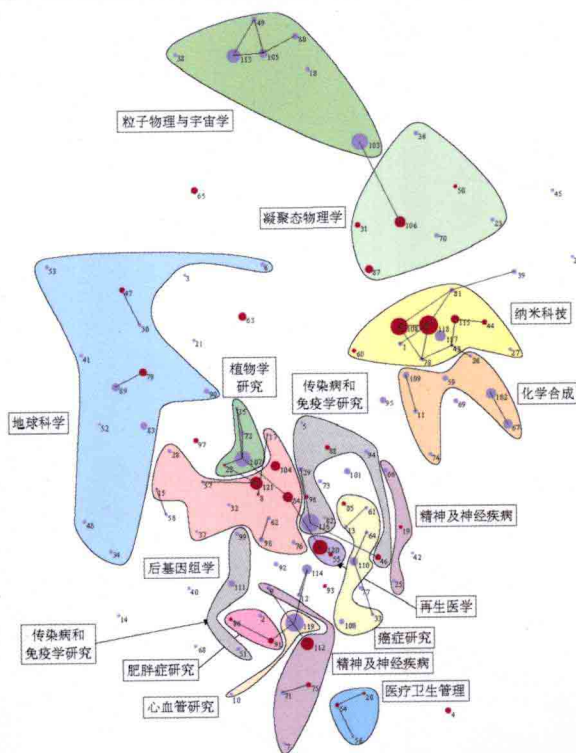


Mapping Science Structure

科学结构地图 2009

潘教峰 张晓林 王小梅 韩 涛 王 俊◎著



科学出版社
www.sciencep.com

Mapping Science Structure

科学结构地图 2009

潘教峰 张晓林 王小梅 韩涛 王俊◎著

科学出版社

北京

内 容 简 介

本书基于文献计量学理论,利用权威科学论文的同被引关系进行聚类分析,绘制了科学结构地图,分析了科学领域结构和各个学科领域以及主题的联系,揭示研究热点,发现潜在交叉领域,反映科学发展态势。本书进一步分析了研究领域的学科交叉性和新颖性,分析了中国和10个代表性国家在各个领域的活跃程度,分析了这些国家的国际合作情况。本书提供了一个客观和系统的分析方法,为科学家和决策者提供一种新视角来观察科学发展态势和科学研究结构。

本书可供各级科技管理人员、科技研究人员、高等院校师生阅读、参考。

图书在版编目(CIP)数据

科学结构地图 2009/潘教峰等著. —北京:科学出版社, 2010

ISBN 978-7-03-026546-3

I. 科… II. 潘… III. 科学体系学-研究 IV. G304

中国版本图书馆 CIP 数据核字 (2010) 第 014491 号

责任编辑:李晓华 卜 新 / 责任校对:刘亚琦

责任印制:赵德静 / 封面设计:无极书装

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2010年3月第 一 版 开本:787×1092 1/16

2010年3月第一次印刷 印张:14 1/4

印数:1—2 000 字数:328 000

定价:85.00 元

(如有印装质量问题,我社负责调换)



前言

当今世界,科学发展呈现出学科纵向深入和学科间综合、交叉、汇聚加速的特点,新方向、新前沿不断涌现。把握科学的总体结构,发现科学发展中不同领域的发生、发展、转移、汇聚、湮灭等规律,探测科学研究中新问题、新焦点和新领域的萌生、崛起、扩展、形成学科等过程,对于掌握科学发展规律、把握科学方向、发现科学前沿,都有重要意义。

人们曾经采用学科分类、主题聚类、共词聚类等方法,从不同角度促进了对科学结构及其变化的理解。本书利用科学研究论文之间存在的同被引关系,选取同被引程度高的核心论文,分析由同被引所形成的相互关联性,可在一定程度上排除论文或其登载期刊的学科分类、使用词汇、研究项目归类、作者或机构学科归属等影响,客观揭示出科学结构,帮助发现潜在的、尚未被科学界认识或承认的科学方向和前沿。我们借鉴经济合作与发展组织(Organisation for Economic Co-operation and Development, OECD)《抓住科学演化本质,制定新的科学指标及学科分布图》的方法,利用汤森路透科技集团的基本科学指标数据库(Essential Science Indicators, ESI)中的2002~2007年高被引论文和科学引文索引数据库(Science Citation Index, SCI)中的施引论文,揭示出由同被引关系聚类形成的121个研究领域及其相互关系,分析了这些领域的学科交叉性和新颖性,分析了中国和10个代表性国家在各个领域的活跃程度,希望为把握科学结构、发现潜在热点等提供参考。

