposed two method enhancements, namely, *fractional citation counting* to weight citation links according to the number of emitted citations, and *variable level clustering* to define the maximum size of the final clusters to be obtained. In a related second work (Small, Sweeney & Greenlee, 1985) the last methodological improvement, particularly, *iterative clustering of clusters* to map SCI fields in terms of scientific literature covered by this database was introduced. Glenisson, Glänzel, and Persson (2005) prepared a pilot study relying on 19 papers published in a special issue of the journal *Scientometrics,* where they combined text-based and bibliometric methods to test the performance of hierarchical clustering of full-text versus abstract-title-keywords and compared their results with the topic structure created by the guest-editors.

Matrices and networks of journal-level aggregation

Hitherto, we have talked about cluster analysis at the document level. However, tests regarding cluster algorithms and methods on citation measures were developed not only on papers sets and, very early journals have become the object of study in assignment and classification. The reason is that documents can indirectly be assigned to subjects through the scope or classification of journals covered in multidisciplinary databases thus allowing subject delineation at the field and discipline level to prepare accurate bibliometric and scientometric meso- and macro-level studies. Narin (1976) was a pioneer in proposing article classification through the assignment of journals, in which they have been published, to subject categories. Thereby, Carpenter and Narin (1973) performed clustering on cross-citation data from 288 SCI journals in physics, chemistry and molecular biology field looking for a more precise level of journal classification than disciplines. Bassecoulard and Zitt (1999) stated that "journals remain the main substrate for macro-level classifications", and on this basis, they presented an automatic multi-level classification of scientific journals conducting a hierarchical clustering on a journal-journal cross-citation matrix. They obtained 141 specialties and 32 sub-disciplines that were mapped and compared with other classifications schemes, namely, the CHI subfields and ISI subject categories showing a large degree of similarity and compatibility with the latter

one. Chen (2008) used the *affinity propagation* cluster algorithm on a journal-journal citation links from Thomson Reuters' Journal Citation Reports (JCR) in order to achieve an automatic classification of SCI and SSCI journals. The algorithm was applied to distances among journals by using a cutoff parameter to constrain a maximal distance that permitted to determine different levels of resolution in the final classification schemes. Rafols and Leydesdorff (2009) made a comparative study of two content-based and two algorithmic decomposition of aggregated journal-journal citation matrix based on 7,611 journals from 2006 volume of the JCR. Results evidenced a more homogeneous and overlapped classification from content-based decomposition and a limited matching between similar categories of different classification schemes. However, maps depicted and relied on four classification schemes showed a good fit and correspondence. Janssens et al. (2009) developed a hybrid clustering method based on text and citation analysis of about 8,300 journals covered by Thomson Reuters' Web of Science (WoS) in order to improve WoS journal classifications and to get a comparable classification scheme to compare and validate this versus existing "intellectual" subject classification schemes. Here, Ward's hierarchical clustering was used. Later, Zhang et al. (2010) conducted a journal cross-citation analysis on a similar WoS journal set to improve a journal-based subject classification scheme developed earlier by Glänzel and Schubert (2003). They used a combination of statistical, computerized and bibliometric techniques including Ward's hierarchical method for clustering journals, text-mining to label obtained clusters, and several indicators as journal link strength, journal entropies or PageRank algorithm to determine important leading journals within the clusters.

In the present study, we focus on cluster analysis of Scopus journals included in the Scimago Journal & Country Rank (SJR) platform (Scimago, 2007) using a combination of three citation measures, namely, direct citation, co-citation and bibliographic coupling. We follow the suggestion by Persson (2010) to combine these citation measures in a raw and in a normalized way (Weighted Direct Citation and Normalized Weighted Direct Citation respectively).

Material

We retrieved all data from the SJR database. A set of 18,891 Scopus journals included in the SJR platform and covering a period of two years (2009-2010) were captured. Citations were collected and counted at the level of individual papers published in a two-year period and assigned to the journals in which they appeared. No threshold was set regarding citations and document type. Only cited references from 2000 on were taken into account while journal self-citations were discarded since this type of citations tend to have a strong effect on similarity measures without contributing to the actual classification issue.

Methods

Citation-based measures calculation and integration

In the first step a pair-wise journal list containing direct citation, bibliographic coupling and co-citation journal networks was built. These lists represent journal relationships on the basis of raw values expressing the strength of their relations. The three citation-based measures were calculated at the document level and then aggregated to journals. For co-citation and bibliographic coupling calculation, references co-occurring were counted only once per paper

following the binary counting proposed by Rousseau and Zuccala (2004) and avoiding what Vargas-Quesada and Moya-Anegón (2007) named *latent co-citation*. Moreover, fractional counting was applied for three citation measures by considering the total number of links given and received by journals in calculating relation strength. In this way, the power of attraction exerted by certain journals, for instance, with a high quantity of citations or pertaining to research fields with a considerable influence could be offset.

Then, we integrated the three lists of journals covering fractionalized direct citation, co-citation and bibliographic coupling in a new one according to Persson's approach to identify research themes through this new combined measure which he named *normalized weighted direct citation (NWDC)* links (Persson, 2010). However, we limited the final list to those journal pairs for which all the three citation-based measures were present. In this way, we took into consideration the Persson's idea of those citation links that are not sharing references and are not being cited together might be out of topic. Hence, the integration of the normalized direct citation, co-citation and bibliographic coupling lists was executed following the next rule:

If $x > 0$ AND $y > 0$ AND $z > 0$ Then $R = x + y + z$
But
If $x = 0$ OR $y = 0$ OR $z = 0$ Then $R = 0$

Where x = direct citation, y = co-citation, z = bibliographic coupling and R = result.

Consequently, the three normalized measures were combined by summing up the values on the condition of all of them are non-zero. After that, the network list consisted of 16,258 citing and 15,266 cited journals.

Clustering performance

The resulting asymmetrical journal-journal list was finally converted into a journal-journal symmetrical matrix by applying cosine similarity on the cited side of the matrix. Due to the large citation network, which includes several thousand journals from different science fields characterized by different citation behaviors, we found an extremely sparse matrix with only 1.81% non-zero values. Cosine let us to make the matrix symmetrical and set similarities among journal vectors at once. Nevertheless, similarities were then transformed into distances according to the *1–cosine* formula and Ward's hierarchical clustering was conducted at last. All these steps were executed in "R" statistical software with the 'lsa' package (Wild, 2011) for Latent Semantic Analysis aimed at calculating cosine similarities.

We executed several tests based on Ward's hierarchical agglomerative clustering in order to determine the best solution adapting to our main goal: to refine journal assignment and improve the SJR journal classification system. Solutions giving 200 to 250 clusters were analyzed and, finally, by considering the number of joined clusters over the different partitions generated from down to top in the dendrogram we found that the solution providing 232 clusters was a good and tailored choice to our final classification objective. This point was reinforced by the fact that the higher the number of clusters generated in each solution, the higher the number of small clusters with a poorer relatedness among the journals comprised.

Labelling clusters

The definitive number of clusters given was labeled by adopting tags from the set of original SJR categories. By doing so, we would facilitate one of the most prevalent problems in clustering analysis such as cluster labeling. Thus, for each cluster we calculated the number of links from journals to former SJR categories and, subsequently, we transformed frequencies into *tf-idf* scores by an adaptation of Salton and Buckley (1988) formula:

$w_{i,j} = catf_{i,j} \times Log\ (N / cluf_i)$

Where $w_{i,j}$ = total weighted score; $catf_{i,j}$ = raw frequency of category 'i' into cluster 'j'; N = total number of clusters; and $cluf_i$ = number of clusters containing category 'i'

At the same time, we converted the number of links into percentages. Finally, we ranked journals according to *tf-idf* scores and selected those category tags amounting at least 33% of the links which would be designed to delineate the subject of clusters accurately. During the labeling process SJR *'Miscellaneous'* and *'Multidisciplinary'* category tags were discarded. However, using this approach we found some clusters thematically delimited by just the same tags. In these cases, we re-labeled these clusters using significant words extracted from journal titles. Through this combined approach we were keeping core, well-defined and representative categories from the original SJR classification system according to the citation links and detecting emerging categories through the alternative text component at once. In closing, it is worthy to remark that journal multiple assignments were a consequence of cluster labeling process rather than Ward's clustering method, which is hard clustering methods that performs assigning only one journal per cluster.

Assessment of the new classification

In a last phase, we proceeded to assess the performance of the new SJR journal classification by contrasting it with the original SJR classification system. To do so, we prepared a series of indicators applicable to both classification systems including: (1) *number of categories*, (2) *mean number of journals per category*, (3) *mean number of categories per journal*, (4) *overlapping percentage*, (5) *journals changing their classification*, (6) *journals adding categories* and (7) *journals losing categories*. Apart from these indicators we examined the (8) *distributions of journals over categories* given in both systems in order to display some points such as the distribution curve or the aggregation of journals in some particular categories. To make a reliable and reasonable comparison, only the set of journals included in both classification systems were taken for indicators (5), (6) and (7).

Results

Firstly, we examined the various partitions generated providing 200-250 clusters and could observe that everyone had a common factor: the appearance of two especially large clusters covering journals of the areas of 'Social Sciences & Humanities' (cluster #2) and 'Medicine' (cluster #12). Generally, Ward's method is featured by this phenomenon and one or various

large clusters can appear when is executed. Everitt, Landau, Leese, and Stahl (2011) stated that "In standard agglomerative or polythetic divisive clustering, partitions are achieved by selecting one of the solutions in the nested sequence of clusterings that comprise the hierarchy, equivalent to cutting a dendrogram at a particular height (sometimes termed the best cut)". Based on their idea that "Large changes in fusion levels are taken to indicate the best cut" we analyzed the down-top fusions of clusters occurred during the clustering performance, and found that the partition supplying 232, which coincide with the lowest number of joined clusters from the whole set, were appropriate and fitted to our classification goal. Moreover, this argument is underpinned by the fact that the higher the number of clusters generated, the wider the set of small-size clusters obtained resulting less thematically consistent and delineated.

The overall hierarchical structure derived from Ward's clustering method is represented through the dendrogram in figure 1. Due to space constrains and aimed at making clearer and easier its analysis and visualization, we cut the dendrogram at the height of 50 in order to keep only its topside. In this manner we got a simplified version of the hierarchical structure where four core subject areas of Scopus database (Elsevier, 2014), i.e. 'Health Sciences', 'Life Sciences', 'Physical Sciences' and 'Social Sciences' have been identified. In addition, some smaller journal aggregations corresponding to subject categories included in these four large areas are displayed in other branches of the dendrogram. In any case, there seems to be a good deal of coherence among adjacent branches which, at the same time, are in some consonance with Scopus classification at this high level of aggregation. Thus, the dendrogram shows two principal branches. The first one on the left side is an isolated branch covering the area of 'Computer Sciences and Information Systems'. The second principal branch on the right side is much more extensive and comprises of various hierarchical levels. A first main subdivision of this branch is dealing with *Biomedical Sciences*. Thus, from left to right, we can find a branch referring to 'Cellular and Molecular Biology' together with two other branches on 'Medical Research' and 'Health Sciences'. On its side, the second main subdivision on the right side presents seven branches. The first and second branches respectively cover a specific aggregation of journals on 'Optics' as a component of a subsequent higher aggregation of journals on 'Physical Sciences'. Similarly, the third and fourth branches include journals on 'Zoology' as a fraction of a higher group of journals on 'Life Sciences'. The fifth branch shows a separated group of journals on 'Mathematics' which is just followed by two new branches consisting of journals on 'Engineering' and, finally, a wide set of journals from 'Social Sciences and Humanities' areas.

