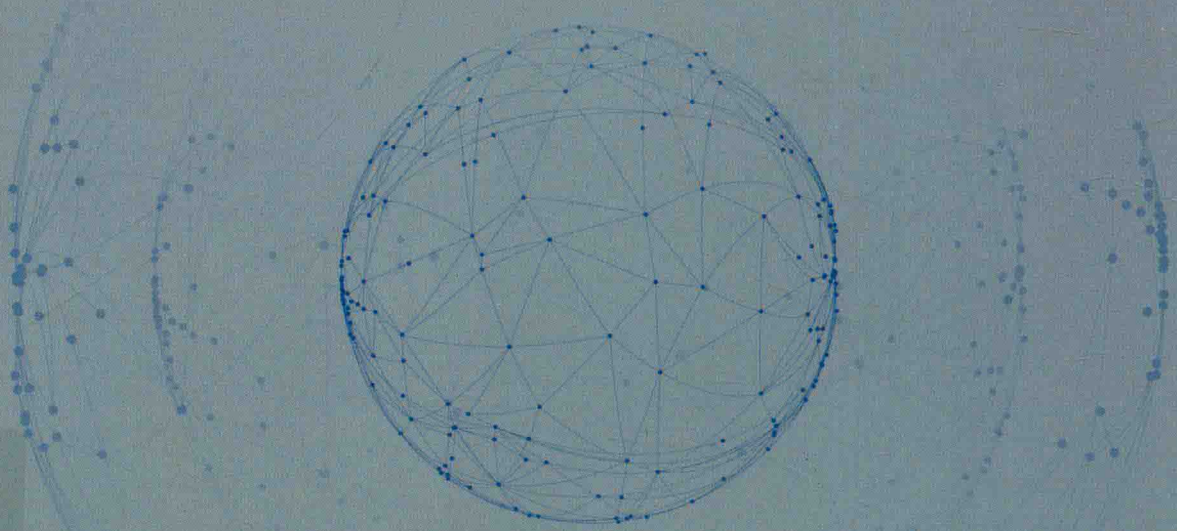
科学知识图谱原理及应用

——VOSviewer 和 CitNetExplorer 初学者指南

Principles and Applications of Mapping Knowledge Domains
A Beginner's Guide to VOSviewer and CitNetExplorer

李 杰 著

本书顾问：
（荷）尼斯·杨·凡·艾克（Nees Jan van Eck）
（荷）卢多· 瓦特曼（Ludo Waltman）



高等教育出版社

科学知识图谱原理及应用

——VOSviewer 和 CitNetExplorer 初学者指南

Principles and Applications of Mapping Knowledge Domains

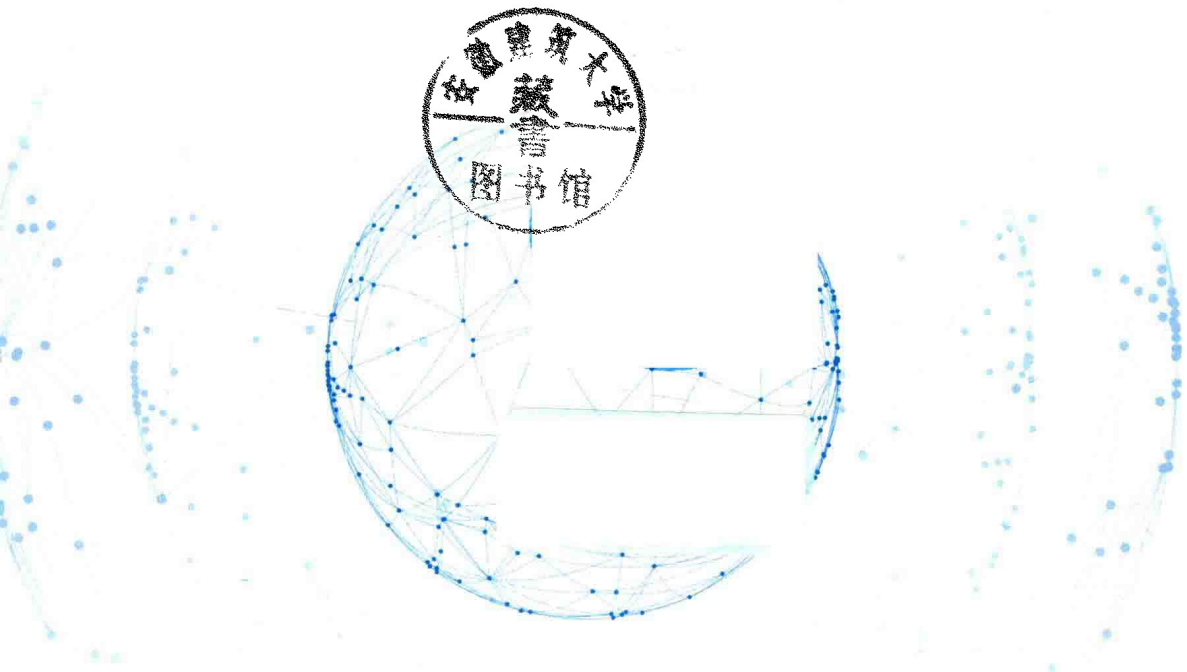
A Beginner's Guide to VOSviewer and CitNetExplorer