需要指出,这项研究带有探索性。引用关系只是揭示了科学研究中复杂关系的一个侧面,引用关系所汇聚的领域往往涉及交叉、变化和非共识

的方向，引用数据滞后于当前科研活动，且我们所使用的数据比较偏重基础科学和生物医学领域，再加上我们也是初次开展这方面的研究，所以分析结果肯定存在不完善之处，我们希望今后改进、优化和深化，更好地揭示科学结构及其演变规律，支持科技决策和科技创新。

潘教峰

2009年11月

目 录

前言

第一章 引 言	001
第二章 研究方法	004
一、理论基础	004
二、数据来源	005
三、利用聚类算法确定研究前沿和研究领域	006
四、可视化方法	007
五、研究领域的内容分析	008
六、热点研究领域的确定	009
第三章 绘制科学结构地图	012
第四章 研究领域的学科交叉性度量	019
一、研究领域与 ESI 学科的关系图	019
二、在科学结构地图中观察学科交叉性	025
第五章 研究领域的新颖性度量	026
第六章 中国及代表性国家的份额度量	030
一、世界科学格局	031
二、按研究大类观察国家份额	034

三、按 121 个研究领域在科学结构地图中观察国家份额	046
四、小结	082
第七章 中国及代表性国家的国际合作度量	083
一、世界总体情况	083
二、按研究大类度量国际合作	085
三、按 121 个研究领域在科学结构地图中度量国际合作	091
四、小结	106
第八章 结语	107
附录 A 121 个研究领域列表	108
附录 B 121 个研究领域的内容分析	112
附录 C 各国在各 RA 中的核心论文份额及国际合著情况	179
附录 D 分数计数与整体计数的比较	219
致谢	221

第一章 引言

对科学知识体系的结构及其演化规律的研究一直是人们高度重视的课题。近期，科学文献与专利的获取和分析能力得到了空前发展，为绘制科学知识结构地图提供了有利的条件，开辟了研究科学知识结构和演化的新途径。

本书运用文献计量学的理论和方法绘制科学结构地图，力图反映当前科学发展态势，分析科学研究特性，考察世界主要国家科学研究优势领域，揭示科学研究热点领域，以期捕捉科学研究活动质变提供客观的参考途径，为科学家和决策者提供一种新视角。

本书的理论基础是科学论文同被引分析。论文引证是科学家们信息交流的一种直接表现，这些交流体现了知识的创造和流动。同被引则指一组论文共同被其他论文引用，反映了该组在学科分类、发表期刊、作者机构、研究项目等方面可能毫无关联的论文存在着某种关系，而当同时引用该组论文的文章数增加时，往往反映该组论文的内在关联不断加强。因此，同被引现象从一个角度反映了科学研究活动的聚合关系。早在 20 世纪 70 年代，普赖斯（Derek de Solla Price, 1922~1983）就提到同被引分析可以反映科学研究所形成的隐形科学共同体——无形学院（invisible college）。

通过同被引，具有相同或相似主题的论文被聚合在一起，代表了科学家通过论文引证的交流而自发涌现的研究领域，反映了科学结构的演变与自组织。利用同被引关系的聚类分析，使我们可以超越传统的学科分类来了解科学研究的结构及其变化，构建科学结构地图。基于科学结构地图的进一步分析，还可以从不同角度诠释科学发展特性——对跨学科和多学科特性的度量有助于监测科学研究交叉融汇，对各个研究领

域引证强度及其时间变化规律的分析可以揭示新兴或高速发展领域,等等。上述分析方法为我们提供了一个探索科学结构本身及研究领域内部相互关系和发展脉络的全新思路。

本书以汤森路透科技集团的 ESI 为信息源,提取了 2002~2007 年 6 094 个研究前沿,其中含 38 117 篇高被引论文,通过同被引分析对它们进行聚类,得到了 121 个研究领域,即相互关联的高被引论文簇。通过这些研究领域的关键词分析以及相关领域专家对这些研究领域的分析,确定了这些研究领域的名称。