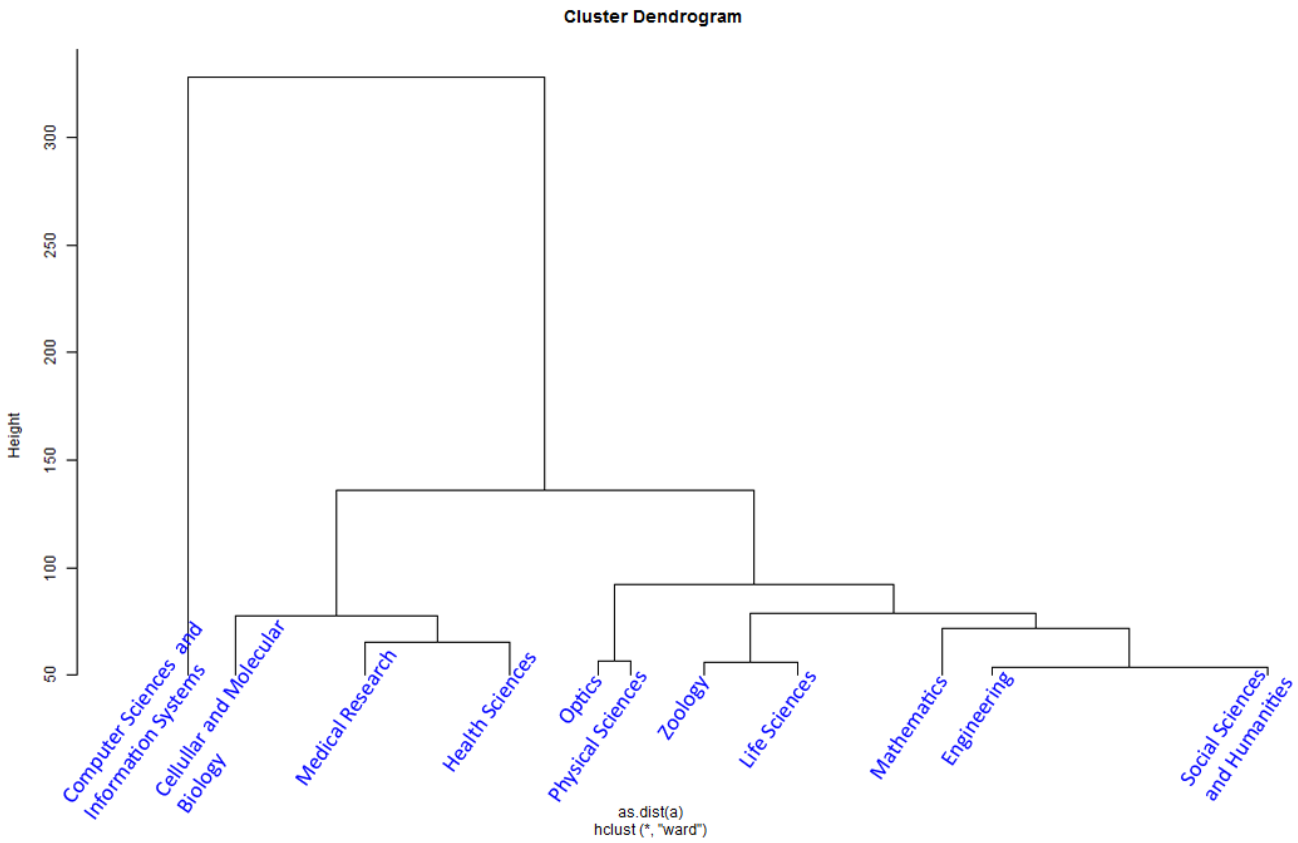


Figure 1: Dendrogram of Ward's SJR journal clustering

As aforesaid, the two large clusters on 'Social Sciences & Humanities' and 'Medicine' are present throughout the distinct solutions generated by the method. In spite of being a frequent issue in Ward's method performance, this fact is not satisfactory to our classification objective since, similarly to relatively small clusters, high-size clusters do not constitute well-defined and delimited subject groups of journals. Besides, the distribution of journals over the 232 subject clusters categories produced a great aggregation of journals in only two categories producing a far skewed distribution. To solve this shortcoming we decided to repeat the full methodological procedure on the sub-matrices referring to clusters #2 and #12 of our 232-cluster solution. For both clusters this implied (1) to extract raw data of the three citation-based measure networks for each journal included only in these clusters, (2) to calculate fractional counting, (3) to combine the measures following the rule which binds to the three citation-based measures to appear together in the matrix, (4) to apply cosine similarity on cited side of the matrix, (5) to execute Ward's clustering method, (6) to determine a good solution by analyzing large changes in fusion levels to set the best-cut and, finally, (7) to label the clusters.

After executing the whole process, cluster #2 originally formed by 1788 journals was then reduced to 804 which were spread over 35 new clusters. On its side, from the initial 951 journals in cluster #12 only 385 left spread over a set of 25 clusters. In summarizing, the initial network of 18,891 SJR journals was pruned to a final number of 13,716 journals after performing the second phase of our methodological procedure while the final number of subject clusters raised to 290 resulting from the 232 original clusters of the first clustering

phase minus the clusters #2 and #12 then divided into 35 and 25 respectively. The new final classification originated as a consequence of our method is accessible at the following link: http://www.ugr.es/local/benjamin/new-sjr-classification.pdf

Table 1 presents some relevant indicators comparing the original SJR classification to the new SJR classification system generated.

	ORIGINAL SJR CLASSIFICATION	NEW SJR CLASSIFICATION
Number of classified journals	18,891	13,716
Number of categories	308	298
Mean number of journals per category	61.33	46.03
Mean number of categories per journal	1.61	1.42
Overlapping percentage	60.73	42.26

Table 1: Some relevant indicators about the original and new SJR classification systems

The first indicator refers to the set of journals included in each system. Original SJR covers a total of 18,891 journals that were then decreased to 13,716 in the new system because of the combination of citation-based measures used in the 2-phase clustering applied. A more interesting data is related to the final set of active categories in both systems, the original SJR has 308 categories while the new cluster derived system amounts a total of 298 categories. However, a relevant issue derived from the comparison is that only 159 categories are being shared by both systems. Therefore, a total of 139 (Annex I) have been generated as a consequence of the method applied, especially, by the labeling process executed. It is important to clarify that the number of clusters derived from our method (290) does not agree with the number of categories specified in table 1 (298) due to the multiple assignment of category tags occurred in some clusters. For instance, cluster #148 was labeled as 'Cardiology and Cardiovascular Medicine' + 'Internal Medicine' category tags, while cluster #210 was assigned to 'Cardiology and Cardiovascular Medicine' and #175 'Endocrinology' + 'Cardiology and Cardiovascular Medicine' categories. The several possible combinations and the distinct number of category tags assigned to the set of clusters is, therefore, the only reason behind the difference in the number of categories included in table 1 and the number of clusters given by Ward's clustering method.

Admittedly the set of journals of original SJR overcomes in more than 5,000 journals to the new one by contrast the number of categories in both original and new SJR systems are quite similar, amounting 308 and 298 respectively. This fact could explain the much better results of new classification system regarding the indicator on 'Mean number of journals per category', and, to a lesser extent, on indicators on 'Mean number of categories per journal' and 'Overlapping percentage'. However, this subset of over 5,000 journals is characterized by having a scant influence in the whole journal network of combined citation-based measures and very low interaction among them. Therefore, the better outcomes of new SJR classification system seem not merely a consequence of the reduction of the number of journals but rather the good results derived from the methodology proposed. This argument can be contrasted by comparing the results of the above mentioned indicators only on the 13,716 journals matching in both classification systems (table 2).

	SJR CLASSIFICATION REDUCED VERSION (13,716 JOURNALS)	NEW SJR CLASSIFICATION
Number of classified journals	13,716	13,716
Number of categories	297	298
Mean number of journals per category	46.18	46.03
Mean number of categories per journal	1.62	1.42
Overlapping percentage	62.18	42.26

Table 2: Some relevant indicators about the 13,716 journals matching in the original and new SJR classification systems

The *'Mean number of categories per journal'* and *'Overlapping percentage'* are complementary indicators which differs only in the way to express their values, i.e. as decimal and percentages. The overlapping percentage was calculated using the next formula:

B - A / A * 100

Where A = journals covered by the system (or in other words, the set of journals under consideration); and B = journal multi-assignment (set of classified journals including multi-assigned journals, that is, assigned to more than one category).

A very significant issue not covered in tables 1 and 2 is in connection with the number of journals suffering changes in their original classification. From the full set of 13,716 journals under study, a total of 1,988 altered their classification by changing one or several of their assigned categories. Apart from these changes, many other journals suffered either increases or decreases in the number of categories they covered. Thus, 2,426 journals adding at least one category in the new SJR classification system while a number of 3,951 lost categories with respect to the original SJR classification. This means that taking into account the journals changing their original classification and the journals adding/losing categories the percentage of SJR journals modifying their category assignment from the original to the new system amounts 60.99% which is a really high figure.

Linking with changes in the category assignment of journals and the overlapping indicators of table 1, the figure 2 represents the multi-assignment of journals in the original SJR system and new SJR one. In this point, also the new SJR system returned better results, by decreasing the number of multi-assgigned journals. Thus, in the new SJR system, over 72% of journals were assigned to only 1 category, almost 19% to 2 categories, a little bit more of 5% to 3 categories, and approximately the same percentage (around 2%) of journals allocated to 4 and 5 categories. There are not journals with more than 5 categories in this system. On its side, the reduced SJR system amounted 63.7% of journals in one category, 20.6% of journals were assigned to 2 categories, 9.9% to 3 categories, almost 4% to 4 categories whereas over 1.2% were allocated to 5 categories. In contrast to the new system, here a residual quantity of journals (around 0.7%) was ascribed to more than 5 categories.

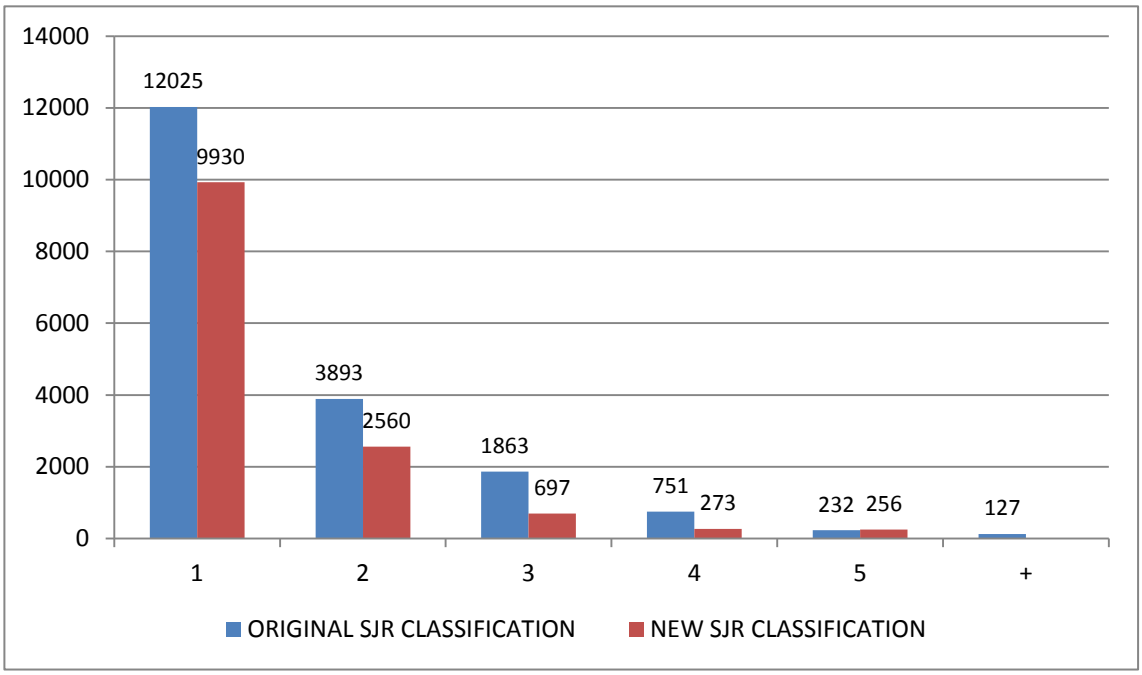


Figure 2: Journal multi-assignment in the original and new SJR classification systems

Another important issue to be analyzed is the distribution of journals over the 298 categories obtained from the new journal classification. Figure 3 shows the distributions of the original and new SJR systems. A simple glance is enough to uncover a much more balanced and softer distribution of journals over categories in the latter. The original system, basically, revealed a high skew in the first part of the distribution where 'Medicine (miscellaneous)' category appears amounting 5.2% of the total set of journals classified. However, although the initial part of the distribution is too flatter for the new classification system, approximately on the middle point of the distribution both systems holds quite similar and parallel.

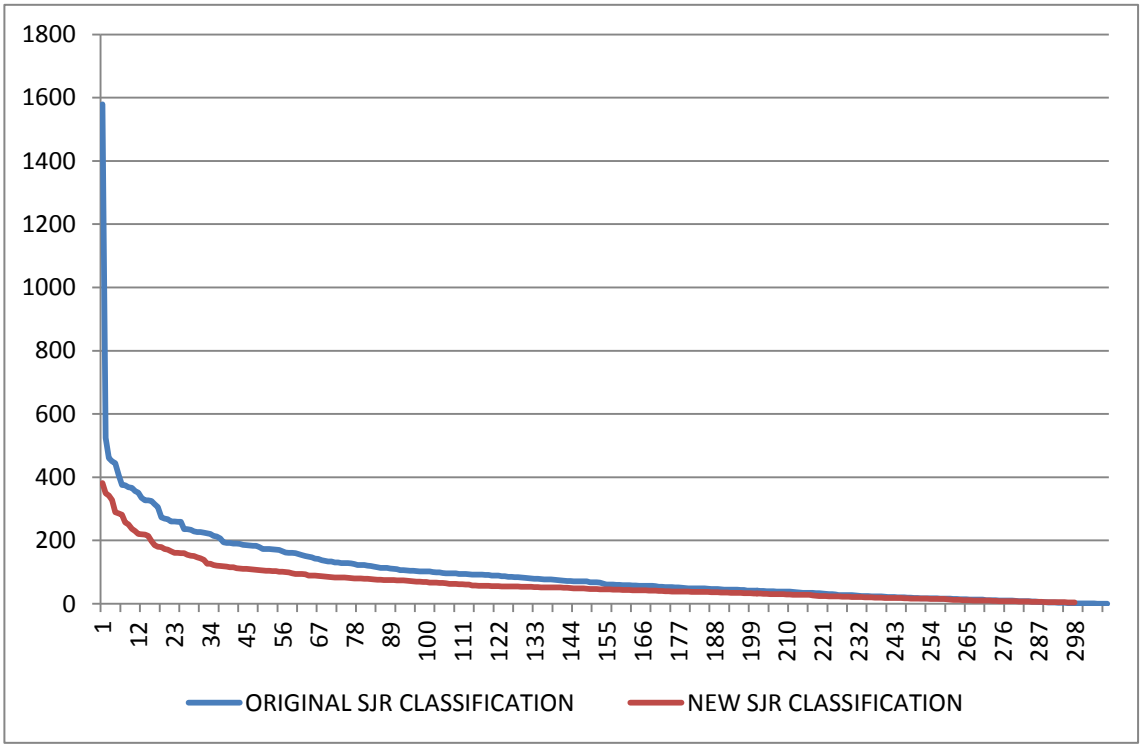


Figure 3: Distribution of journals over categories in the original and new SJR classification systems

Figure 3 is supplemented by the table 3, which reflects the top-20 categories in the ranking of the distribution of journals over categories in both systems. By observing the number of journals per category in each system it can be seen that as a general rule, the original SJR classification offers higher concentrations of journals per category than the new classification system. According to figure 3, this concentration is extremely notable in the first category of the reduced SJR classification where 'Medicine (miscellaneous)' amount 1579 journals whereas in the new one only 381 journals were included by the category 'Social Studies'. Despite the high number of new categories given by the new SJR classification system, a total of 4 categories are shared in the top-20 table, namely, 'Electrical and Electronic Engineering', 'Sociology and Political Science', 'Public Health, Environmental and Occupational Health' and 'Psychiatry and Mental Health'. This point indicates certain coherence and concordance between both classifications systems even more when the set of 'Miscellaneous' categories are not being used in the new classification system.