李 杰 著

本书顾问：

（荷）尼斯·杨·凡·艾克（Nees Jan van Eck）

（荷）卢多· 瓦特曼（Ludo Waltman）



高等教育出版社·北京

图书在版编目（CIP）数据

科学知识图谱原理及应用：VOSviewer和CitNetExplorer初学者指南 / 李杰著. -- 北京：高等教育出版社, 2018.8

ISBN 978-7-04-049166-1

Ⅰ. ①科… Ⅱ. ①李… Ⅲ. ①数据处理软件-高等学校-教学参考资料 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第331272号

科学知识图谱原理及应用

——VOSviewer和CitNetExplorer初学者指南

KEXUE ZHISHI TUPU YUANLI JI YINGYONG

——VOSviewer HE CitNetExplorer CHUXUEZHE ZHINAN

策划编辑 李雅悠　责任编辑 徐 阳 李雅悠　封面设计 张 志　版式设计 张 志

插图绘制 邓 超　责任校对 窦丽娜　责任印制 毛斯璐

出版发行 高等教育出版社　网　址 http://www.hep.edu.cn

社　址 北京市西城区德外大街4号　http://www.hep.com.cn

邮政编码 100120　网上订购 http://www.hepmall.com.cn

印　刷 高教社（天津）印务有限公司　http://www.hepmall.com

开　本 787mm × 1092mm 1/16　http://www.hepmall.cn

印　张 13.75

字　数 280千字　版　次 2018年8月第1版

购书热线 010-58581118　印　次 2018年8月第1次印刷

咨询电话 400-810-0598　定　价 98.00元

物 料 号 49166-00

序

信息技术溪流已淌过计算机丛林，经互联网汇成滚滚洪流形成“数字地球”，并海纳百川积成“大数据”海洋。于是奇景涌现：波普尔“三个世界”理论中，由“世界2”（精神世界）认识“世界1”（物理世界）而形成“世界3”（知识世界）的顺序，被颠覆为通过“世界2”认识“世界3”来认识“世界1”的本质与规律；“数据密集型科学发现”范式横空出世，继实验科学、理论科学、计算科学范式而成为科学研究的“第四范式”……科学知识图谱理论与应用技术随信息技术大潮在21世纪初应运而生、迅速发展，并向着“知识发现”的问题导向性应用、专题应用性交互设计、通用性软件开发等需求洼地呈梯度式奔涌，势不可挡。

该书正是科学研究第四范式的一种实现手段和工具。它基于特定大数据，按问题导向进行数据关联分析，力求发现事物的本质和关系，通过可视化技术显示知识单元或知识群之间的网络、结构、互动、交叉、演化、衍生等复杂关系，从而解决大量具体的知识发现问题。正如刘则渊教授指出的：科学知识图谱（Mapping Knowledge Domains，MKD）是显示科学知识的发展进程与结构关系的一种图形，是科学计量学从数学表达转向图形表达，由知识地图转向以图像展现知识结构关系与演进规律的结果。作为知识发现工具，VOSviewer长于主题挖掘、文献耦合、共被引以及合作网络等分析，CitNetExplorer可进行文献时序引证网络分析，能有效地揭示科学知识的结构关系和发展进程并予图像化展现，满足相关的应用需要。

该书作者长期潜心于安全科学原理、科学计量学及科学知识图谱应用交叉领域的研究，与国内外相关知名学者交流合作频繁，获博士学位之际已发表论文50余篇，包括在《科学计量学》（*Scientometrics*）、《安全科学》（*Saferty Science*）等国际权威期刊发表的数篇高水平SCI论文，已出版4部专著，应邀到多所高校和研究机构进行知识图谱相关交流和学术讲座。