本书方法参照了 OECD 和日本科学技术政策研究所 (NISTEP) 关于科学结构演化的类似研究。^{①~③}

本书的结构如下:第一章,引言;第二章,研究方法,绘制科学结构地图的理论和方法;第三章,绘制科学结构地图,展示了 121 个研究领域及 14 个研究大类的绘制结果;第四章,研究领域的学科交叉性度量,通过 121 个研究领域与 ESI 的 22 个学科之间的关系揭示研究领域的学科交叉性;第五章,研究领域的新颖性度量,通过各个研究领域平均年的分布情况观察研究领域的新颖性;第六章,中国及代表性国家的份额度量,通过观察中国及代表性国家在每个研究领域中的论文份额分布讨论各国在不同研究领域的相对优势;第七章,中国及代表性国家的国际合作度量,通过国际合著分析提供对国际知识流的度量;第八章,结语。

本书术语解释:

高被引论文 (highly cited paper): ESI 对过去 10 年 SCI 数据库中的论文被引频次进行统计,将 22 个学科领域中论文被引频次最高的前 1% 论文遴选为高被引论文。本书使用了 2002~2007 年的高被引论文。

研究前沿 (research front): ESI 以 SCI 前 6 年的高被引论文为基础,利用论文之间的同被引关系聚类产生一系列论文集合。

研究领域 (research area): 在研究前沿基础上的再次聚类得到一系列高被引论文集合。

同被引 (co-cited): 一组论文共同被其他论文引用。

核心论文 (core paper): 研究领域中的高被引论文。

热点研究领域 (hot research area): 包含 4 个及以上热点研究前沿且包含热点研究前沿的比例排在前 50% 的研究领域。

① Masatsura Igami (OECD), Ayaka Saka (日本科学技术政策研究所). 2007-04-25. Capturing the evolving nature of science, the development of new scientific indicators and the mapping of science. <http://www.oecd.org/dataoecd/11/40/38134903.pdf>

② NISTEP. 2007-07-24. Science map 2004. <http://www.nistep.go.jp/achiev/ftx/eng/rep100e/pdf/rep100e.pdf>

③ NISTEP. 2009-02-02. Science map 2006. <http://www.nistep.go.jp/achiev/ftx/eng/rep110e/pdf/rep110e.pdf>

重力模型 (gravity model): 对象之间关联程度的可视化方法。基本思想是把对象间关联强度看做原子之间的引力和斥力。当力平衡时, 描绘各个对象之间的相对位置。

平均年 (mean year): 一组论文出版年的平均值。

第二章 研究方法

本书的理论基础是科学论文同被引分析，通过对高被引论文的同被引关系进行聚类分析，产生若干研究领域；对所有研究领域进行内容分析，并从中确定热点研究领域；利用重力模型可视化方法绘制当前科学研究活动中研究领域的结构图，以期帮助了解科学结构演变趋势。

一、理论基础

科学论文间的引用反映了科学研究的动态交互。同被引是指一组论文共同被其他论文引用。如图 2-1 所示，论文 A、B、C 同时被论文 1、2、3 引用。如果论文 A 和论文 B 频繁同被引，可以推测它们拥有相同或相近的研究主题。

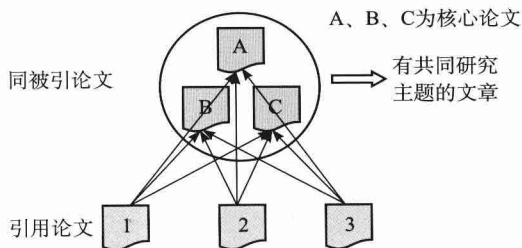


图 2-1 通过同被引分析确定研究领域

使用同被引的方法，计算高被引论文两两之间的同被引关系，并根据同被引关系对高被引论文进行聚类形成若干论文簇，称为“研究前沿”（research front, RF）；在此基础上利用同被引关系对上述 RF 再次进行聚类，得到的若干论文簇，称为“研究

