科学结构分析方法优化研究

任红娟 著

KEXUE JIEGOU FENXI FANGFA YOUHUA YANJIU

中国出版集团 世界图书出版公司

科学结构分析方法优化研究

任红娟 著

世界图书出版公司
上海·西安·北京·广州

图书在版编目(CIP)数据

科学结构分析方法优化研究 / 任红娟著. —上海:
上海世界图书出版公司,2013.7
ISBN 978-7-5100-6339-8

Ⅰ.①科… Ⅱ.①任… Ⅲ.①科学体系学—研究 Ⅳ.①G304

中国版本图书馆 CIP 数据核字(2013)第 117495 号

科学结构分析方法优化研究

著　　者　任红娟

出 版 人　陆　琦
策 划 人　姜海涛
责任编辑　吴柯茜
装帧设计　车皓楠
责任校对　石佳达

出版发行　上海世界图书出版公司　　www.wpcsh.com.cn
地　　址　上海市广中路 88 号　　www.wpcsh.com
电　　话　021-36357930
邮政编码　200083
经　　销　各地新华书店
印　　刷　上海市印刷七厂有限公司　　如发现印装质量问题
开　　本　787×960　1/16　　请与印刷厂联系　021-59110729
印　　张　11.5
字　　数　150 000
版　　次　2013 年 7 月第 1 版
印　　次　2013 年 7 月第 1 次印刷
书　　号　ISBN 978-7-5100-6339-8/G·403
定　　价　35.00 元

序　言

美国著名哲学家卡尔纳普(Carnap)1928年在《世纪的逻辑结构》一书中指出:“一切科学的解释从根本上说都是结构的解释”。数字化和网络化的飞速发展扫清了科学交流过程中时间和空间的障碍,使得科学交流变得更加顺畅,也大大地促进了科学文献的膨胀。在科学交流体系中,文献是作者发表知识声明的重要平台,是作者进行知识扩散和知识转移的重要渠道,利用表征文献的内容和引用的各种特征来揭示隐含其中的有价值信息一直是文献计量学和科学计量学的重要研究内容,而准确和有效地识别科学结构更是研究的重中之重。科学结构是科学知识长期形成的、固有的、不以人们的意志为转移的客观存在,是科学内在逻辑结构的集中体现,既可以反映整体科学的逻辑关系,也可以针对某一研究领域形成其科学结构。

对于科学家来说,准确的科学结构可以帮助他们定位自己所需要的信息,了解学科研究的主流和分支,识别学科的交叉和融合,辨识学科发展的态势和前沿,为科学家从整体上把握领域全貌提供了强有力的支撑。同时,对于政策制定者而言,科学结构的分析有利于他们理解科学的复杂关系,能够更加有效地进行资源优化配置,

实施研究的规划和资助，从宏观上把握科学发展的脉络，推动科学的均衡发展。这一切都表明在信息急剧膨胀的环境中，准确、快速的识别领域的科学结构，对于确定领域的主要研究内容、领域的研究前沿和领域的发展态势具有非常重要的研究价值和意义。

揭示科学结构的方法有很多，包括各种共被引分析方法、文献耦合方法、直接引用和互引关系分析方法、共词方法、文本挖掘方法等，这些方法都能够从不同程度和不同层面揭示领域一定时间段的科学结构以及科学结构的演变。然而，这些方法大多都只从一个视角或一种关系来揭示不同层面的科学结构。马林斯(Mullins)在 *The Structural Analysis of a Scientific Paper* 一书中提到：要从科学文献的每一个方面来展开研究，从标题到参考文献，从图表到写作风格以及词的利用，分析文献的每个方面都能得到有价值的信息。莫里斯(Morris)在 *Mapping Research Specialties* 一文中用一张盲人摸象图形象地隐喻从一个特征或关系来刻画领域的科学结构的局限性和片面性。他认为每种方法都能够反映科学知识体系的一个侧面，尽管这些方法得到的结果不是完全分离的，但至少可以肯定是存在偏差的。詹森斯(Janssens)举了一个根据词“nano”文本相似的文献集，根据文献之间的引用关系无联系的实例，这是一个比较特殊但又很具代表性的例子，说明从文本的角度揭示的文献集之间的相似性不但片面，甚至可能是完全错误的，引入了很大的噪音。这些学者的观点都表明科学结构分析的方法亟须进行改进和优化。同时，在科学结构分析的整个流程中，利于数据集的构建、分析对象的选择、是否采用共现矩阵、共现矩阵形成方法、矩阵的归一化方法、科学结构划分方法的选择以及最终科学结构的解释等问题都有很多更优方案的选择，研究科学结构方法优化对于文献计量学方法体系的完善具有重要的意义。