科学研究第四范式之数据驱动型科学发现，将引领研究者把握科技前沿领域、研究热点，适应未来科技方向的需求，该洼地性需求势必大力推动科学知识图谱原理、方法、可视化软件工具的快速发展，由此带来MKD应用文献数量的暴发性增长、MKD实现软件的可视化和互动设计不断完善以及MKD通用性互动设计软件的发展。无论您有上述何种需求，都可从该书中获得相关知识和灵感。可以说，第四范式将成为科学研究的新常态，该书则是第四范式践行者的科学寻宝指南，有志的工具使用者、技术开发者都将由此获得学术和技术先机，勇立科技创新潮头，强国惠民，让人类生活更幸福，让世界更美丽。

上海市科学技术情报学会副理事长、教授

陈伟炯

2018年3月于上海浦东

前　言

在科学计量学理论、大数据和数据可视化技术的推动下，科学知识图谱的理论和方法在近十年来取得了飞速的发展。特别是在科学知识图谱的发展过程中，相继诞生了一批操作简便、功能强大、结果可靠的科学知识图谱分析工具。这使得该方法不仅仅局限在科学计量领域内的实践研究上，也为其他领域的科学知识图谱的绘制提供了广阔的空间。当前，科学知识图谱在医学、体育、教育以及经济学等多个领域有了系统性的应用研究，为这些领域的学者带来了理解知识的新模式。这种以可视化的方式去认识、理解和解读相关知识域的方法之所以受到科研人员的青睐，说到底就是因为它为我们提供了一种新的认识知识世界的方式。

据笔者不完全统计，目前直接或间接用于科学知识图谱绘制的工具不少于30种。这些工具各具特色，都融入了开发者和设计者的心血。本书介绍的VOSviewer和CitNetExplorer以及与之关联的科学知识图谱工具（HistCite、CRExplorer和RPYS i/o等），是笔者在这些工具群中有意挑选的一组。本书重点介绍的VOSviewer和CitNetExplorer的开发者凡·艾克（Nees Jan van Eck）和瓦特曼（Ludo Waltman）来自享有盛誉的国际科学计量研究机构——荷兰莱顿大学科学技术研究中心。更值得一提的是，两位开发者都出生在1982年，且同年博士毕业于荷兰鹿特丹伊拉斯谟大学。在博士就读期间，他们就已经开始有关合作，从理论、算法和可视化上对科学知识图谱进行系统研究，并共同在多个国际知名期刊上发表系列论文。如今开发者凡·艾克已经是多个国际期刊的编委，且负责莱顿大学科学技术研究中心的IT部门。瓦特曼不仅是多个期刊的编委，而且在2014年10月1日成为国际权威期刊《计量情报学期刊》（*Journal of Informetrics*）的主编。在科学计量学界利奥·埃格赫（Leo Egghe）和罗纳德·鲁索（Ronald Rousseau）被称作科学计量界的“双子星座”[①]，那么凡·艾克和瓦特曼将很可能成为新的“双子星座”。

本书分为三大部分，共8讲内容：

第一部分：第0讲至第2讲是关于知识图谱分析的基础内容。该部分依次介绍了科学知识图谱的基本理论、方法和相关的历史资料，VOSviewer和CitNetExplorer的基本情况和数据分析的准备。

第二部分：第3讲至第5讲为VOSviewer1.6.4软件的详细介绍。主要从软件的下载、安装、基本原理以及核心的数据分析功能等方面对VOSviewer进行了全方位的系统介绍。

第三部分：第6讲和第7讲对文献引证网络分析软件CitNetExplorer进行系统介绍。与此同时，对加菲尔德[②]开发的引文历史分析软件HistCite以及莱兹多夫[③]近期开发的文献年谱分析工具CRExplorer和RPYS i/o进行了介绍。

---

① 埃格赫和鲁索为师生关系，两人于2001年获得国际科学计量学界的最高奖——普赖斯奖。
② 加菲尔德（Eugene Garfield）博士是SCI的创始人，1984年获得普赖斯奖。
③ 莱兹多夫（Loet Leydesdorff）是阿姆斯特丹大学教授，2003年获得普赖斯奖。