REDUCED ORIGINAL SJR CLASSIFICATION SYSTEM				**NEW SJR CLASSIFICATION SYSTEM**			
Category	**Num. of Journals**	**%**	**Cum. %**	**Category**	**Num. of Journals**	**%**	**Cum. %**
Medicine (miscellaneous)	1579	5.20	5.20	Social Studies	381	1.95	1.95
Education	524	1.73	6.93	Electrical and Electronic Engineering	349	1.79	3.74
Sociology and Political Science	460	1.52	8.44	Computer & Information Systems Engineering	342	1.75	5.49
Geography, Planning and Development	450	1.48	9.92	Mechanical Engineering	328	1.68	7.17
History	444	1.46	11.39	Immunology	289	1.48	8.66
Electrical and Electronic Engineering	406	1.34	12.72	Cancer Research	285	1.46	10.12
Cultural Studies	375	1.24	13.96	Oncology	281	1.44	11.56
Social Sciences (miscellaneous)	374	1.23	15.19	Neurology (clinical)	257	1.32	12.87
Economics and Econometrics	368	1.21	16.40	Cell Biology	251	1.29	14.16
Literature and Literary Theory	366	1.21	17.61	Orthopedics and Sports Medicine	237	1.21	15.37
Engineering (miscellaneous)	356	1.17	18.78	Computer Systems	230	1.18	16.55
Psychology (miscellaneous)	351	1.16	19.94	Sociology and Political Science	221	1.13	17.69
Public Health, Environmental and Occupational Health	335	1.10	21.04	Psychiatry and Mental Health	220	1.13	18.81
Plant Science	327	1.08	22.12	Development Studies	219	1.12	19.94
Language and Linguistics	327	1.08	23.19	Public Health, Environmental and Occupational Health	215	1.10	21.04

Psychiatry and Mental Health	325	1.07	24.26	Strategy and Management	198	1.01	22.05
Animal Science and Zoology	315	1.04	25.30	Applied Mathematics	185	0.95	23.00
Mathematics (miscellaneous)	306	1.01	26.31	Molecular Medicine	180	0.92	23.92
Cardiology and Cardiovascular Medicine	273	0.90	27.21	Biochemistry	179	0.92	24.84
Agricultural and Biological Sciences (miscellaneous)	269	0.89	28.09	Genetics	173	0.89	25.73

Table3: Ranking of the top-20 categories in the original and new SJR classification systems

Up to now we have contrasted the results of the new SJR classification with the original SJR classification system in order to reveal the improvements obtained through our proposal. Nevertheless, it might also be interesting to compare the results generated by this methodology with those coming from a previous work based on a different combination of citation measures and distinct clustering methods, namely, Louvain and VOS community detection algorithms (Gómez-Núñez, Batagelj, Vargas-Quesada, Moya-Anegón & Chinchilla-Rodríguez, 2014). To this end, we have introduced the table 4 collecting the same indicators exposed in tables 1 and 2. Furthermore, indicators referring to journals changing, adding or losing categories in relation to the original SJR classification have also been added.

	NEW SJR CLASSIFICATION	LOUVAIN 18	VOS 15
Number of classified journals	13716	17287	17729
Number of categories	298	272	267
Mean number of journals per category	46.03	63.56	66.40
Mean number of categories per journal	1.42	1.48	1.50
Overlapping percentage	42.26	47.58	49.89
Journals changing their classification	1988 (14.49%)	5784 (33.46%)	5874 (33.13%)
Journals adding categories	2426 (17.69%)	3820 (22.10%)	4192 (23.64%)
Journals losing categories	3951 (28.81%)	4540 (26.26%)	4603 (25.96%)

Table 4: Some relevant indicators about the new SJR classification based on Ward's clustering and hard combination of citation measures and previous classifications developed by Louvain and VOS community detection and soft combination of citation measures.

Several important points may be remarked from the indicators shown in table 4. First of all, we can notice that classification based on Louvain and VOS provided larger sets of classified journals. However, this was not a result of algorithm performance, which in all cases allow clustering the whole set of journals, but rather a consequence of the harder combination of the citation measures used in the present work (citation AND co-citation AND coupling) in comparison to the former one (citation OR co-citation OR coupling) and other methodological procedures. Furthermore, the 2-phase clustering conducted in this work favored the creation of a higher number of new categories with respect to Louvain and VOS methods, overcoming them in 26 and 31 categories respectively. Obviously, a smaller set of journals in tandem with a higher set of categories will give rise to a lower 'mean number of journals per category' in the new SJR classification system. On its side, indicators concerning the 'Mean number of categories per journal' and 'Overlapping percentage' which are less influenced by the size of

the journal and category sets revealed the best results for new SJR classification system, followed not far by Louvain and VOS system.

More significant findings can be extracted by analyzing the figures on journals changing, adding or losing categories in the three systems under comparison, especially when figures are transformed into percentages. It should be emphasized that while the new SJR classification presents the lowest values in journals changing and adding categories by far, conversely, it has the highest value in relation to journals losing categories. This fact uncovers a general refinement of classification through a decrease of overlapping in journal classification which is validated and corroborated by the data supplied by the earlier indicators on 'Mean number of categories per journal' and 'Overlapping percentage'.

Discussion and conclusions

In a previous study (Gómez-Núñez et al., 2014) we used the Person's citation-based measure combination (Weighted Direct Citation) to cluster journals of SJR in order to improve their classification regardless only one or several measures were present at the same time during their integration. In contrast to it, we have now used an alternative hard combination of normalized citation-based measures (Normalized Weighted Direct Citation) where only couples of journal of the network meeting values for direct citation, co-citation and coupling links at once were contemplated. This novel proposal has emerged as a solid method to refine and update SJR journal classification system by several reasons, namely:

1. Persson (2010) applied the combination of citation-based measures with the aim of detecting research themes within LIS field. Here, we have adopted his approach as a reliable proposal to (a) update the classification systems of SJR platform and to (b) refine the journal adscription to the SJR categories. A total of 139 new categories representing fitted and emerging topics like 'Social Work', 'Nanoscience and Nanotechnology' or 'Plastic Surgery' among others, have been introduced into the SJR classification system. Regarding the refinement of journal assignment our method provoked that more than 8000 journals suffered changes in their classification by either altering their adscription or by losing/adding categories.
2. In relation to the latter point, the reduction of (a) journal multi-assignment together with an (b) overall lower concentration of journals over categories suggests an improvement in the journal classification in spite of having a smaller set of journals in the new system. The representation of the journal distribution showed a marked enhancement in the new system obtained with a flatter and less skewed final distribution of journals.
3. In our view, the combination of the three citation-based measures helps to maximize the cluster effect, even more when the hard combination proposed here is applied. We based this argument on the obviousness of the higher the number of variables to compare between two items, the richer the likelihood to identify similarities or dissimilarities between them by widening the possibilities of the comparison.
4. Apart from the advances aforementioned, our method allowed to replace very general categories as 'Miscellaneous' characterized by covering a broad scope and an big concentrations of journals in favor of new categories discovered as a consequence of our method performance, specially by means of the labeling stage.

Although our proposal has reported reliable and solid findings in updating and improving the SJR classification system by offering a fine-tuning in the assignment of journals to categories, a lower overlapping and a refinement and reduction of the number of categories in use, we have also found two main shortcomings in its performance concerning the following points:

1. Over 5000 journals from the initial network of SJR journals were isolated from the study because they did not satisfy the condition of having direct citation, co-citation and coupling together during the hard combination followed by our method. However, this fact should not be considered as fail of the method rather a requirement to enrich the comparison and extend the difference among the journal clusters obtained. Moreover, some other method, such as reference analysis or sibling journals should be proposed in order to classify the journals left. Earlier we have already used the analysis of references to improve SJR classification with optimal results (Gómez-Núñez, Vargas-Quesada, Moya-Anegón & Glänzel, 2011). Therefore, these all journals excluded by this method could be integrated in the set of clusters or subject categories generated now.
2. Then, two huge clusters have emerged from Ward's method over the different partitions analyzed. Indeed, this is a very usual issue in Ward's hierarchical agglomerative clustering performance. In addition, He (1999) noticed that this phenomenon was also favored when sparse matrices are being used, such as in the case of the citation matrix coming from citation patterns among SJR journals, with over 98% of zero values. To solve this question we have opted for using a pragmatic approach based on achieving our main purpose of refining and updating SJR journal classification. Thus, we have implemented a second round of clustering on two huge clusters detected which allowed refining the assignment of journals included in them. At the same time, the second round of clustering produced a complementary classification with new specific categories which enriched and improved the previous set of categories derived from single clustering procedure.

In relation to incoming and future research works, it seems to be interesting to address other journal classification procedures from different approaches. Network analysis arises as a quite feasible choice. Also, it would be interesting to enhance the methodology described here by combining citation-links with text-based methods. To conclude with avenues of future research, we would like to make a final reflection in looking for a reliable and consistent technique to assess the goodness-of-classification of the cluster-based solution provided by our method. Finally, it would be meaningful to focus on the classification dilemmas concerning multidisciplinary scientific journals.

Acknowledgements

We would like to express our appreciation to Diego Guzmán Morales for supporting in preparing data set as well as easing to solve computational problems and providing orientation in technical questions. Also, we want to thank to Wolfgang Glänzel and Bart Thijs for their patient and inestimable help.

References

Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multilevel classification of journals. *Scientometrics*, 44(3), 323–345.

Bichteler J., Parsons, R.G. (1974). Document retrieval by means of an automatic classification algorithm for citations. *Information and Storage Retrieval*, 10(7-8), 267-278

Borko, H., & Bernick, M. (1963) Automatic Document Classification. *Journal of the ACM,* 10(2), 151-162.

Borko, H., & Bernick, M. (1964) Automatic Document Classification Part II. Aditional Experiments. *Journal of the ACM,* 11(2), 138-151.

Borko, H. (1964). Measuring the Reliability of Subject Classification by Men and Machines. *American Documentation*, 15(4), 268-273.

Boyack, K.W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351–374.

Carpenter, M.P., & Narin, F. (1973). Clustering of Scientific Journals. *Journal of the American Society for Information Science*, 24(6), 425-436.

Chen, C.M. (2008), Classification of Scientific Networks Using Aggregated Journal-Journal Citation Relations in the Journal Citation Reports. *Journal of the American Society for Information Science and Technology*, 59(14), 2296–2304.

Cormack, R.M. (1971) A Review of Classification. *Journal of the Royal Statistical Society. Series A (General),* 134(3), 321-367.

Croft, W.B. (1980). A model of cluster searching based on classification. *Information systems*, 5(3), 189-195.

Elsevier (2014). Subject Area Categories of Scopus. Retrieved from http://help.scopus.com/Content/h_subject_categories.htm [Accessed: 06-06-2014].

Everitt, B.S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). London, Wiley. ISBN 978-0-470-74991-3.

Garfield, E. (1975). Clusters and classification. *Essays of an Information Scientist*, 2, 354-355.

Garfield, E., Malin, M.V., & Small, H. (1975). A System for Automatic Classification of Scientific Literature. *Journal of the Indian Institute of Science,* 57 (2), 61-74.

Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, 63(1), 163-180.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics* 56(3), 357–367.

Glänzel, W., & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, 91(2), 399–416.

Gómez, I., Bordons, M., Fernández, M.T., & Méndez, A. (1996). Coping with the Problem of Subject Classification Diversity. *Scientometrics*, 35(2), 223-235.

Gómez-Núñez, A.J., Vargas-Quesada, B., Moya-Anegón, F., & Glänzel, W. (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, 89(3), 741–758.

Gómez-Núñez, A.J., Batagelj, V., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Moya-Anegón, F. (2014). Optimising SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, 8(2), 369–383.

He, Q. (1999). *A review of clustering algorithms as applied to IR* (Doctoral dissertation). Univ. Illinois at Urbana-Champaign, Tech. Rep. UIUCLIS-1999/6+IRG.

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid Clustering for Validation and Improvement of Subject-Classification Schemes. *Information Processing & Management*, 45(6), 683-702.

Klavans, R., & Boyack, K.W. (2007). Is there a convergent structure of science? A comparison of maps using the ISI and Scopus databases. In D. Torres-Salinas, H. Moed (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics* , Madrid, Spain, 437–448.

Klavans, R., & Boyack, K.W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3), 455-476.

Leydesdorff, L. (2002). Dynamic and evolutionary updates of classificatory schemes in scientific journal structures. *Journal of the American Society for Information Science and Technology*, 53(12), 987–994.