本书首先对研究背景、相关的概念、主要的研究内容和研究方

法进行了说明，对当前的科学结构分析方法从“基于内容特征的科学结构研究”“基于引用的科学结构研究”和“基于融合方法的科学结构研究”三个层面进行了系统的梳理。根据本研究科学结构方法优化的方向，对文献的内容和引用特征融合优化方法从哲学、科学知识社会学和科学交流学相关理论为根基，确立了其理论基础。然后，从二元关系融合和多元分析融合以及操作流程优化的角度，从实验中证明了本研究提出方法的优化效果，同时在“观点挖掘”领域进行了应用研究，进一步对本研究提出的方法进行了验证，最后总结了本研究的主要研究内容和贡献，同时也提出了本研究的不足之处，以及后续的研究方向。

本书主要是在文献的内容以及引用特征的二元和多元融合方法上进行了科学结构方法的优化，并对科学结构的分析流程中会影响融合效果的方法也进行了筛选，但是本书提出的融合函数还比较有限，且需要在更大的数据集中进一步进行验证，并不是对所有的科学结构方法进行优化，对于科学结构优化效果方法的选择也需要更加科学和系统。

本书可以作为科学学、科学管理和文献计量学以及图书情报档案专业系统了解科学结构分析方法以及分析流程的辅助工具，也可以作为文献计量学科学研究人员进行科学结构分析方法优化的参考书。在当今社会，越来越多的学科面临着信息爆炸和知识缺乏的严峻局面，学习和掌握抽取领域科学结构、研究重点、热点和研究趋势的方法对于相关人员进行知识的学习和进行科学研究具有重要的研究意义，本书可以作为他们从更广层面上理解和学习文献计量学相关知识的工具书。

任红娟

2012 年 10 月

目　录

1 引　言

1.1 研究背景

历史学家阿尔弗雷德·克罗斯比(Alfred Crosby)曾经说过：形象化是促进所有的现代科学爆炸性发展的两个因素之一，而另一个因素是测量。① 而在20世纪末期开始逐步发展并快速成长的科学知识图谱研究，与阿尔弗雷德提出的促进现代科学发展的两个因素达到了非常好的契合和匹配。科学知识图谱是将传统的文献计量方法与现代的文本挖掘以及复杂网络理论、数学、统计学、计算机科学、社会学等学科领域的方法有机地整合在一起的一种综合分析科学发展的知识发现方法。②

科学知识图谱是我国学者对英文“mapping knowledge domain(s)”一词的翻译，而实际上英文“science mapping, knowledge domain(s) visualization, mapping(of) science”等词的内涵和外延与科学知识图谱是一致的。这种方法主要是从大量的抽象文献或者科学信息中，抽取出能够表征其内容、关系和演化等特征的对象，并利用一定的方法揭示出它们的关系，最后采用可视化的方法形象化表达出纷繁芜杂的信息

① 见林聚任《社会网络分析：理论、方法与应用》，北京师范大学出版社2009年版，第252页。

② 见陈悦、刘则渊《悄然兴起的科学知识图谱》，《科学学研究》2005年第23卷，第2期，第149－154页。

中间隐藏的有序关系，从而帮助人们从大量的杂乱无序的信息中识别主要关系，促进知识的理解、吸收以及知识的扩散、转移，继而促进知识创新的发展。早期的文献计量学方法都是科学知识图谱的雏形，主要是采用相对简单的手段对相对数量较小的信息进行处理，并从中识别出隐含的关系。然而随着信息处理技术和数字化技术的不断发展，为从大量的数据当中挖掘有效的信息并利用可视化的形式展示出来提供了更多的可能，资源的丰富和技术的支撑促使科学知识图谱研究在20世纪90年代后期迅速的发展和壮大起来。它作为文献计量学、科学计量学、网络计量学、可视化技术以及社会网络分析、统计物理学、数据挖掘和人工智能等多个学科方法融合的一个研究领域，无论是研究内容的丰富性、研究方法的广博性，还是从应用的广泛性以及分析效果的强大性而言，都是一种极具发展前景的研究领域。同时，科学知识图谱方法作为情报学方法的集中体现，其发展对于情报学的发展具有非常重要的推动作用。