本书在附录中增加了常见的科学计量和图谱分析工具列表以及其他信息，以帮助读者更全面地认识和使用科学知识图谱的工具和方法。

本书在写作过程中得到了多方面的大力支持。首先，要感谢凡·艾克和瓦特曼。我在代尔夫特理工大学安全系的办公室与凡·艾克谈及出版该书的想法时，得到了他以及后来瓦特曼的大力支持。在写作过程中，两位更是提供大量的原始资料、意见和建议，使得本书能够顺利完成。感谢CiteSpace科学知识图谱软件的开发者陈超美教授，陈教授是全球科学前沿图谱研究的领军人物。从我2013年8月参加陈老师的科学知识图谱讲座到现在，陈老师持续地为我在科学知识图谱的学习和研究中提供帮助。感谢国内科学知识图谱研究的开创者大连理工大学刘则渊教授，刘老对我在知识图谱研究中的支持、鼓励以及无私的帮助，是我能够坚持进行知识图谱研究和实践的重要动力来源。感谢上海市科学技术情报学会副理事长陈伟炯教授在百忙之中阅读书稿，并拨冗作序。感谢科学计量学和知识图谱研究的同行：安徽财经大学魏瑞斌博士、中国科学院文献情报中心杨立英博士、浙江大学李江博士以及四川大学李睿博士，感谢他们为本书撰写的推荐语。最后，特别感谢我的爱人隗合佳女士，对于我完成的所有书稿，她永远是第一位读者。学习法律专业的她，总是能从不同的视角建议我去修改、补充、完善每一版文稿。

笔者衷心地希望，本书能够帮助读者绘制出可靠、清晰、美观的科学知识图谱，使读者节约出大量的时间来仔细分析、解读和总结知识图谱的意义，从而使绘制的知识图谱不仅被专业领域的专家认可，更能够长时间地为专业领域服务。

李杰

2017年夏于上海浦东

lijie_jerry@126.com

目　录

第0讲　认识科学知识图谱

0.1　科学知识图谱的兴起及发展概况

国外的科学知识图谱（Mapping knowledge Domains, MKD）绘制起源于2003年5月美国国家科学院组织的一次研讨会。当时会议的组织者和参与者包括了多个最知名的关于科学计量和数据可视化的专家学者，如史蒂夫·莫利斯（Steven Morris）、陈超美（Chaomei Chen）、尤金·加菲尔德（Eugene Garfield）以及凯蒂·博纳（Katy Börner）等。会议的议题包括数据库、数据格式和存取（Session 1: Data-bases, Data Format & Access）、数据分析算法（Session 2: Data Analysis Algorithms）、可视化与交互设计（Session 3: Visualization & Interaction Design）以及应用前景（Session 4: Promising Applications），共四个部分（如图0.1）。会议结束后，2004年4月6日，《美国科学院院刊》（*Proceedings of the National Academy of Sciences of the United States of America*, *PNAS*）发表了一期科学知识图谱专刊。

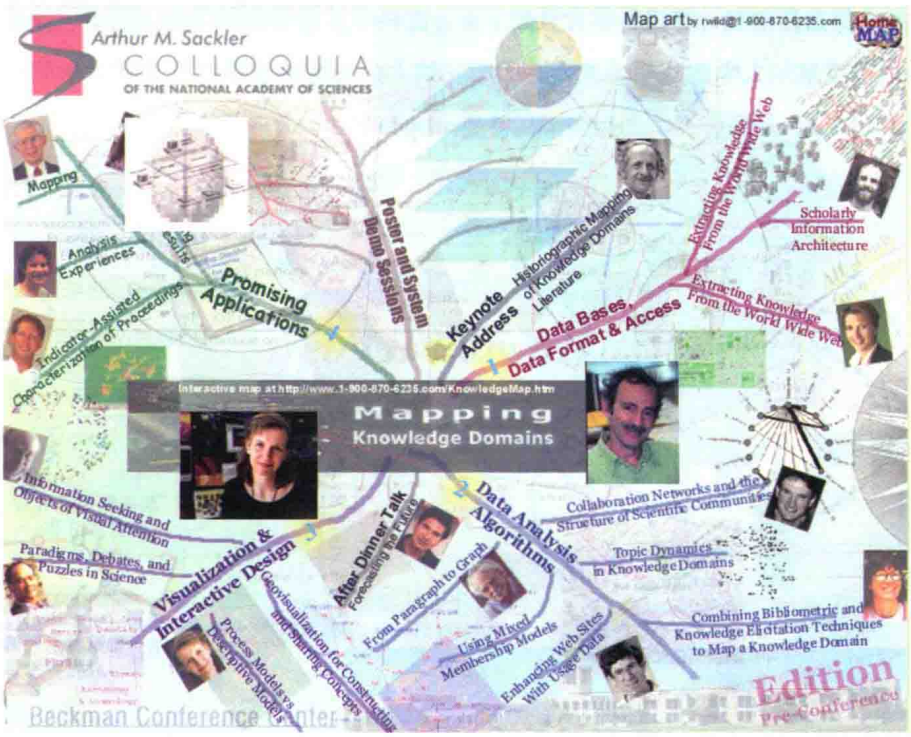


图0.1　2003年科学知识图谱会议的成员及主题[①]