Leydessdorff, L., & Rafols, I. (2009), A Global Map of Science Based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology*, 60(2):348–362,.

Leydessdorff, L., Moya-Anegón, F., & Guerrero-Bote, V.P. (2010). Journal Maps on the Basis of Scopus Data: A Comparison with the Journal Citation Reports of the ISI. *Journal of the American Society for Information Science and Technology*, 61(2), 352–369.

Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., González-Molina, A., Muñoz-Fernández, F. J., et al. (2007). Coverage analysis of Scopus: a journal metric approach. *Scientometrics,* 73(1), 57-58.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Munoz-Fernández, F.J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129–145.

Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Gonzalez-Molina, A., Munoz-Fernández, F.J., et al. (2007). Visualizing the Marrow of Science. *Journal of the American Society for Information Science and Technology*, 58(14), 2167-2179.

Narin, F. (1976). *Evaluative Bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. New Jersey: Computer Horizons.

Persson, O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, 4(3), 415–422.

Price, D.J.D. (1963). *Little Science, Big Science*. New York: Columbia University Press.

Price, N., & Schiminovich, S. (1968). A clustering experiment: first step towards a computer-generated classification scheme. *Information and Storage Retrieval*, 4(3), 271-280.

Pudovkin, A.I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology,* 53(13), 1113-1119.

R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org/ [Accessed: 21-04-2014].

Rafols, I., & Leydesdorff, L. (2009). Content-Based and Algorithmic Classifications of Journals: Perspectives on the Dynamics of Scientific Communication and Indexer Effects. *Journal of the American Society for Information Science and Technology*, 60(9), 1823–1835.

Rousseau, R., & Zuccala, A. (2004). A classification of author co-citations: definitions and search strategies. *Journal of the American Society for Information Science and Technology,* 55(6), 513-629.

Salton, G., & Buckley, C. (1998). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.

Schiminovich, S. (1971). Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm. *Information and Storage Retrieval*, 6(6), 417-435.

Scimago. (2007). *Scimago Journal & Country Rank*. Retrieved from http://www.scimagojr.com/ [Accessed: 21-04-2014].

Small, H., & Griffith, B.C. (1974). The structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies*, 4, 17-40.

Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations I. A comparison of methods. *Scientometrics*, 7(3-6), 391-409.

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the science citation index using co-citations II. Mapping science. *Scientometrics*, 8(5-6), 321-340.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. *Journal of Information Science*, 11(4), 147–159.

van Rijsbergen, C.J., & Croft, W.B. (1975). Document clustering: an evaluation of some experiments with the Cranfield 1400 collection. *Information Processing and Management*, 11, 171-182.

van Rijsbergen, C.J. (1974). Further experiments with hierarchic clustering in document retrieval. *Information Storage and Retrieval* , 10(1), 1-14.

Vargas-Quesada, B., & Moya-Angeón, F. (2007). *Visualizing the Structure of Science*. Berlin: Springer.

Wild F. (2011). *lsa: Latent Semantic Analysis. R package version 0.63-3*. Retrieved from http://cran.r-project.org/web/packages/lsa/index.html [Accessed: 21-04-2014].

Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing & Mangement*, 24(5), 517-597.

Yagi, E., Badash, L., & Beaver, D.B. (1996). Derek J. de S. Price (1922-83): Historian of science and herald of scientometrics. *Interdisciplinary Science Reviews*, 21(1), 64-84.

Zhang, L., Janssens, F., Liang, L., & Wolfgang, G. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics*, 82(3), 687–706.

Annex I: *New SJR categories emerging from our classification proposal based on Ward's hierarchical agglomerative*

	CATEGORY
1	Aerospace Science, Technology & Engineering
2	Agribusiness
3	Agricultural Engineering
4	Agricultural Science
5	Alcohol & Drug Abuse
6	American History
7	Applied Physics
8	Aquaculture
9	Behavioural Sciences
10	Bioethics
11	Biomedical Research
12	Business Ethics
13	Cardiothoracic surgery
14	Ceramic materials
15	Chemical Engineering
16	Chemical Science & Technology
17	Child & Developmental Psychology
18	Chinese Medicine
19	Classical Studies
20	Combinatorics
21	Composites
22	Computer & Information Systems Engineering
23	Computer Systems
24	Contemporary European History
25	Criminology & Criminal Justice
26	Deafness
27	Dentistry
28	Development Studies
29	Developmental Disabilities Research
30	Earth Sciences
31	Economic History
32	Economics
33	Educational Research
34	Electromagnetics & Microwave
35	Electronics
36	Energy & Environmental Policy
37	Engineering Geology
38	Engineering Science & Technology
39	Environmental Chemical Engineering
40	Environmental economics
41	Environmental Pollution

42	Evidence-Based Medicine
43	Evolutionary Biology
44	Family Counseling & Psychology
45	Forensic Psychology
46	Forensic Science & Legal Medicine
47	Geophysics and Planetary Research
48	Geotechnical Engineering
49	Health Administration
50	Health Communication
51	Health Psychology
52	Healthcare
53	Higher Education
54	History of France
55	History of Science
56	Human Biology
57	Imaging
58	Infection Control
59	Infectious Diseases and Clinical Microbiology
60	International Law
61	Language
62	Learning & Educational Technology
63	Linguistics & Language
64	Literary Studies
65	Literature
66	Macroeconomics & Finance
67	Management Engineering
68	Materials Engineering
69	Mathematical Analysis
70	Mathematical Biology
71	Mathematics (general)
72	Mathematics Education
73	Medical Societies
74	Medical Ultrasound
75	Medicine, General
76	Medicine, General (Ibero-America)
77	Medicine, General (India)
78	Medicine, General (Iran)
79	Medicine, General (Middle East)
80	Medicine, General (Poland)
81	Music Education
82	Mycology
83	Nanomaterials
84	Nanoscience and Nanotechnology
85	Neuropsychology & Cognitive Neuroscience
86	Neuroscience & Brain Research
87	Neurosurgery

88	Nuclear & High Energy Physics
89	Nuclear & Pharmaceutical Chemistry
90	Nuclear Medicine
91	Numerical Methods
92	Nursing (general)
93	Nutrition
94	Oral & Maxilofacial Surgery
95	Paleoecology (quaternary)
96	Pathology
97	Pediatric Primary Care
98	Pediatrics, Women and Child Health
99	Petroleum Engineering
100	Philosophy of Psychology
101	Physics
102	Phytopathology
103	Plant & Agricultural Research
104	Plant Biotechnology
105	Plasma Physics
106	Plastic Surgery
107	Politics
108	Power Electronics
109	Proteomics
110	Psychology
111	Psychopedagogy
112	Psychotherapy
113	Public Health
114	Radiology & Imaging
115	Risk Management
116	Science, Technology and Society
117	Sismology
118	Sleep Medicine
119	Social & Economic History
120	Social & Political Philosophy
121	Social Economics
122	Social Policy
123	Social Studies
124	Social Work
125	Sociology
126	Sport Psychology
127	Sports & Leisure
128	Terramechanics
129	Textile Technology
130	Therapeutics
131	Traumatology
132	Tribology
133	Vascular Surgery

134	Veterinary
135	Veterinary Research and Animal Science
136	Waste Technology & Management
137	Water Resource Management
138	Wildlife Biology
139	Wounds

Informação e/ou Conhecimento: as duas faces de Jano. I Congresso ISKO Espanha e Portugal / XI Congreso ISKO España. Porto, 7 a 9 de novembro de 2013.
http://www.youblisher.com/p/749221-I-Congresso-ISKO-Espanha-e-Portugal-XI-Congreso-ISKO-Espana/

VISUALIZACIÓN Y ANÁLISIS DE LA ESTRUCTURA DE LA BASE DE DATOS SCOPUS

Antonio J. Gómez-Núñez ✉
CSIC, Unidad Asociada Grupo SCImago. Granada, España
anxusgo@gmail.com
Benjamín Vargas-Quesada
CSIC, Unidad Asociada Grupo SCImago. Granada, España
Departamento de Información y Comunicación, Universidad de Granada. Granada, España
benjamin@ugr.es
Teresa Muñoz-Écija
CSIC, Unidad Asociada Grupo SCImago. Granada, España
teresamunyozecija@gmail.com
Félix de Moya Anegón
CSIC, Unidad Asociada Grupo SCImago. Granada, España
CSIC, Instituto de Políticas y Bienes Públicos (IPP). Madrid, España
felix.demoya@cchs.csic.es

Resumen

Introducción: La visualización de grandes redes de citación extraídas de bases de datos multidisciplinares como Web of Knowledge y Scopus es un tema de investigación recurrente en la investigación generada dentro de las ciencias de la información. La visualización de los elementos de la red y su agrupamiento en *clústeres* temáticos permite mapear la estructura de la investigación y la interrelación entre sus disciplinas, equiparables a los clústeres temáticos detectados.
Objetivos: Se pretende representar la estructura de Scopus en base a la extensa red de citación establecida entre las numerosas revistas Scopus incluidas en la plataforma SCImago Journal & Country Rank (SJR), que en nuestro estudio ascienden a 18891. Mediante técnicas de clustering y visualización, se procederá a la re-clasificación de las revistas.
Metodología: En base a la citación de trabajos, se obtuvieron listas de adyacencia agregadas a nivel de revistas para la *citación*, *co-citación* y *coupling*. Estas listas muestran parejas de revistas del SJR relacionadas mediante un valor numérico que expresa la fuerza de su relación. Las tres listas fueron integradas en una nueva resultante de su suma, y sus valores fueron normalizados mediante la *geo-normalización*. Por último, se ejecutó el algoritmo de clustering de *VOSviewer*. Los clústeres de revistas obtenidos se etiquetaron con las categorías originales del SJR junto con las palabras significativas más repetidas en los títulos.
Resultados y Discusión: El mapa resultante refleja la estructura de Scopus en función de un conjunto de categorías que representan el contenido temático de las revistas científicas incluidas en la base de datos. La reducción del conjunto de categorías en relación con el número inicial del SJR, así como el elevado número de cambios en la clasificación de las revistas sugiere un refinamiento y una optimización de la clasificación original.
Conclusiones: El *cienciograma* presentado constituye una representación fiable y precisa de la estructura de la investigación basada en revistas científicas, puesto que se fundamenta en la opinión de los expertos, reflejada por medio de sus citas.

Keywords: Clasificación; Visualización de información; Clustering

VISUALIZATION AND ANALYSIS OF SCOPUS DATABASE STRUCTURE

Abstract

Introduction: Visualization of big citation networks extracted from multidisciplinary databases as Web of Knowledge and Scopus is a recurrent topic in Library and Information Science research. Visualization and clustering of network items enable to map science and research structure on the basis of thematic clusters detected as well as their relations.
Objectives: We pretend to map Scopus database structure based on the extensive citation network derived from the full set of Scopus journals included in SCImago Journal & Country Rank (SJR) platform, which rise to 18891. We will re-classify the journals analysed using visualization and clustering techniques.
Method: Working from citation of papers we constructed three journal adjacency lists covering citation-based measures, namely, direct citation, co-citation and bibliographic coupling. These lists are showing journal couples related through a numeric value which express the strength of the relation. Then, the three lists were combined in a new one resulting from summing up their values which were later normalized through *geo-similarity* measure. Finally, VOSViewer clustering algorithm was executed and journal clusters obtained were labelled using original SJR category tags together with the most repeated significant words from journal titles.
Results and Discussion: The resulting map reflects the Scopus structure through a set of categories that represents thematic content of scientific journals included in the database. The reduction of categories as well as the high number of shifts in journal classification originated from our method suggests a refinement and optimization of SJR journal original classification.
Conclusions: The *scientogram* displayed arise like a reliable and accurate picture of science and research structure based on scientific journals, since it is built upon expert opinions, revealed by means of their citation patterns.

Keywords: Classification; Information Visualization; Clustering

Introducción

La visualización de información surge como una disciplina de enorme interés en el ámbito de la *Bibliometría* y de la *Cienciometría*, al proporcionar diferentes representaciones visuales y *cienciogramas* o mapas de la ciencia que facilitan el análisis de un dominio mostrando la estructura de la ciencia y de la investigación a través de las distintas disciplinas temáticas que la componen(elementos representados) junto con sus relaciones e interacciones(Moya-Anegón et al., 2007). Principalmente, estos mapas se construyen a partir de la literatura científica compilada por las bases de datos, utilizando diferentes unidades de análisis (papers, revistas, categorías…) y distintas unidades de medida, tanto basadas en la citación y sus derivados (citación directa, co-citación, coupling…), como en el texto de las publicaciones. Pero además de mostrar la estructura disciplinar de la ciencia, permiten contemplar la evolución temporal de la investigación, detectar frentes de investigación, áreas de interdisciplinariedad, temas emergentes o en decadencia, etc.

Respecto a la bases de datos, en la actualidad destacan especialmente dos: Web of Knowledge (Wok) (Thomson Reuters, 2009) y Scopus (Elsevier, 2002), consideradas por la mayoría de la comunidad científica como las fuentes de información con una cobertura más exhaustiva y con mayor prestigio y reconocimiento a nivel internacional. Estas bases de datos, tienen carácter multidisciplinar y dan cabida a un elevado número de revistas científicas de prestigio de las que no sólo proporcionan información bibliográfica detallada, sino también índices de citas que permiten construir numerosos indicadores bibliométricos. Estos instrumentos, que pueden ser tanto cualitativos como cuantitativos, resultan de gran valor en tareas de evaluación de la ciencia y la investigación, y en especial, para los encargados de la toma de decisiones y el diseño de la política científica de los países.