领域”(research area, RA)。高被引论文、研究前沿以及研究领域之间的关系如图 2-2 所示。

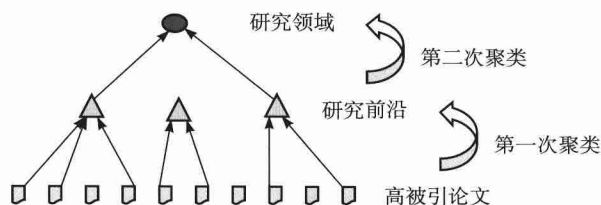


图 2-2 高被引论文、研究前沿、研究领域的关系

二、数据来源

本书使用的高被引论文和研究前沿取自 ESI 2008 年 7 月公布的 2002 年 1 月至 2008 年 4 月的数据。其中，高被引论文 53 892 篇，经过一次聚类的研究前沿 6 094 个，在研究前沿中包含 38 117 篇高被引论文。引用这些高被引论文的论文集合来自 SCI 2002~2007 年光盘。这三个数据集都来源于汤森路透科技集团。

表 2-1 研究前沿中的高被引论文在 ESI 的 22 个学科中的逐年分布 (单位: 篇)

ESI 22 个学科	2001 年	2002 年	2003 年	2004 年	2005 年	2006 年	2007 年	2008 年	总计
农业科学	1	60	104	94	115	148	94	0	616
生物与生物化学	0	249	364	388	367	419	347	15	2 149
化学	4	478	807	922	891	1 103	733	0	4 938
临床医学	4	841	1 333	1 340	1 460	1 519	1 084	0	7 581
计算机科学	2	78	177	255	269	242	95	0	1 118
经济与商业	4	72	102	120	124	99	84	0	605
工程科学	3	300	484	528	603	607	358	0	2 883
环境/生态学	0	106	162	186	191	219	156	33	1 053
地球科学	0	100	165	212	226	264	219	49	1 235
免疫学	0	52	75	79	78	90	101	21	496
材料科学	1	163	269	309	329	343	273	0	1 687
数学	4	96	152	155	169	184	75	0	835
微生物学	0	72	101	124	123	131	110	6	667
分子生物学与遗传学	0	111	174	202	196	205	224	15	1 127
多学科	0	7	10	14	14	17	17	3	82
神经科学与行为科学	0	121	204	223	225	228	193	14	1 208
药理学与毒物学	1	62	90	87	101	112	98	0	551
物理学	4	489	686	756	740	901	775	38	4 389
植物学与动物学	1	212	344	373	351	467	327	0	2 075
精神病学/心理学	0	95	149	161	169	186	157	19	936
社会科学	4	169	247	231	266	262	138	0	1 317
空间科学	0	43	91	102	97	110	114	5	562

表 2-1 显示了 38 117 篇高被引论文在 ESI 的 22 个学科中的逐年分布情况。ESI 将 6 094 个研究前沿划分到 ESI 的 22 个学科中，每个研究前沿至少属于 1 个学科，至多

属于 2 个学科, 图 2-3 显示了 6 094 个研究前沿在 ESI 的 22 个学科中的分布情况, 其中与“生命科学”相关的学科包括生物与生物化学、临床医学、免疫学、微生物学、分子生物学与遗传学、神经科学与行为科学、精神病学与心理学、药理学与毒物学、植物学与动物学在内的 9 个学科共有 2 981 个研究前沿, 占 48.9%。这反映了 SCI 数据库的学科结构不平衡, 来源期刊中生物医学类占较大比例。