① 参见www.1-900-870-6235.com/KnowledgeMap.htm，2015-11-12。

2004年4月10日，大连理工大学刘则渊教授受到《参考消息》上一篇题为《科学家拟绘制科学门类图》的文章启发，在国内率先带领自己的团队开始了科学知识图谱研究工作，并创建了大连理工大学网络-信息-科学-经济计量实验室（WISE Lab of DaLian University of Technology），为我国培养了一批专门从事科学知识图谱理论与实践研究的专业人才。刘则渊教授将科学知识图谱定义为：以知识领域为对象，显示知识的发展进程与结构关系的一种图形。科学知识图谱具有“图”和“谱”的双重性质与特征：既是可视化的知识图形，又是序列化的知识谱系，显示了知识单元或知识群之间网络、结构、互动、交叉、演化或衍生等诸多复杂的关系。知识图谱通常都是以知识网络形态展现的知识图形与知识谱系，它有许多不言自明的概念。

科学知识图谱研究以科学学为基础，是涉及应用数学、信息科学以及计算机科学的交叉领域（如图0.2），是科学计量学（Scientometrics）的新发展领域。2015年11月28日以检索式TOPIC：（"map* knowledge domain*"）OR TOPIC：（"Biblio* map*"），在Web of Science（简称WoS）中检索了有关知识图谱的95篇论文。对这些论文进行领域的叠加分析，结果如图0.3。可以发现，科学知识图谱涉及的领域中，来自信息科学、计算机科学以及应用数学领域的学者往往研究的是基础性的理论，如科学知识图谱的数学算法和图谱可视化的设计。来自科学计量学和科学学领域的学者通常具有文科背景，主要对知识图谱的哲学原理和表达含义进行深层次的解读。当然，具有信息科学和计算机科学背景的学者，在科学计量学和科学知识图谱领域就显得更有优势，如德雷塞尔大学的陈超美教授、印第安纳大学的博纳教授以及莱顿大学的尼斯·杨·凡·艾克（Nees Jan van Eck）和卢多·瓦特曼（Ludo Waltman）研究员。

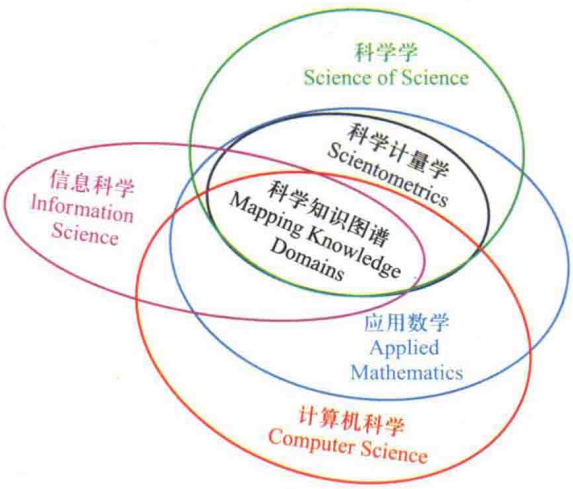


图0.2　科学知识图谱的学科背景[①]