情报工作需要大量的信息和知识作为支撑。面对浩如烟海的信息，如果能够很好地利用各种技术和方法把复杂的关系简单和明了化，不仅有助于情报工作更深入的开展，还更加有利于情报服务效率的提高。国家的创新体系、经济和社会的发展都需要情报工作来提供有力的支撑，帮助决策者和各种信息、知识需求的用户提供更加丰富、更加精准的信息一直以来都是情报工作的宗旨和目标所在。而科学信息，尤其是以期刊为代表的书面交流系统的信息的有效挖掘和揭示，对于了解领域的研究结构、研究前沿和研究重点、学科交叉、交流模式以及识别潜在的研究对象都是至关重要的。网络化、数字化时代的各种技术推动科学知识图谱从幕后走向了前台，从少数科学家的研究视角走向了逐步大众化的道路。情报学领域对于科学知识图谱方法并不陌生，如何把这个强大的知识发现工具融入到高层次的情报服务当中，对于情报学的发展，尤其是对情报深入的挖掘和分析工作来说非常关键。

在信息膨胀、情报知识匮乏和信息严重不对称的社会中，谁占有了

有效信息，谁就有可能成为财富和资源的拥有者，甚至有可能在某个领域独占鳌头。尤其是信息社会的到来，更使得信息作为一种重要的稀缺资源，成为现代社会的三大支柱之一。人们急切地盼望能够拥有更多的信息和知识，从而拥有更多的知识资本和财富资本。随着计算机技术和网络技术的飞速发展，信息瞬息传遍地球的每个角落不再是梦想，遵循“摩尔定律”的计算机硬件性能也在不断地提升，而价格却不断地跌落，给更多的人提供了利用计算机通过网络来获取资源的机会。受到数字化技术的冲击和推动，各种形式的内容资源都能够以数字化的形式获取，包括传统的出版单位、新闻媒体、政府等主体都积极地参与到数字化信息提供、发布以及传播的行列中来。而 Web 2.0 时代的到来，更是掀起了全民参与数字化内容生产和传播的浪潮，汹涌而至的信息洪流使得人们被淹没在信息的海洋之中，无法自由的呼吸。面对浩如烟海的信息，人们变得无所适从，不知道该如何在这样的信息汪洋中获取自己真正所需要的信息。虽然功能不断强大的搜索引擎带给人们很多的惊喜，但是我们也不难发现由于搜索引擎的查全率和查准率与我们的需求还有很遥远的距离，而且搜索引擎本身能够检索到的内容只是信息海洋的冰山一角，所以我们仍然对于如何获取全面的信息和相关的信息感到非常的茫然。“信息丰富，知识匮乏”的窘境愈发严重。而学科的交叉和融合以及学科的细分程度越来越高，又使得找到了解本学科或者本专业的“通才”成了天方夜谭。而科学知识图谱方法正是顺应了时代的需求而发展起来的，既能从宏观上了解一个学科或者领域的整体研究概貌，又能从更加细粒度来剖析领域发展细节的一个有力的工具。

因此，无论从科学知识图谱方法本身的潜力和知识发现能力，还是从国家、组织、个人的信息和情报需求来看，利用科学知识图谱方法来进行科学信息抽取和挖掘都是非常重要的。而利用科学知识图谱方法从纷繁和交织的科学文献集合中识别出主要的内容和关系，发掘有用的信息、汲取有价值的情报和信息、识别出其中的结构关系和揭示潜在的发展趋势等应用对于学科和科学的发展都至关重要，也是情报学研

究的一个永恒的主题。

1.2 问题的提出

任何一种方法都不是完美无缺的，而且大多数事物的发展都需要经历从萌芽到逐步成熟的过程，科学知识图谱方法也不例外。从其现在的研究方法体系来看，主要是基于文献计量方法的共词方法、各种共被引方法、文献耦合以及社会网络分析方法，利用传统的计量方法分析流程，采用多维尺度分析、各种聚类分析方法等对文献计量分析的结果进行展现。在文献计量学研究当中，多数研究注重传统方法在不同领域的应用，所以使得出现在20世纪六七十年代的方法流程和操作到现在依然一成不变或者只是有了细微的改动，对于方法本身的修正和改进关注的相对比较少。而方法的科学性是保证结果的准确性和可靠性非常关键的因素。科学发展的历史证明，任何一门科学的理论和应用研究，只有应用科学的研究方法，才能揭示事物的最本质的规律，建立完美的科学体系。而科学知识图谱研究继承了优秀的文献计量学的方法，从计量学的历史考验来证明其应用在该研究中的可靠性，人们在利用这些方法的时候无形中把方法的科学性作为应用的前提假设，因此也很少有人怀疑它们的可靠性和合理性。