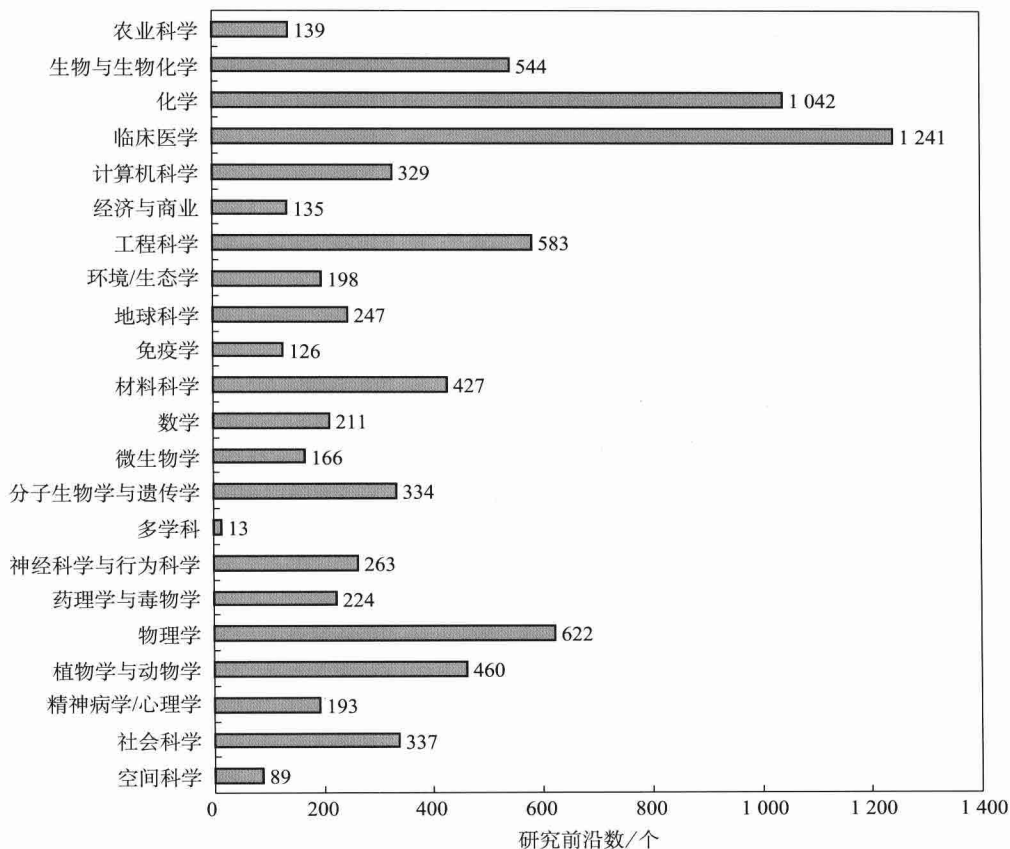


图 2-3 ESI 的 22 个学科的研究前沿数

三、利用聚类算法确定研究前沿和研究领域

研究前沿和研究领域都是高被引论文簇, 通过单连接聚类方法^{①②}确定。首先利用 SCI 论文对高被引论文的引用, 计算高被引论文两两之间的同被引次数。然后采用皮尔逊相关系数对同被引次数进行归一化处理, 形成论文间的同被引强度。最后, 设定

① H. Small, E. Sweeney. 1985. Clustering the Science Citation Index using co-citations. I. A comparison of methods. *Scientometrics*, 7 (3~6): 391~409

② H. Small, E. Sweeney, E. Greenlee. 1985. Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, 8 (5, 6): 321~340

聚类参数,进行单连接聚类,形成高被引论文簇。

聚类过程分为两个步骤:

第一步,确定研究前沿。本书的研究前沿直接取自 ESI 提供的 2002~2008 年的研究前沿,共 6 094 个,其中包含 38 117 篇高被引论文。

第二步,确定研究领域,对研究前沿进行二次聚类。聚类方法与第一次基本相同,主要的区别是此时研究前沿被当成一个“超级论文”来处理。研究前沿的被引用次数按如下方法确定:引用研究前沿中高被引论文的所有论文视作引用该研究前沿的论文,这些施引论文去重后的数量视作该研究前沿的被引次数。

单连接聚类方法的使用中主要有四个参数需要设定:①簇内成员间关系阈值;②关系阈值增量;③簇内成员数最大值;④簇内成员数最小值。这四个参数决定了聚类结果:

- 1) 簇内成员间关系阈值决定簇内成员最小关系值。
- 2) 关系阈值增量、簇内成员数最大值和最小值决定簇内成员数的多少。
- 3) 四个参数共同决定了簇的数量。

这四个参数的设定需要在聚类过程中反复调整,以适应特定的数据集,并结合聚类结果的判读确定一组最佳参数。本书在此次聚类中设定的四个参数为:①簇内成员间关系阈值为 0.02;②关系阈值增量为 0.009;③簇内成员数最大值为 100 个研究前沿;④簇内成员数最小值为 6 个研究前沿。研究前沿二次聚类后产生了 121 个研究领域,其中包含 2 300 个研究前沿的 18 203 篇高被引论文(即核心论文)。这些研究领域成为本书的分析对象。