① 刘则渊、陈悦、侯海燕等：《科学知识图谱：方法与应用》，人民出版社2008年版，第12页。

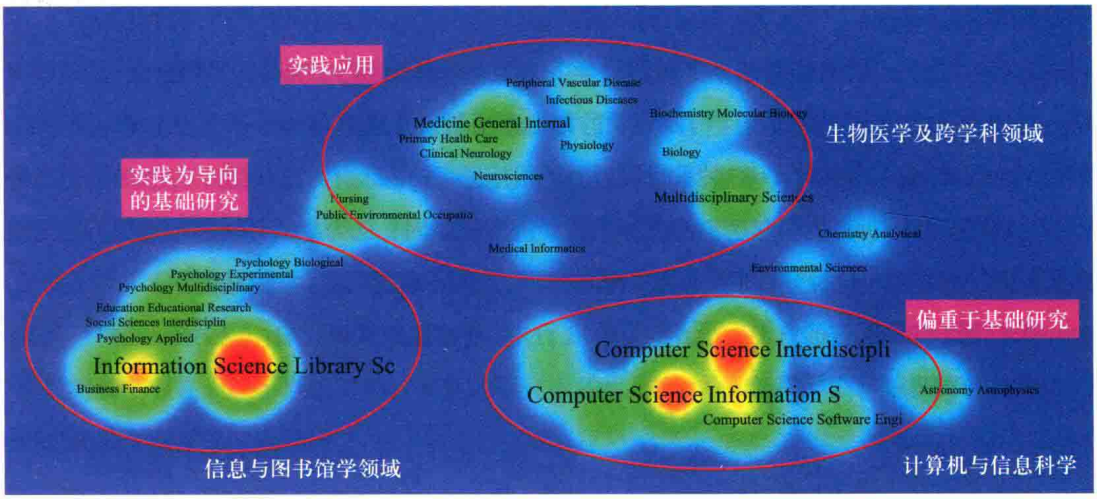


图0.3　科学知识图谱研究的领域分布

刘则渊教授进一步基于卡尔·波普尔（Karl Popper）的“三个世界”理论给出了科学知识图谱的深层次意义的关系图，即知识图谱与视觉思维的关系（如图0.4）。在波普尔的三个世界理论中，传统的看世界的方式是从世界2到世界1并形成世界3。即人类通过视觉来认识客观世界，并由此产生对客观世界认识的知识世界。科学知识图谱的方法则是从世界3出发，通过对世界3的视觉化来认识世界1。这种分析方法为研究人员提供了一种新的认识客观世界的方法和科

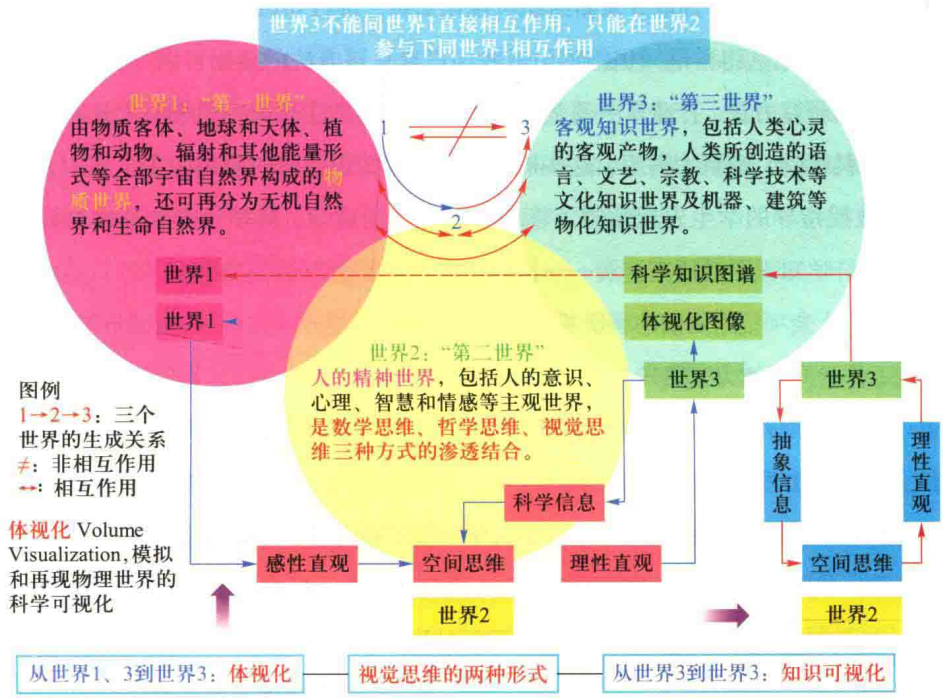


图0.4　科学知识图谱与视觉思维[①]

① 刘则渊：《研究和应用科学知识图谱的意义——知识图谱的科学学源流》，第三期科学知识图谱与科学计量学方法与应用高级讲习班讲义。

学发现的模式。

2015年一个比较有趣的基于文献的科学发现发表在《自然》(*Nature*)[①]上——德国马克斯·普朗克鸟类研究所和新西兰梅西大学的科学家通过对《世界鸟类手册》(*Handbook of the Birds of the World*)中共计5 983种雀形目鸟类的图片进行扫描，并对这些图片的RGB值进行分析和比较评分，分析可能引起鸟羽颜色种间差异的演化原因，认为生活环境和性选择可能是造成这种差异的主因。这种分析正是体现了从“世界3”向“世界1”的认识过程。