Ahora bien, en el desarrollo y diseño de herramientas basadas en las publicaciones científicas albergadas en las bases de datos, conviene tener presente que la correcta clasificación de la literatura resulta de vital importancia para conseguir productos y resultados coherentes, fiables y sólidos. Por lo general, en la construcción de los *cienciogramas* subyacen procesos de asociación y distribución espacial de los ítems representados en función de su similaridad. La asociación puede calcularse, por ejemplo, en base a la co-ocurrencia de palabras significativas o al número de referencias bibliográficas compartidas. Utilizando técnicas estadísticas como el *clustering* o el *análisis factorial* es posible detectar grupos temáticos interrelacionados, que pueden interpretarse como un reflejo de las diferentes disciplinas en las que puede descomponerse el conocimiento científico. En la actualidad, varias herramientas de visualización y análisis de redes, como Pajek (Batagelj & Mrvar, 1999) o VOSViewer (Eck & Waltman, 2010), integran diferentes algoritmos para la detección de *clústeres* o *comunidades* dentro de una red, descomponiéndolas en grupos de ítems similares y fuertemente relacionados entre sí. Así, las herramientas de visualización aparecen también como una solución efectiva para la optimización y el refinamiento de la clasificación de la literatura en las bases de datos.

Clustering y visualización de información

El clustering emerge como una de las técnicas estadísticas más utilizadas en la clasificación e identificación de grupos temáticos. Son muchos los métodos de clustering que han sido utilizados con frecuencia por investigadores del ámbito de la visualización de la información con el fin de delinear la estructura del conocimiento científico y de la investigación. Para ello, resulta indispensable disponer de un esquema de clasificación consistente que represente de forma efectiva las diferentes disciplinas y/o subdisciplinas que integran la ciencia. Algunas propuestas significativas en el uso de clustering para la construcción de mapas de la ciencia (basados tanto en literatura de WoK como de Scopus) fueron recogidas y estudiadas en un

trabajo de Klavans y Boyack(2009). Su propósito era desarrollar un mapa de la ciencia de consenso a partir de los mapas previamente analizados.

Numerosos investigadores han aplicado también algoritmos de clustering a matrices y redes de citación, co-citación y/o coupling de revistas. Chang y Chen(2011) proponen aplicar el método de *mínimum span clustering (MSC)* a una matriz cuadrada de citación de aproximadamente 1.600 revistas del Social Science Citation Index (SSCI). Leydesdorff, Hammarfelt y Salah (2011) utilizaron el *algoritmo k-core* para representar las 25 categorías específicas del área de *Arts & Humanities Citation Index* y tratar de integrar su representación en un mapa global de la ciencia desarrollado previamente (Rafols, Porter, & Leydesdorff, 2010). Leydesdorff y Rafols (2012) publicaron un estudio en el que una matriz de citación compuesta de 9162 revistas extraídas del Science Citation Index Expanded de 2009 se utilizó para elaborar mapas interactivos. Entre los diferentes métodos llevados a cabo utilizaron diversos algoritmos de clustering para detectar grupos de revistas relacionadas e incluirlas en grupos temáticos bien definidos.

A lo largo de la cuantiosa literatura existente sobre clustering pueden encontrarse experimentos ejecutados a distintos niveles de agregación. A nivel de documento, Small (1999)desarrolló un mapa de la ciencia jerárquico mediante un método combinado de conteo fraccionalizado de los documentos citados, single-linkage clustering y ordenación bidimensional conforme a un proceso de triangulación geométrica. Moya-Anegón y otros(2004)propusieron la co-citación de categorías como unidad de análisis y representación, combinándola más adelante con técnicas de reducción del espacio como pathfinder networks (PFNET), y de identificación de grupos como análisis factorial(Vargas-Quesada & Moya-Anegón, 2007). Ahlgren y Colliander (2009)analizaron diferentes métodos de cálculo de similaridad entre documentos en base a texto, coupling y combinación de ambos, así como varios métodos para representar y clasificar un pequeño conjunto de 43 documentos de la revista *Information Retrieval*. Para ello utilizaron *complete-linkage clustering* y compararon los clústeres generados automáticamente con una clasificación previa realizada por expertos. De forma similar, Boyack y otros (2011) aplicaron una combinación de presentaciones gráficas y *average-link clustering* sobre varias matrices de similaridad basadas en palabras significativas extraídas del título, resumen y descriptores del Medical Subject Headings (MeSH) de un total de más de 2 millones de artículos científicos extraídos de la base de datos Medline. Más recientemente Börner y otros (2012) presentaron una metodología para el diseño y actualización de un mapa de la ciencia y un sistema de clasificación elaborado para la Universidad de California, San Diego (UCSD) utilizando técnicas de clustering sobre matrices de similaridad de revistas de Web of Science (WoS) y Scopus.

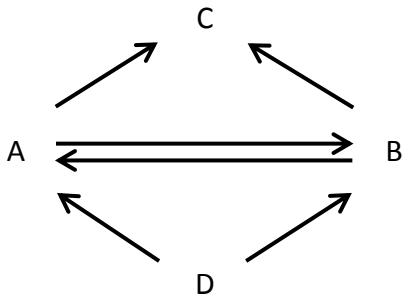
Por lo general, la ejecución de procedimientos de clustering sobre grandes redes y matrices de datos implica complejas operaciones de cálculo además de altos requerimientos de hardware y software. Además, la visualización de los datos debe ser lo más óptima y clara para su comprensión y manipulación. Programas como Pajek o VOSViewer surgen como excelentes herramientas para el análisis y la visualización de grandes masas de datos que, además, integran algoritmos de clustering para clasificar los datos analizados. En este trabajo, se utilizan ambas herramientas para llevar a cabo la representación de la estructura de la base de datos Scopus, utilizando Pajek para la preparación previa de los datos y VOSViewer para ejecutar la visualización final. Los detalles del proceso son pormenorizados en la siguiente sección de este trabajo.

Material y Método

En primer lugar, se diseñó una ventana temporal de dos años correspondiente a 2009 y 2010 y se recuperaron los datos de citación de un total de 18891 revistas de la plataforma SCImago

Journal & Country Rank (SJR) (Scimago Lab, 2007). Esta plataforma aglutina todas las revistas contenidas en la base de datos Scopus y permite elaborar indicadores para la producción de análisis de dominio y rankings tanto de revistas como de países. Para este conjunto de datos sólo se contaron las citas incluidas en el período temporal comprendido entre 2000 y 2010. Las citas se calcularon a nivel de artículos y posteriormente fueron agrupadas por revistas. A partir de estos datos se construyeron tres listas de adyacencia, compuestas por parejas de revistas y un valor numérico que expresa la fuerza de su relación, representando medidas basadas en la citación, como son la citación directa, la co-citación y el coupling.

Por último, estas tres listas fueron integradas en una lista final mediante la suma de las tres medidas de citación mencionadas y siguiendo para ello el modelo propuesto por Persson (2010),que él mismo denominó *Weighted Direct Citation (WDC)*. Conforme a su trabajo, se presenta a continuación el diagrama utilizado por Persson para calcular el WDC. No obstante, se ha introducido una sensible modificación con respecto al original, al representar en nuestro caso los dos sentidos en que puede considerarse la citación directa.



Siguiendo el diagrama, la fórmula para la integración de las tres medidas de citación sería la siguiente:

c_ij = ABC + DAB + max(AB,BA)

Donde ABC hace referencia al coupling, DAB a la co-citación, y AB o BA a la citación directa.

Seguidamente, se normalizaron los resultados de la red utilizando la *geo-normalización*, medida cercana al *coseno* de Salton y que funciona dividendo los elementos de la matriz por la media geométrica de ambos elementos de la diagonal (Batagelj & Mrvar, 1999). Se detalla a continuación, la fórmula utilizada para su cálculo:

s_ij = c_ij / sqrt(c_i*c_j), c_i = sum{ j: j!=i: c_ij }

La siguiente fase de nuestro método se corresponde con la ejecución del algoritmo de clustering sobre la red normalizada de revistas que integra las tres medidas basadas en la citación. VOSViewer permite llevar a cabo no sólo la visualización de información, sino también ejecutar un algoritmo de clustering que permite establecer una clasificación de los datos que posteriormente serán mapeados. En palabras de sus creadores (Waltman, Eck, & Noyons, 2010) este algoritmo incluye un parámetro de resolución capaz de detectar grupos o clústeres de pequeño tamaño si se proporciona un valor adecuado para configurarlo. También indican que un mayor parámetro de resolución implica un incremento paralelo del número de clústeres generados. Teniendo en cuenta estas consideraciones, decidimos ejecutar varias pruebas introduciendo distintos valores en el parámetro de resolución del algoritmo. De esa forma,

podríamos obtener diferentes soluciones ofreciendo distintas descomposiciones (subredes) de la red de revistas y, por lo tanto, produciendo diferentes conjuntos de clústeres o comunidades temáticas. Nuestro objetivo final fue obtener un sistema de clasificación consistente para representar de forma eficiente las diferentes disciplinas de la ciencia y la investigación a partir de la literatura científica compilada por Scopus.

La Figura 1 muestra la correlación existente entre el valor del parámetro de resolución y el número de comunidades o clústeres proporcionados por el algoritmo de VOSViewer a lo largo de las diferentes pruebas ejecutadas. Atendiendo a nuestros propósitos de visualización, una solución aportando alrededor de 250-300 clústeres resultaría efectiva para representar la estructura de la ciencia eficazmente. Además, se estimó que con independencia de la solución final escogida, el tamaño mínimo de clúster no podría ser inferior a 10 revistas, garantizando de esta forma una serie de grupos temáticos con un mínimo de consistencia y delimitación. Una vez analizados los resultados de las diferentes pruebas, se consideró que el parámetro de resolución con valor 15, aportando 270 clústeres útiles con más de 10 revistas, resultaba ser una solución óptima para la elaboración del *cienciograma* que representaría la estructura de Scopus.

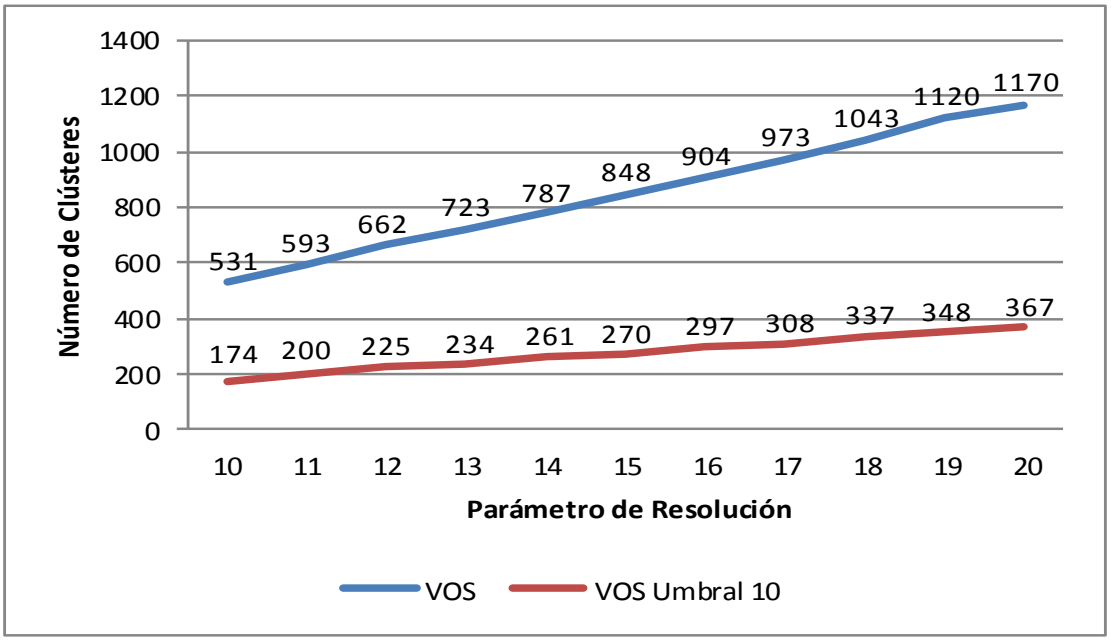


Figura 1: Número de clústeres por parámetro de resolución con y sin umbral 10

En la clasificación original de la plataforma SJR, sus 18891 revistas se distribuyen entre un total de 304 categorías temáticas, con un número medio de revistas por clúster de 62,14. Nuestra propuesta, en base al algoritmo de VOSViewer con parámetro de resolución 15 y un tamaño mínimo de clúster de 10 revistas, arrojó un número medio de revistas que asciende a 65,66 y que, por lo tanto, se antoja bastante similar a la distribución de revistas por categorías de la clasificación original del SJR. Las distribuciones referentes al número de revistas clasificadas y al número medio de revistas por clúster a lo largo de los distintos parámetros de resolución configurados pueden observarse en las Figuras 2 y 3 respectivamente.