四、可视化方法

本书通过重力模型^①形象描述研究领域之间的相互关系。重力模型的基本思想是:原子之间存在引力和斥力,所有原子达到受力平衡时,原子之间的相对位置将保持稳定。借用重力模型的思想,可以把研究领域视作原子。两个研究领域之间的引力由以下公式计算:

$$N_{ij}^{\text{nom}} r$$

其中, N_{ij}^{nom} 是研究领域 i 和 j 之间的同被引强度, r 是研究领域 i 和 j 在图中的距离。因此它们之间的引力与同被引强度和距离成正比,两个研究领域之间的斥力由以下公式计算:

$$N_{\max}^{\text{nom}} r^2$$

其中, $N_{\max}^{\text{nom}}$ 是所有研究领域中最大的同被引强度。两个研究领域之间的斥力与这两个

^① N. Quinn, M. Breuer. 1979. A forced directed component placement procedure for printed circuit boards. IEEE Transactions on Circuits and Systems, 26 (6)

研究领域之间的距离成反比。

重力模型的计算流程是：首先，将所有研究领域（一个圆点代表）随机摆放在一个范围固定的二维空间中；其次，计算每个点受到的引力和斥力，根据受力情况并结合 Velocity Verlet Algorithm^① 对点进行位移，这里不考虑圆点半径，以圆的中心位置作为该点的位置进行计算，反复移动并在新布局下重新计算每个点的合力；最后，当作用到每个点的合力都小于预设的最小值时则结束。通过重力模型计算得到的是各个研究领域间的相对位置，通过相对位置表现研究领域间的相互关系：聚集在一起的研究领域有相似的研究内容，距离近的研究领域相关性强。

从重力模型得到研究领域的相对位置，而相对位置如何展现是可视化的一个重要内容。本书采用两种方式展现研究领域相对位置：一种是点线图，另一种是地形图。两者都使用重力模型得到的相对位置对研究领域进行布局，区别在于，前者用圆圈的半径表示研究领域的扩展度（即研究领域中包含核心论文的数量），而后者使用了高斯函数表示研究领域的扩展度。用圆圈表示研究领域的扩展度会有一些问题，因为研究领域中包含的核心论文数最大/最小之比达到了 44，采用圆圈难以真实地反映其大小，而采用地形图的方法可以避免这个问题。具体方法是，把研究领域的核心论文以映射程序（重力模型）求出的位置为中心呈高斯函数状态分布。科学结构地图中的核心论文分布情况，表现为对应各研究领域所得的高斯函数的叠加 $\rho(x, y)$ 。 $\rho(x, y)$ 对全平面的积分值就是全部核心论文数。实际进行可视化时，为了用计算机在二维平面上表示核心论文的分布，把科学结构地图分割成面积为 $dx \times dy$ 的网格。然后，计算出网格内包含的核心论文数，核心论文的密度（核心论文数 / ($dx \times dy$ ））则近似于 $\rho(x, y)$ 。研究领域地形图就是把各网格用核心论文密度对应颜色着色后得到的结果。

五、研究领域的内容分析

对高被引论文的聚类产生了 121 个研究领域，因为每个领域是一个论文簇（包含多篇密切关联的论文），确定这些研究领域的研究内容并给它们命名是很有必要的。本书采用专家分析法进行研究领域内容分析，先由相关领域专业人员参考科学结构地图和核心论文列表对研究领域命名与判读，然后请领域专家对上述判读结果进行审核并提出修改意见。

判读工作包括两个方面。

（一）研究领域命名

根据该研究领域所包含的论文内容，用一个短语为研究领域命名，以准确描述研

^① W. C. Swope, H. C. Andersen, P. H. Berens et al. 1982. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. Journal of Chemical Physics, 76: 637~649

究领域的内容。

(二) 研究领域内容与结构判读

1. 研究领域中的研究前沿的分组

一个研究领域包含若干个研究前沿，将研究领域中的研究前沿按照内容的相似性划分成若干组，在“该 RA 中包含 RF 的图谱”中进行标识，并对每一组研究前沿的内容做简要描述。图 2-4 (a) 是提供给专家的、未对研究前沿分组的图谱；图 2-4 (b) 是专家在内容判读后，把 10 个研究前沿分成 3 组并分别给出内容简述而形成的新图谱。