关于科学知识图谱的意义，学者们也给出了一些有价值的认识，如：

科学知识图谱改变你看世界的方式。

——陈超美

一图展春秋，一览无余；一图胜万言，一目了然。

——刘则渊

科学知识图谱用于发现科学新知识的逻辑基础与第四范式“数据密集型科学发现”不谋而合。2007年1月11日，图灵奖得主、关系型数据库的鼻祖吉姆·格雷（Jim Gray）在他留给世人的最后一次演讲《科学方法的革命》中，提出将科学研究分为四类范式（Paradigm），依次为实验归纳、模型推演、仿真模拟和数据密集型科学发现。其中，最后的“数据密集型”，也就是现在所称的“科学大数据”[②]。科学知识图谱的绘制和分析，基本的理念正是基于此，即从以往发表的大量科学研究的文献中，提取并重新组织可视化知识，进行知识发现。

自从刘则渊教授及其团队将科学知识图谱引入我国以来，该方法的应用可谓是遍地开花，国内也逐渐形成了一批研究力量，产生了一系列研究成果。国内的主要研究机构有大连理工大学、武汉大学、中国科学院、南京大学等，这些机构多以科学知识图谱的应用为主。其中，大连理工大学刘则渊教授指导的学生先后绘制了管理学、科学计量学、科学学等领域的知识图谱，引起了一系列的科学知识图谱研究涟漪——国内由此产生了上百篇科技论文、近百篇学位论文、数十部著作以及十余项国家资助的科研基金项目。陈超美教授开发的CiteSpace软件也随着科学知识图谱绘制工作的开展被大家熟知。

与国内的机构相比，国外的科学知识图谱研究机构则致力于理论、算法及软件和工具的开发，如德雷塞尔大学陈超美教授开发了CiteSpace，荷兰莱顿大学凡·艾克和瓦特曼两位研究员开发了VOSviewer和CitNetExplorer，美国印第安纳大学博纳教授等开发了SCI2。科学知识图谱领域专门工具的开发，极大地促进了科学知识图谱的广泛应用。

基于对国内外科学知识图谱研究情况的梳理，下面对我国今后科学知识图谱的研究提出几点建议：

---

① Dale J, Dey C J, Delhey K, et al. The effects of life history and sexual selection on male and female plumage colouration. *Nature*, 2015, 527: 367–370.

② 参见http://blog. sciencenet.cn/blog–502444–931155.html，2015–11–12。

（1）从MKD 1.0走向MKD 2.0

科学知识图谱的研究方法和理念引入我国以来，产生了大量的以科学知识图谱实践为导向的研究成果。虽然一部分科学知识图谱在科学性上欠佳，但整体上科学知识图谱研究的质量在不断提升。到目前为止，我国科学知识图谱的应用已经涉及管理学、工学、农学以及医学等领域，且应用范围还在不断扩大。科学知识图谱的应用已经在我国有了广度，但相比国外还缺少深度。为了区分我国过去的MKD研究和将来的MKD研究，这里将上一个阶段的科学知识图谱研究简称为MKD 1.0，下一个阶段简称为MKD 2.0，如图0.5。MKD 1.0到MKD 2.0之间的过渡阶段将长期存在。在不同的时期，其他的研究形式也是存在的，科学共同体科研产出在不同时期的成果会有显著的差异。

图0.5　科学知识图谱研究阶段示意图

MKD 2.0与MKD 1.0的区别在于，2.0时代更加注重以问题为导向的科学知识图谱研究，强调实际科研价值及知识发现，要尽量避免浅显的图谱解答。2.0时代我国需要开发具有知识产权且被广泛使用的科学知识图谱工具，这是科学知识图谱在我国继续发展的保障；我国学者也要能绘制出经得起时间考验且被广泛使用的科学知识图谱。

（2）图谱绘制与解读质量并重

对于初次接触科学知识图谱的学者而言，来自科学知识图谱“炫丽”视觉美感的吸引要大于科学知识图谱自身的科学内涵的价值。这是长期以来我国学者在科学知识图谱研究中存在的普遍问题，当然，不可否认一些学者将科学知识图谱结果结合学科进行了完美解读。这一问题是未来科学知识图谱研究的第一个瓶颈，且该瓶颈需要得到年长学者的重视。因为年轻的研究生和青年教师是科学知识图谱绘制的一线“工人”，大多数对学科理解尚欠全面，对科学知识图谱得到的结果往往解读不够准确、深入。虽然知识图谱绘制仅仅是一项技能，但是缺乏背景知识可能会使得本来具有重大发现的图谱被年轻学者忽略。可见，年长学者在科学知识图谱绘制中的作用是非常重要的。