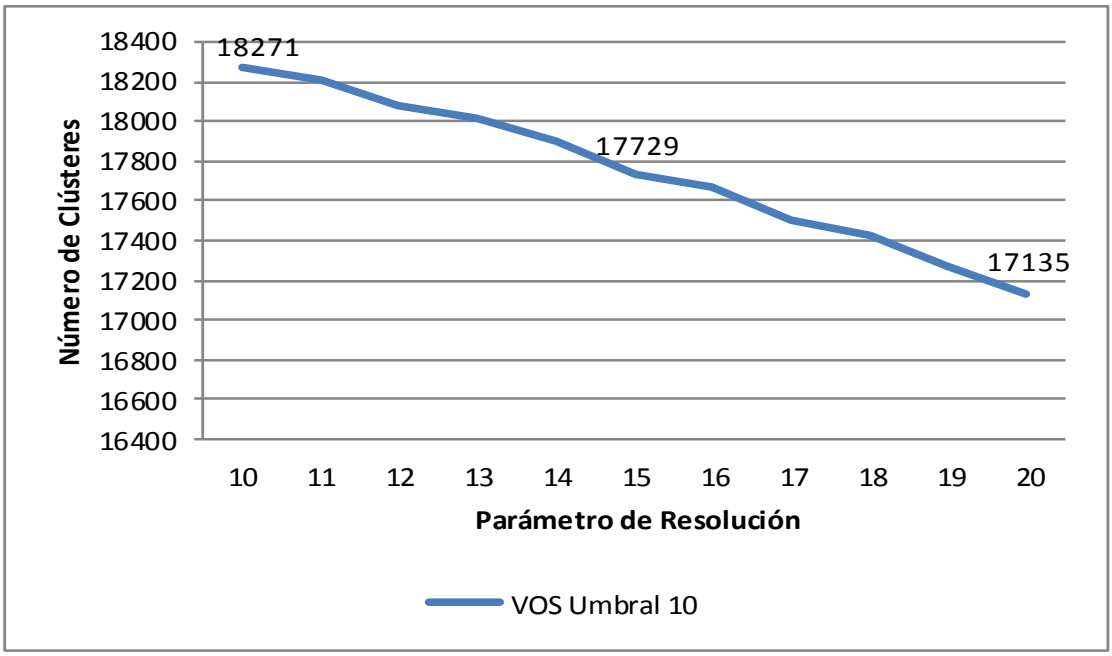


Figura 2: Número total de revistas clasificadas

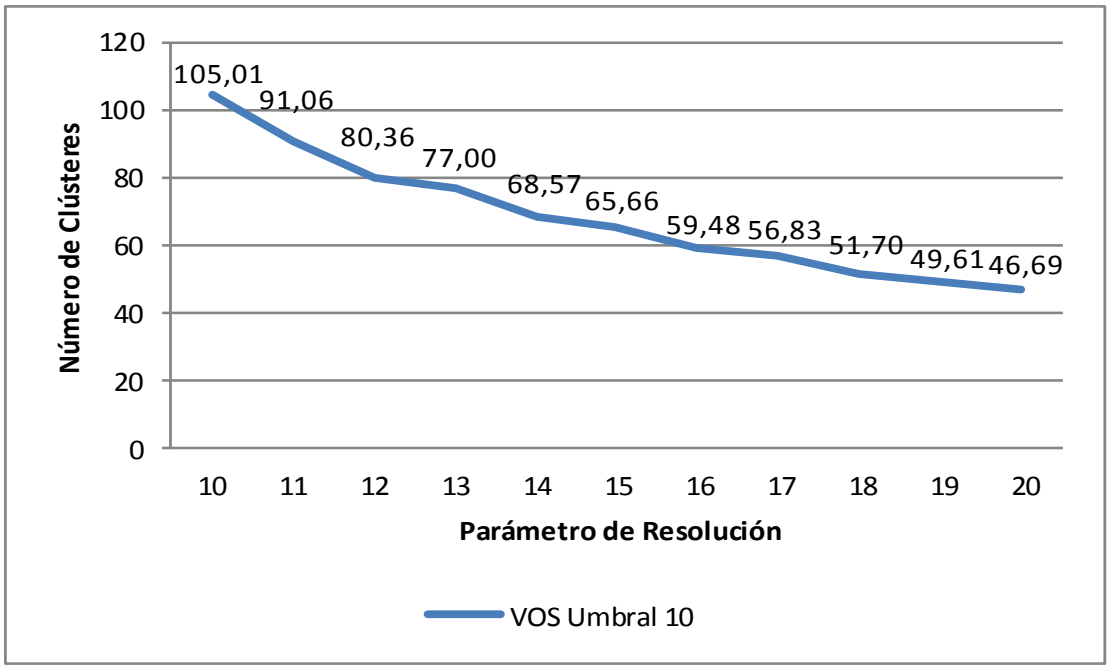


Figura 3: Número medio de revistas por clúster

Al mismo tiempo que el número de clústeres generados por el algoritmo se incrementa al aumentar el valor del parámetro de resolución, si aplicamos el umbral de 10 revistas como tamaño mínimo de clúster, comprobamos que se incrementa también el número de clústeres que no superan dicho umbral y que, por consiguiente, no son útiles para nuestros propósitos. En otras palabras, existe correlación positiva entre el incremento del valor del parámetro y el aumento del número de clústeres tanto que superan el umbral como que no. En relación con el número de revistas asignadas, tal como afirman los creadores de VOSViewer, todas las revistas son asignadas a alguna comunidad o grupo durante el procedimiento de clustering. No obstante, si comparamos el incremento del parámetro de resolución con el total de revistas

asignadas a clústeres que superen el umbral de 10 como tamaño mínimo, entonces encontraremos que existe una correlación negativa entre ambas variables.

El último paso de nuestro método está relacionado con el etiquetado de los diferentes clústeres o comunidades detectadas por el algoritmo de VOSViewer. Para tal fin nos propusimos reutilizar los nombres o etiquetas de las categorías originales del SJR, asignando a cada clúster el nombre de las etiquetas más repetidas y derivadas de las revistas incluidas. La frecuencia de aparición de dichas etiquetas fueron transformadas en porcentajes y en pesos *tf-idf* (Salton & Buckley, 1988). A continuación, las categorías fueron ordenadas conforme a los pesos *tf-idf* y se seleccionaron sólo aquellas que representaban al menos un 33% del conjunto total de categorías citadas por las revistas incluidas en cada clúster para su delimitación temática. En el proceso de etiquetado se desecharon todas las categorías originales marcadas como *Misceláneas* o *Multidisciplinar.* Este hecho provocó que algunos clústeres tuvieran que ser etiquetados a posteriori, utilizando para ello una combinación de etiquetas de las categorías SJR originales junto con términos significativos extraídos del título de las revistas agrupadas en los clústeres.

Antes de continuar, conviene destacar dos asuntos importantes derivados del proceso de etiquetado desarrollado en nuestro método. En primer lugar, el uso de etiquetas de categorías prediseñadas dio lugar a la aparición de varios clústeres nombrados con exactamente las mismas etiquetas. Estos clústeres fueron fusionados en otro clúster nuevo, al considerarse que sus revistas abarcaban temas análogos y colindantes. Como consecuencia, el esquema temático inicial generado en base al algoritmo de VOSViewer derivó en un nuevo esquema temático basado en las etiquetas de las categorías y compuesto por un total de 219 categorías temáticas. En segundo lugar, habría que resaltar que la multi-asignación de revistas no proviene del funcionamiento del algoritmo propiamente dicho, sino que es una consecuencia de nuestro proceso de etiquetado, que permite asignar una revista a más de una categoría provocando un solapamiento en determinadas categorías temáticas. No obstante, esta multi-asignación no resulta demasiado elevada. Así, si observamos la Figura 4, descubrimos que alrededor de un 60% de las revistas fueron asignadas a una sola categoría temática, mientras que cerca de un 30% se asignaron a dos categorías, casi un 7% a tres categorías y un valor casi residual de revistas fueron asignadas a cuatro categorías.

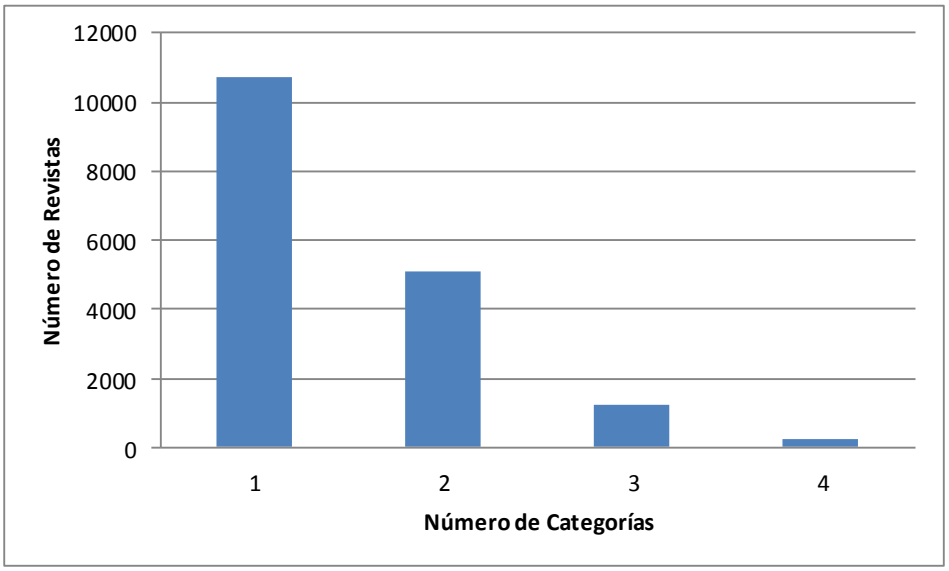


Figura 4: Asignación de revistas a categorías

Resultados y Discusión

La clasificación final del conjunto de revistas analizadas y sometidas al proceso de clustering comentado, puede consultarse en la siguiente dirección web:
http://www.ugr.es/local/benjamin/vos15_classification.pdf

Desde el punto de la visualización de la información a continuación mostramos y analizamos de forma breve, por las limitaciones lógicas de espacio, el cienciograma completo que representa la clasificación general obtenida, así como el de otras disciplinas que se pueden distinguir a simple vista.

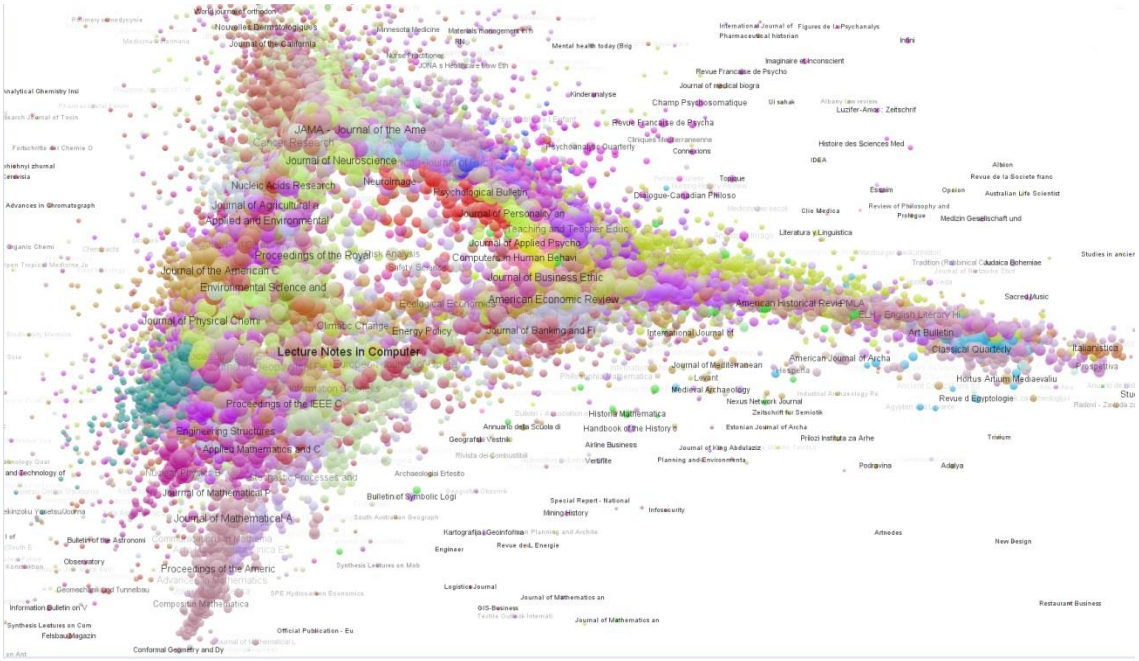


Figura 5: Cienciograma de revistas Scopus

El cienciograma de la Figura 5 muestra cómo se agrupan las revistas de la base de datos Scopus en base a sus medidas de citación, co-citación y coupling. El tamaño de las esferas (revistas) y de sus etiquetas, es proporcional a su grado de interacción con el resto. A mayor interacción, mayor tamaño. El color de cada esfera muestra su adscripción, es decir, el clúster (categoría) al que ha sido adscrita cada revista mediante el algoritmo de Vosviewer (VOS). El cienciograma está construido en base a los principios de VOS, que puede ser considerado como una especie de MDS (Multidimensional Scaling) ponderado mediante proximidades y pesos, que evita los dos problemas/artefactos característicos del MDS: la tendencia a colocar los ítems más importantes en el centro, y la propensión a crear representaciones circulares(Eck, Waltman, Dekker, & Berg, 2010). Esto se consigue haciendo que la proximidad entre dos ítems sea igual a la inversa de su similaridad, y que su peso sea igual a su similaridad.

Grosso modo, si nos fijamos en el cuerno que aparece en la parte inferior del cienciograma, de color marrón claro, observaremos como es en ese sitio donde se agrupan las revistas de Matemáticas. Siguiendo el sentido de las agujas del reloj, y por tanto a su izquierda, podemos observar de color rosa y verde a distintos tipos de Ingenierías, la Física de color amarillo, justo por encima de ella, y sobre esta última a la Agricultura. En el cuerno superior, y justo encima de la Agricultura, encontramos las Neurociencias, la Psicología, la Medicina y la Biología. Desde esta posición y siguiendo de nuevo el sentido de las agujas del reloj, podemos detectar la Sociología, la Lingüística, la Historia y la Literatura (extremo derecho del cienciograma).