2003年一批学者关于共现矩阵的相似性计量方法和共现矩阵是否需要归一化处理展开了一系列激烈的讨论，激起了学者对传统方法本身的有效性、科学性以及准确性研究的兴趣，国内外的很多学者相继对这些传统的科学知识图谱方法提出了质疑，特别是对关于数据处理过程中的方法以及结果解释的可靠性给予了更多的关注。

事物的复杂性是普遍存在的特性，而21世纪是复杂性的世纪，复杂性研究被预言将在新世纪获得重大突破①。作为复杂性研究的一个

① 见杨波《复杂社会网络的结构测度与模型研究》，上海交通大学2007年论文。

很重要的内容，信息融合技术是研究多源信息的加工和协同利用，形成多形式信息互补，获得对同一事物或目标更客观、更本质的认识的信息综合处理技术，信息融合的理论和方法以及思想对于科学知识图谱的研究主题拓展提供了新的视野。1991 年，布拉姆（Braam）等人提出在共被引聚类的基础上利用引用这些共被引聚类的引文来进行共词分析能够提高共被引分析的质量，但是这种融合的思想并没有引起足够的重视。2005 年，格伦尼松（Glenisson）等人利用全文文本挖掘方法和文献计量方法融合，证明了融合的方法比起单一的方法能够更好地揭示领域的专业主题，得到的主题分类更加的准确。近几年来，这些复合方法的研究也逐渐引起了一些学者的关注，但是从这些研究的数量和比例来看，在文献计量学的研究当中几乎可以忽略不计，虽然该方法已经初步展示了其优势，但是目前仍没有形成一定的研究气候，更进一步地研究各种融合方法对于文献计量学来说是十分必要的。笔者和张志强教授根据科学知识图谱近几年主题的演变也揭示出科学知识图谱方法逐步向融合方向发展的趋势，即把共词、共引以及各种可视化技术、数据挖掘技术融合在一起，注重方法的融合和数据源的融合以及数据处理方法的融合，是未来科学知识图谱发展的趋势所在。①

但是，科学知识图谱作为一个包容性非常大的研究领域和方法体系，存在的问题也有很多，需要很多领域的学者和专家不懈地奋斗和不断地努力。而本研究则主要关注科学知识图谱研究当中的极具代表性的科学结构的揭示和分析研究。

目前，揭示科学领域科学结构的方法有很多，包括各种共被引方法、文献耦合方法、直接引用关系、互引关系、共词方法、文本挖掘方法等，这些方法都能够从不同程度和不同层面揭示领域一定时间段的科学结构以及科学结构的演变。然而这些方法也正如科学知识图谱方法

① 见任红娟、张志强《基于文献计量的科学知识图谱发展研究》，《情报杂志》2009 年第 28 卷，第 12 期，第 86－90 页。

存在的一些问题一样，主要存在以下几方面的问题：

（1）采用单一的方法相对比较多，缺少方法的融合。主要是利用相对少量的高频研究对象从词或者从引用的角度来揭示领域的科学结构，例如作者共被引分析方法，利用领域的几十位高被引作者，形成作者共被引共现矩阵，然后利用一定的标准化方法对其进行归一化处理，并利用多维尺度分析方法或者层次聚类等方法，把结果利用坐标图或者树形图的可视化形式展现出来。

莫里斯在 *Mapping Research Specialties* 一文中形象地采用盲人摸象来隐喻采用单一计量方法揭示领域信息的有限性和片面性，如图 1－1 所示①。作者认为：mapping specialties 是一个非常复杂的问题，可以采