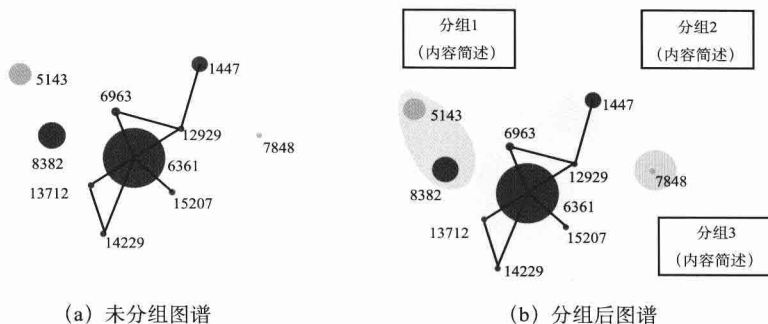


图 2-4 “研究领域”中的“研究前沿”分组命名示意

2. 研究领域重点研究前沿标识

如果研究领域中的若干研究前沿无法分组，就将重要的研究前沿（与其他研究前沿的关联较多，或者是热点研究前沿）标出，并给出内容简述。

表 2-2 是研究领域（ID50）内容分析结果的实例。

六、热点研究领域的确定

热点研究领域是活动频度高的研究领域，由以下方法确定。

首先，确定热点研究前沿，计算所有研究前沿 2002~2007 年的篇均被引用次数及其年均增长率，同时计算 ESI 中 22 个学科的篇均被引用次数和年均增长率。被引用次数及年均增长率均大于所属学科平均值的研究前沿被视作热点研究前沿。本书从 6 094 个研究前沿中确定了 1 794 个热点研究前沿，约占 29.4%。图 2-5 显示 ESI 的 22 个学科中热点研究前沿的分布。

其次，通过热点研究前沿确定热点研究领域，评判标准有两条：包含 4 个及以上热点研究前沿，包含热点研究前沿的比例排在前 50%。在本书的 121 个研究领域中，有 35 个研究领域被确定为热点研究领域。热点研究领域的分布见第三章，具体内容参见表 3-1。

表 2-2 内容分析实例

RA 的名称	关于材料的磁电、铁电、多铁和介电性质的研究		RA ID	50
RA 的关键词	thin films; BiFeO ₃ thin films, ferroelectric polarization; multiferroic BiFeO ₃ ; in multiferroic BiFeO ₃ ; oxygen vacancies in the SrTiO ₃ substrate on the electrical properties			
包含 RF 数 (热点 RF 数)	核心论文数	被引用篇数	总被引次数	平均年
10 (7)	94	2 563	6 547	2004.9
该 RA 中包含 RF 的图谱				
RF 的关键词				
RF ID	关键词			
6361	epitaxial BiFeO ₃ multiferroic thin film heterostructures; multiferroic BiFeO ₃ films; BiFeO ₃ epitaxial thin films; (111) (c)-oriented BiFeO ₃ thin films; (111) BiFeO ₃ thin films			
8382	epitaxial ultrathin ferroelectric films; perovskite ultrathin films; ferroelectric phase transition; ultrathin perovskite films; strained BaTiO ₃ thin films			
5143	LaAlO ₃ /SrTiO ₃ interface; two-dimensional electron gas; high-mobility electron gas; magnetic oxides; atomic-scale perovskite titanate superlattices			
1447	apparent colossal dielectric constants; high dielectric constants; coarse-grained CaCu ₃ Ti ₄ O ₁₂ ceramics; lattice dielectric response; giant barrier layer capacitance effects			
6963	multiferroic BaTiO ₃ -CoFe ₂ O ₄ nanostructures; magnetoelectric CoFe ₂ O ₄ /Pb (Zr _{0.52} Ti _{0.48})O ₃ double-layer thin film prepared; Pb (Zr _{0.52} Ti _{0.48})O ₃ -Co _{0.9} Zn _{0.1} Fe ₂ O ₄ multilayered thin films prepared; multiferroic nanostructures; Pb (Zr _{0.52} Ti _{0.48})O ₃			
7848	multiferroic BiFeO ₃ ceramics finite-temperature properties; polar phonons, magnetodielectric effect; infrared			
12929	colossal magnetocapacitive coupling; magnetoelectric coupling; relaxor ferroelectricity; ferromagnetic CdCr ₂ S ₄ ; magnetocapacitance			
13712	thin film ferroelectrics spontaneous polarization epitaxial strain; multiferroic bismuth ferrite; oxygen vacancies			
14229	Cr-doped BiFeO ₃ thin films grown; enhanced ferroelectric properties; multiferroic properties; chemical solution deposition; La doping			
15207	monoclinic-to-monoclinic phase transition; evidence; BiMnO ₃ ; centrosymmetric crystal structure; likely occurrence			