Siguiendo el cienciograma hasta nuestro punto de origen detectamos la Economía, la Documentación, y la Ciencias de la Computación, que conectan con las Matemáticas.

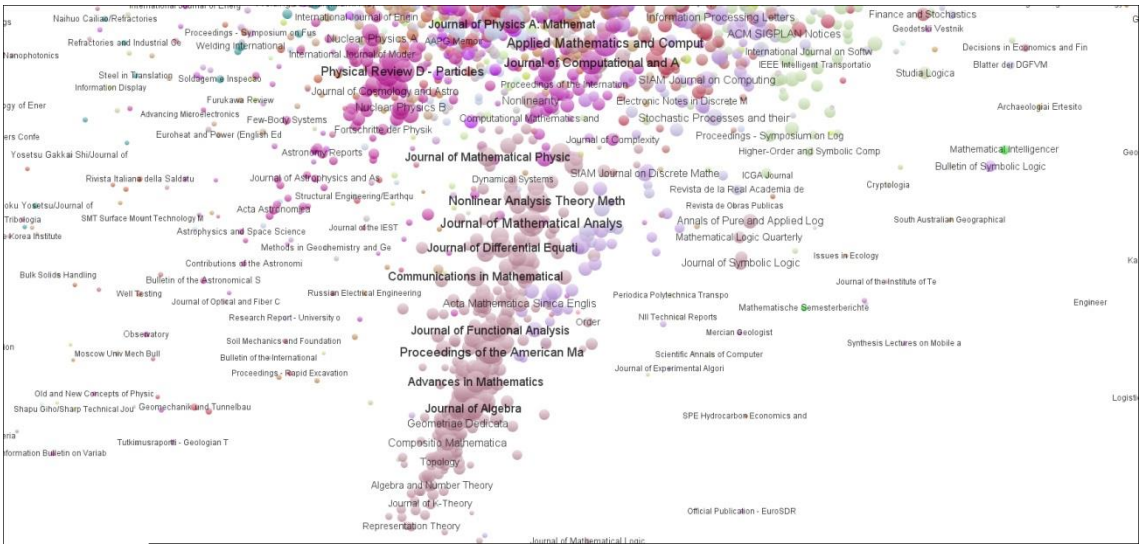


Figura 6: Cienciograma de revistas de Matemáticas de Scopus

La Figura 6 muestra claramente cómo se agrupan y se estructuran las revistas de Matemáticas en base a sus relaciones de citación. Si se observa de forma detenida, se puede ver como en la parte inferior aparecen las revistas de Matemáticas básicas y cómo, a medida que se asciende en el cienciograma, empiezan a aparecer las revistas de Matemáticas aplicadas, llegando a mezclarse con las de Ciencias de la Tierra, Física y Astronomía en la zona superior izquierda, y con las de Ciencias de la Computación en el área superior derecha.

Otras disciplinas, como ocurre con Library & Information Sciences (LIS), representada por revistas como *JASIST* o *Scientometrics* que se sitúan en el centro de la Figura 7 en color verde, no se muestran tan cohesionadas ni forman un clúster tan bien definido como sucedía con las Matemáticas en la figura anterior.

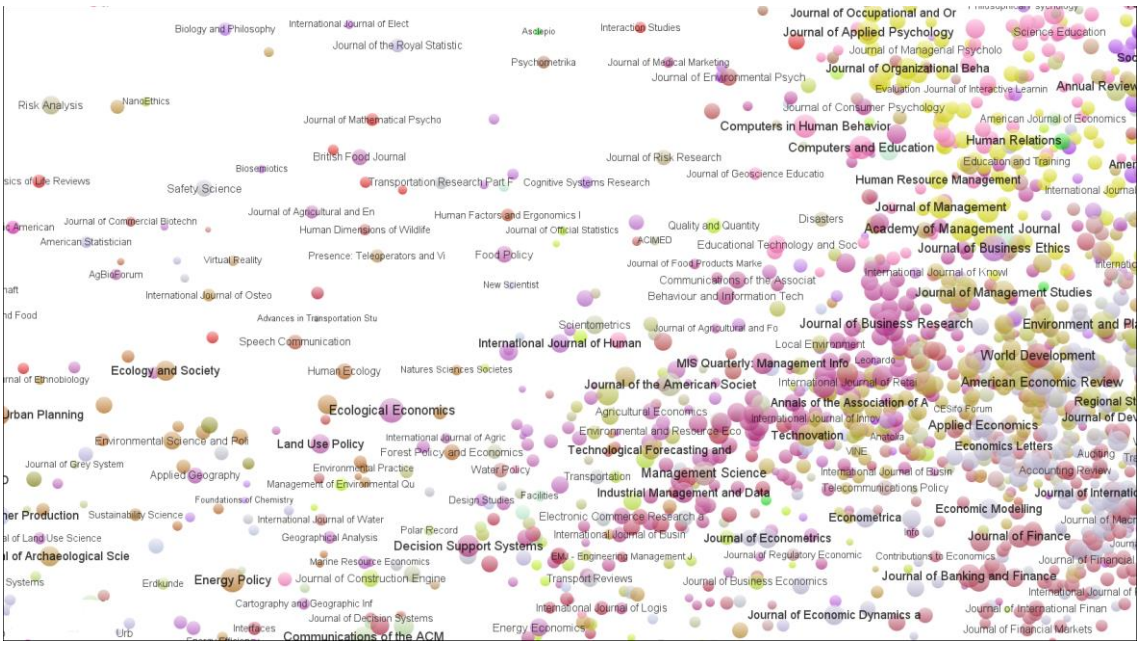


Figura 7: Cienciograma de revistas de Library & Information Science de Scopus

Esta dispersión de las revistas de LIS, al igual que ocurre con muchas otras disciplinas, se debe fundamentalmente a que muchas de estas revistas son interdisciplinares y acaban publicando contenidos que pertenecen a diferentes disciplinas científicas del mismo área. Desde el punto de vista de la visualización de la información, esto hace que las revistas de esta disciplina aparezcan cerca y, por tanto, mezcladas con otras de Ciencias de la Computación, Gestión, Planificación, etc. Este efecto disgregador de las revistas de una disciplina se ve además aumentado por la necesidad y obligatoriedad de tener que representar los cienciogramas en dos dimensiones (2D), puesto que su fin último es ser visualizados mediante una pantalla de ordenador de forma estática, o a través de un soporte como el papel. Esto provoca que la dimensión profundidad (Z) desaparezca, y que disciplinas que claramente están separadas, es decir, lejos de otras en un cienciograma en tres dimensiones, acaben estando juntas e incluso mezcladas, en otro de 2D.

Conclusiones

La visualización de la información es una herramienta muy potente para el análisis y corroboración de resultados en favor de la clasificación. No obstante, por si sola, no puede ni debe ser utilizada como única herramienta, pues la multidisciplinariedad e interdisciplinariedad de las unidades que se representan, en combinación con las limitaciones propias de un espacio de 2 dimensiones (papel o pantalla de ordenador) provocaran un falseamiento de los resultados. Por ello, al igual que hacemos en este trabajo, recomendamos generar esquemas de clasificación tradicionales apoyados y validados mediante técnicas de visualización de información.

La propuesta aquí realizada facilita la creación de un nuevo esquema de clasificación equilibrado en cuanto a la distribución de revistas por categorías, número de categorías útiles, y concentración moderada de revistas en las categorías o grupos temáticos con mayor poder de atracción. Así, en un trabajo desarrollado anteriormente(Gómez-Núñez, Vargas-Quesada, Moya-Anegón, & Glänzel, 2011) basado en un proceso iterativo de análisis de referencias bibliográficas citadas por las revistas del SJR, tan sólo cinco categorías resultaron suficientes para aglutinar un 25% de las 14166 revistas clasificadas. Con la propuesta de clasificación que presentamos aquí, se necesitan 18 categorías para alcanzar algo más de ese 25% de revistas clasificadas. Teniendo en cuenta que el número de revistas ahora es bastante más elevado (17729), este asunto adquiere aún mayor importancia, puesto que ese amplio margen podría favorecer todavía más las concentraciones de revistas en categorías con mayor poder de atracción.

Referencias Bibliográficas

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49–63. doi:10.1016/j.joi.2008.11.003

Batagelj, V., & Mrvar, A. (1999). Pajek – Program for Large Network Analysis. Retrieved from http://pajek.imfm.si/doku.php

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., Larivière, V., et al. (2012). Design and update of a classification system: the UCSD map of science. *PloS one*, *7*(7), e39464. doi:10.1371/journal.pone.0039464

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., et al. (2011). Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PloS one*, *6*(3), e18029. doi:10.1371/journal.pone.0018029

Chang, Y. F., & Chen, C. (2011). Classification and Visualization of the Social Science Network by the Minimum Span Clustering Method, *62*(12), 2404–2413. doi:10.1002/asi

Eck, N. J. Van, & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. doi:10.1007/s11192-009-0146-3

Eck, N. J. Van, Waltman, L., Dekker, R., & Berg, J. Van Den. (2010). A Comparison of Two Techniques for Bibliometric Mapping : Multidimensional Scaling and VOS, *61*(12), 2405–2416. doi:10.1002/asi

Elsevier. (2002). Scopus. Retrieved April 12, 2013, from http://www.scopus.com/home.url

Gómez-Núñez, A. J., Vargas-Quesada, B., Moya-Anegón, F., & Glänzel, W. (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, *89*(3), 741–758. doi:10.1007/s11192-011-0485-8

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476. doi:10.1002/asi.20991

Leydesdorff, L., Hammarfelt, B., & Salah, A. (2011). The Structure of the Arts & Humanities Citation Index : A Mapping on the Basis of Aggregated Citations Among 1 , 157 Journals, *62*(12), 2414–2426. doi:10.1002/asi

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from Web-of-Science data. *Journal of Informetrics*, *6*(2), 318–332. doi:10.1016/j.joi.2011.11.003

Moya-anegón, F., Vargas-Quesada, B., Chinchilla-rodríguez, Z., Corera-álvarez, E., Munoz-fernández, F. J., & Herrero-solana, V. (2007). Visualizing the Marrow of Science, *58*(14), 2167–2179. doi:10.1002/asi

Moya-Anegón, F., Vargas-Quesada, B., Herrero-solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Munoz-fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, *61*(1), 129–145.

Persson, O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, *4*(3), 415–422. doi:10.1016/j.joi.2010.03.006

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science Overlay Maps : A New Tool for Research Policy and Library Management. *Journal of the American Society for Information Science and Technology*, *61*(9), 1871–1887. doi:10.1002/asi

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Scimago Lab. (2007). Scimago Journal & Country Rank (SJR). Retrieved April 15, 2011, from http://www.scimagojr.com/

Small, H. (1999). Visualizing Science by Citation Mapping, *50*(1973), 799–813.

Thomson Reuters. (2009). ISI Web of Knowledge. Retrieved April 12, 2013, from http://wokinfo.com/

Vargas-Quesada, B., & Moya-Anegón, F. (2007). *Visualizing the structure of science*. New York: Springer.

Waltman, L., Eck, N. J. Van, & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. doi:10.1016/j.joi.2010.07.002

PARTE III: ABSTRACT

1. Introduction

In today´s society, scientific/technological development is clearly understood as a sign of progress and well-being, in economic and in social terms. The activities and strategies destined to uphold research and innovation are considered key elements of the scientific policy of advanced economies. This is not only true for individual countries, but also in the context of supranational organizations or communities, including, for instance, the European Union. Through organizations and specific financing agencies, governments designate vast amounts of public funds to support research projects and innovation programs to make businesses and the economy in general more competitive, to enhance the quality of living of citizens and, of course, to achieve excellence in scientific/technological research institutions.

Notwithstanding, the exponential growth of science and research has been accompanied by a scarcity of resources. As the situation becomes more and more serious, there is greater pressure for the financing agencies that help develop research programs of interest, and distribute public funds among the numerous deserving candidates. Within the framework of such programs, the responsibility for professionals in charge of decision-making can be tremendous.

So as to remain above suspicion in these decisive matters, funding agencies rely on various mechanisms that facilitate evaluation in diverse stages: *ex ante* (selection), *interim* (follow-up) and *ex post* (final assessment of results and potential impact). Bibliometrics have become a supporting discipline of great utility and potential in this sense, making it possible to quantify the success of a project or program using

indicators such as the number of publications generated (articles, patents, conference papers, etc.), the impact of the publications (citation), the visibility and quality of the publications (quartiles of the source journals), and institutional collaboration.

In view of these bibliometric indicators, rankings of entities (regions, countries, institutions, researchers or research groups) can be derived. These increasingly popular listings offer a broad perspective of the different realities analyzed (output, impact, visibility…), facilitating decision-making for those persons in charge of managing, funding or designing the scientific policy of a region, country or institution. After all, rankings and the bibliometric indicators used to build them are by-products of the most tangible result of research: scientific publications. This type of literature may take on different documental types, depending on the subject matter. The scientific article is common, yet other formats are conference papers, reviews, books or patents. Once published or registered, the documents are collected and processed by the services of information and summarizing of vast databases such as WoS, Scopus, PubMed and PatStat.

The technical treatment by information services includes indexing, summarizing, and classification. From the bibliometric standpoint, these processes are crucial for the elaboration of indicators that consistently and reliably represent the reality subjected to analysis. Particularly important is an adequate classification of literature when the objective is domain analysis, ranking the output of institutions, or qualifying the collaboration of authors from a specific discipline.