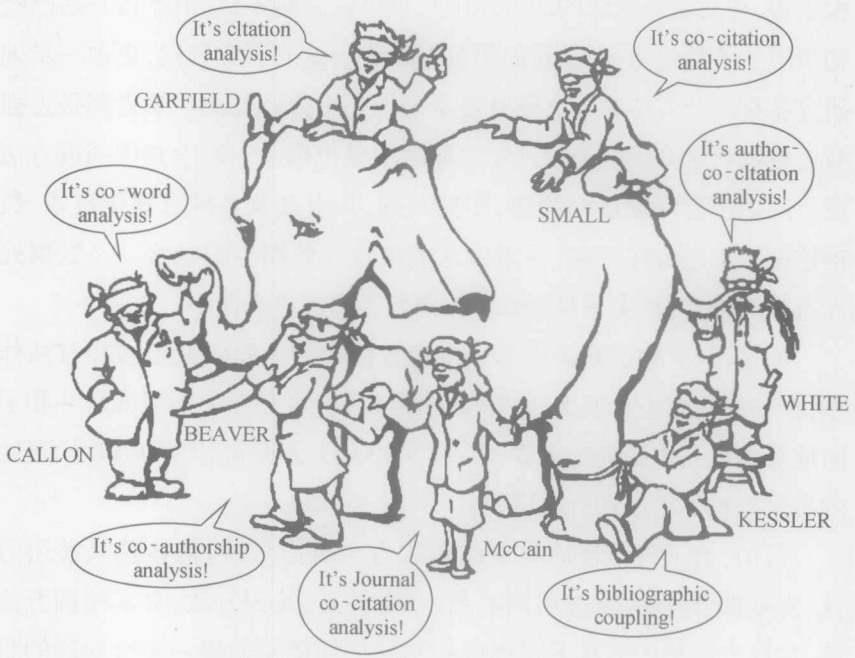

图 1－1　不同文献计量方法形成的盲人摸象图

（引自：Morris. Mapping Research Specialties. ARIST, 2008, 42）

① Morris, Mapping Research Specialties, ARIST, 2008(42).

用许多方法，而每种方法都会发现专业本身的一些不同的方面。例如，共作者分析可以发现专业的科学家之间的协作模式，共被引分析可以发现专业基础知识的结构，文献耦合可以发现研究的子主题。任何一种方法都不可能全面地分析整个研究专业，每一种方法得到的结果就如同盲人摸象一样，只是反映了专业的一个侧面的信息，尽管这些方法得到的结果不是完全分离的，但是可以肯定是存在偏差的。

此外，如图 1－2 所示，詹森斯（Janssens）举了一个根据词“nano”文本相似的文献集，根据文献之间的引用关系无联系的实例，这是一个比较特殊但又很具代表性的例子，说明从文本的角度揭示的文献集之间的相似性不但片面，甚至可能是完全错误的，引入了很大的噪音①。

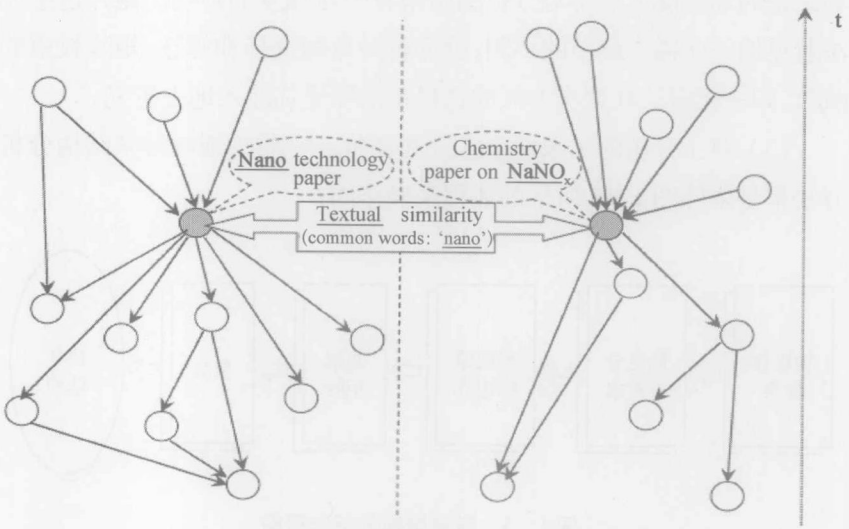

图 1－2 基于文本强相似和基于引用无联系的示例图

（引自：Janssens F. Clustering of Scientific Fields by Integrating Text Mining and Bibliometrics. 2007）

① F. Janssens, Clustering of Scientific Fields by Integrating Text Mining and Bibliometrics, 2007.