2. Objectives

Despite its relevance since ancient times, approached by great figures of scientific knowledge, classification of the sciences —and therefore of scientific output— remains essential for disciplines such as Library and Information Science, and sub-disciplines such as Bibliometrics. A proper classification of scientific literature is key for the elaboration of indicators, rankings, and domain analysis. Potential improvements in the classification of the SCImago Journal & Country Rank (SJR), an open-access platform for the creation of bibliometric indicators useful for evaluating and analyzing scientific domains, are presented in this doctoral thesis. We test different proposals for classification based on semi-automatic processes described in four constituent scientific publications. Improvements entail updating the classification scheme used by the platform, and refining the assignment of journals to the different thematic categories of the system.

3. Material and Methods

The different experiments designed and implemented in order to improve the SJR classification are based on information from the Scopus database (Elsevier) included in the SJR platform. For each experiment, temporal windows and citation windows were established *ad hoc*, to compile concrete data about publications and the citation of "active" SJR journals. The windows designed could vary depending on the date of execution of the experiment, which may imply a direct effect on the total number of citation, for example, or on the final set of active journals involved in the study. This should be stressed, as it bears upon the replication of some of the experiments developed as part of the thesis.

All the methodological considerations used for the different classification experiments have been detailed in the four scientific articles that make up this doctoral thesis. They can be summed up as:

- *Article 1*: It pursues an improvement in the classification scheme and in the assignment of journals to the SJR platform to thematic categories by means of a process based on the iterative analysis of bibliographic references to the journals. Some relevant aspects of the performance of the study were determined on the basis of heuristic and empiric criteria, including:
    - The selection of the total number of iterations to be executed in the analytical process, which ranged from 1 to 12.
    - The choice of a cut-off point appropriate for determining a number of iterations that would lead to the generation of a coherent classification system and a consistent classification of journals.
    - The determination of a threshold (expressed as a percentage) that effectively delimits the vectors of journals-categories, so that each one of the vectors contains only those categories that best define the journals, excluding the least representative categories.
- *Article 2*: Given the results of the analysis of reference, new solutions were sought to improve certain aspects of the SJR classification, especially the total journals classified and their distribution in categories. It was then attempted to confirm the effectiveness of the Louvain and VOS clustering algorithms in detecting communities in a citation network (and derivatives) of SJR journals. Within this

network, three measurements were integrated: direct citation, co-citation and coupling, following the notions presented by Persson (2010), who denominated the combination of the three as Weighted Direct Citation. In addition to corroborating the adequacy of the algorithms, a comparative analysis was carried out of the two, which revealed very similar functionalities. These results were then compared with the results obtained using the original WoS and SJR classification systems.

- *Article 3*: Continuing with the clustering algorithms and the combination of measurements derived from citation, a new clustering procedure was designed using the hierarchical agglomerating Ward method, plus an alternative combination of direct citation, co-citation and coupling, with the obligatory presence of the three measures for each pair of journals related in the matrix. This was meant to guarantee the establishment of more solid relations among all the journals included in the experiment. Again, the results obtained were compared with those of the other classification systems (SJR, Louvain and VOS) applied in *Article 2.*
- *Article 4*: The three classification proposals developed in *articles 1, 2* and *3* underwent a test to validate the classification results, basically by means of a comparative analysis with similar classification systems, and most notably the original SJR. While true that this evaluation model is effective in providing a general view of the solutions generated, a complementary system would ensure more precise evaluation of results. To this end, techniques for the visualization of information were applied to the results from the VOS classification, and a visualization algorithm was used to display the network of journals previously

generated by the clustering algorithm. A detailed analysis of the graphs at different levels allowed us to detect aggregations that were comparable to the clusters obtained, thereby confirming the validity and coherence of the results of the VOS classification algorithm.

4. Results

The main results of the experiments in classification, as expounded in the articles of this thesis, are summed up in the table 1 below. The indicators that facilitate interpretation are:

10. *Total number of journals included in the system.* This indicator refers to the complete set of SJR journals that served as the basis of the study.
11. *Number of journals classified.* That is, the number of classified journals obtained after executing all the methodological procedures.
12. *Number of categories.* It indicates the number of useful categories or categories actually in use that are included in the final classification scheme.
13. *Mean journals per category.* This figure expresses the ratio between the total number of journals in the system and the total number of existing categories.
14. *Mean categories per journal.* The result of the multiple ascriptions of journals, this indicates the average number of categories corresponding to each journal of the system.
15. *Percentage of overlap.* Tied to the previous indicator, but expressed as a percentage, it reflects the overlap produced when there is multiple assignment of journals, to more than one category. It is calculated using the formula: **B - A / A * 100**

Where ***B*** is the total number of journals ascribed to the categories of the system (including repetitions deriving from multiple assignment) and ***A*** represents the real number of journals included.

16. *Total journals that change their classification.* It indicates the number of journals whose original SJR classification has undergone a modification in some ascribed category, as a result of the new classification process.
17. *Number of journals that add categories.* With regard to the original SJR classification, it quantifies the journals assigned one or more new categories as a consequence of the re-classification.
18. *Number of journals that lose categories.* Likewise, it expresses the number of journals that, as a result of re-classification, lose one or more categories with respect to the original SJR classification.

	SJR	Proposal 1	Proposal 2		Proposal 3	Proposal 4
	Original classification	**Analysis of References**	**Detection of Communities: Louvain**	**Detection of Communities: VOS**	**Clustering: Ward**	**Detection of communities (VOS) and Visualization**
Total no. of journals included in the system	18,891	17,158	18,891	18,891	18,891	18,891
Number of journals classified	18,891	14,166	17,287	17,729	13,716	17,729
Number of categories	308	198	272	267	298	267
Mean journals per category	61.33	71.55	63.56	66.40	46.03	66,40
Mean categories per journal	1.61	2.06	1.48	1.50	1.42	1.50
Percentage of overlap	60.73	106.18	47.58	49.89	42.26	49.89
Total journals that change their classification	-	2,872 (20.27%)	5,784 (33.46%)	5,874 (33.13%)	1,988 (14.49%)	5,874 (33.13%)
Number of journals that add categories	-	7,249 (51.17%)	3,820 (22.10%)	4,192 (23.64%)	2,426 (17.69%)	4,192 (23.64%)
Number of journals that lose categories	-	2,488 (17.56%)	4,540 (26.26%)	4,603 (25.96%)	3,951 (28.81%)	4,603 (25.96%)

Table 1: Main results of the experiments in classification

5. Discussion and Conclusions

Analysis of bibliographic references as a journal classification technique

The iterative analytical procedure, as applied to the cited references dealt with in article 1, caused noteworthy modifications in the SJR classification scheme: the number of areas was reduced (from 27 to 24), as were the subject categories (from 308 to 198) with respect to the original classification scheme. The main reason for this is the great power of attraction generated by some categories within the citation network; this may be due to citing patterns or habits characteristic of certain disciplines, a bias in thematic coverage of the database, the degree of consolidation of a given discipline, and, of course, the execution of the multiple iterations according to the method. There is a negative correlation between the number of iterations and the number of categories. In other words, the more the iterations, the fewer the categories.

The reduction in categories favors, in turn, the appearance of journals in higher concentrations, with a small group of categories. Therefore, the method would appear to be adequate for updating the SJR classification scheme, pruning the thematic categories with less weight and influence. However, the successive reduction of categories leads to a concentration of journals in just a few categories —which does not constitute refinement, but rather a readjustment in the journal classification per se. The yield and results of the experiment would be appropriate, for example, for readjusting journals to greater levels of aggregation, as in the case of the subject areas.

Clustering algorithms in networks based on journal citation

The clustering algorithms used in the experiments of *article 2* (detection of Louvain and VOS communities) and *article 3* (Ward´s hierarchical clustering) allowed us to classify the complete set of journals involved in each one of the experiments designed. However the application of considerations and methodological criteria particular to each experiment —e.g. setting a minimal cluster size, selecting the unit of measure to be used, or in this case the a combination of measures— caused certain subsets of journals to be left out to the final set of classified journals.

In *article 2*, many of the clusters generated by the Louvain and VOS algorithms were discarded due to the fact that they did not reach the minimal size of 10 journals required to represent the research disciplines in a precise and consistent way. This led to the exclusion of over 1000 journals in each case. In *article 3*, the alternative combination of measures based on citation, which forced the appearance of direct citation, co-citation and coupling at the same time for each pair of journals included in the matrix (*hard combination*), together with the execution of a second round of clustering into two super-clusters of great dimensions that did not fit the initial classification scheme, meant that over 5000 journals were excluded from the final classification process.

Meanwhile, the tagging system proposed in the experiments carried out in *articles 2* and *3*, based on the reutilization of tags of the original SJR categories according to citation thresholds weighted with *tf-idf* and the extraction of significant or relevant terms from the titles of the journals constituting the different clusters proved successful for updating the SJR classification scheme. This procedure made it possible

to configure balanced classification schemes that accommodated stable, solid categories of the original SJR scheme, along with other new categories obtained through application of the textual component in the tagging process. Even though all the clustering algorithms used in experiments 2 and 3 were *hard clustering* algorithms that proceed by assigning each journal to a single category, our particular tagging process facilitated the ascription of journals to different subject categories (multi-assignment).

Regarding the differences between the classification generated by means of the different algorithms, Louvain and VOS gave very similar classifications: in the number of coincident categories in the journal ranking by categories, in the indicators referring to journals that change their classification, in the number of journals that add a category, and in the number of journals that lose a category with respect to the initial SJR classification. The fact that these indicators offer very similar percentages under both classification schemes means that the workings and the yields are adequate for both algorithms.

The classification obtained by means of Ward´s clustering algorithm was the one providing best results in terms of the overlap of categories. Moreover, it is the one that gives the greatest number of new categories generated in comparison with that of the original SJR, with a total of 139. This finding points to a refinement in the final classification obtained. At the same time, this solution presented the lowest number of journals changing category, which may be understood as a more stable classification with respect to the original SJR one. In contrast, the forced application of the three

measures combined resulted in a drop in the number of journals, which also bears an impact on the final number of journals per category.

Evaluation of the results of classification

In *articles 1, 2* and *3*, the different solutions for classification created were compared with other classification systems in order to gain a more general view of how effectively the methodological procedures worked. This evaluation was carried out applying the following indicators to each one of the classifications generated: (i) Number of journals classified; (ii) Number of categories; (iii) Mean number of journals per category; (iv) Mean number of categories per journal; (v) Percentage of overlap; (vi) Number of journals that change their classification; (vii) Number of journals that add categories; and (viii) Number of journals that lose categories.

In turn, the proposal dealt with in *article 4* served to analyze and corroborate the effectiveness of the information visualization techniques in terms of validating and contrasting the classification results. The graphs generated with VOSViewer software —which integrates the clustering algorithm used to detect communities using VOS and an algorithm of its own to visualize result— allowed us to compare the results of direct clustering with the maps and graphs generated by the VOSViewer program. These maps stand as a logical and organized representation of the structure of the SJR journal network, in agreement with the previously created structure of clusters.

At first analytical glance, the structured journal map served to locate aggregations of journals reflecting the subject areas of the SJR. Most belonged to relatively coherent and well-defined groups of journals presenting considerable similarity with areas of the SJR, such as 'Mathematics', 'Physics', 'Agriculture and Biology', 'Neurosciences', 'Psychology', 'Medicine', 'Social Sciences', or 'Arts and Humanities', among others. The general map presents a high consistency not only regarding the journals that make up the different aggregations, but also the interrelation, the order, and the final pattern in which the different aggregations identified appear. A closer look at the logical groupings of subject matter such as 'Mathematics' or 'Library and Information Science' confirmed the validity and reliability of information visualization as a means of contrasting and validating the classification systems that had been created.

Having analyzed results, responding to the research questions and determining the pros and cons of the different proposals involved, it is time to extract the most relevant conclusions from this work.

- All the proposals presented can be considered viable solutions that improve upon the classification and assignment scheme of journals to SJR categories. Yet they have certain limitations, such as the exclusion of some journals in the process of classification. The most appropriate final solution would be a combination of the different procedures applied.

- The loss of journals in the classification processes cannot be attributed to the automatic techniques used (clustering algorithms or citation analysis). Rather, it is due to the parameters that configure the methodological procedures, e.g. the combination of measures, the threshold of bibliographic references or minimum number of articles per journal, or the minimum cluster size established for each experiment. The use of a completely automated classification procedure would therefore be limited by intrinsic aspects of the methodology itself, and could be overcome by human intervention.
- The mixture of measures based on citation seems to have a positive influence on the final classification of the journals, giving rise to the appearance of aggregations of journals that are more consistent and solid. In the case of the experiments with clustering algorithms, the combination helped to maximize the cluster effect.
- The tagging system used made it possible to mix SJR category tags and text, lending stability to the more consolidated categories while allowing for the integration of new categories deriving from the analysis of text within the journal titles. Tagging also allowed for the multi-assignment of journals, hence the overlap of certain categories. The methods used to validate the classifications proved useful, above all to enhance overlapping, concentrations of journals in the categories, etc. However, more specific techniques should be applied to test the refinement and the improvement of ascription of journals to categories, such as evaluation by a panel of experts.
- Any classification process is oriented toward objectives established within a specific framework. It is difficult to design, develop and implement a classification system that does not pertain to the intended context, or is not tied to the

objectives originally set forth. The ultimate aim should be the adequacy of the system together with precision in the assignment of the entities or objects classified. This clearly calls for a pragmatic focus.