对于科学知识图谱的初学者，往往可以通过使用某一软件，得到一系列的图谱结果（即包含常见的合作网络、主题网络和引文网络等）。每一张科学知识图谱都需要比较长的时间分析和解读，若将某一学科或主题的科学知识图谱全部放在一篇论文中，解读难免无法深入。这就

要求科学知识图谱应该走向问题导向研究，即“科学问题+知识图谱”的模式。

科学知识图谱解读欠缺还有一个可能的原因，就是国内部分期刊审稿学者缺乏责任心和审稿制度的不完善。笔者深有感触，在国际上发表的关于科学知识图谱的论文往往会得到极其详细的反馈意见。而在国内遇到比较多的情形是简单粗暴的“拒稿”或“录用”，更有甚者，投稿之后石沉大海。试想，若一开始审稿专家对每篇文章都认真阅读，并给出合理的录用理由和拒稿理由，那么知识图谱也不会在一些领域变得“烂大街”。此外，国内科学知识图谱论文，又有多少是科学知识图谱领域的专家审读的？特别是跨学科的科学知识图谱绘制，往往都是某个专业领域内的专家审稿，他们有几人真正研究过科学知识图谱？可见，科学知识图谱的论文要提升质量，在专家把关方面是坚决不能忽视的。

（3）辩证理性地看待科学知识图谱的研究

有些学者发现科学研究中存在使用相同的方法发表一系列类似论文的现象，笔者认为就整个科学研究来看，该现象是普遍存在的。对于还未形成定律的问题进行大量的实践分析并无不可，但若实践应用一直没有推进或者提高，那么这样的实践还是少做为好。笔者认为对于科学知识图谱实践研究不能一棒子打死，因为科学知识图谱的绘制毕竟工作量不小，这种工作应该受到尊重，不能因噎废食。另一方面，进行科学知识图谱绘制的研究人员，要尽量在当前已有的知识图谱成果和经验基础上进行科学知识图谱绘制。笔者发现，一些领域知识图谱的绘制还停留在数年之前的水平。

如何才能让后来的知识图谱学习者站在“巨人”的肩膀上？这需要科学知识图谱领域的专业学者协作起来，分享自己关于科学知识图谱的研究经验、心得、方法以及技巧。笔者认为，在这方面做得好的当属路特·莱兹多夫（Loet Leydesdorff）教授，他不仅将自己的论文分享在自己的主页上，而且对于刚刚开发的科学知识图谱工具或者技巧也予以详细介绍。相比之下，国内学者的学术分享和开放程度还有所欠缺。

0.2 科学知识图谱可以做些什么

从所要分析的科技文献的知识单元组成和目的来看，科学知识图谱回答的基本问题可以总结为5W1H，即When、Where、Who、What、Why和How。

When用来回答科学知识图谱所反映的科学研究的时间信息。例如，从科学知识图谱能了解到某一主题随时间的演变，了解整个网络链接的时间分布以及科学研究的增长和周期律等。

Where主要是提取科技文献中包含的作者地理信息，用来展示某项研究的空间分布情况。科技文献的地理信息可视化分析能够快速地帮助研究人员定位某项研究的重要区域，为科技发展和合作提供决策支持。如果将时间信息和空间地理信息结合，还可以认识科学研究中心的转移情况。

Who主要是提取科技文献中的作者信息。作者信息包含在施引文献及其参考文献中，并通过不同的方法来进行分析。如施引文献作者的合作分析、耦合分析，参考文献作者的共被引分析等。此外，作者的频次分布的研究，可以了解作者产出的统计学分布特征。

What主要对科技文献主题进行挖掘和分析，可以探究某一领域有哪些研究主题，当前的研究前沿主题是什么以及各个主题的演进趋势及关系是什么等问题。

Why主要通过知识图谱得到的结果并结合专业发展背景，解释为什么专业领域是这样的并预测未来将会怎样发展。另外，还可以从网络数学模型上分析，以探究科学知识图谱形成的机制和原因。

How主要通过知识图谱得到的结论，分析指导我们科学发展的实际决策。

上面给出的科学知识图谱的功能仅仅是从被分析的科技文献所包含的内容及目的出发。在具体实践中，科学知识图谱的研究还被拓展、引申到研究前沿、研究热点、研究基础、学科发展、科学结构等方面的研究。因此，关于科学知识图谱到底能回答哪些问题、具体能做些什么，除了从大量的文献中来学习外，还可以结合科学知识图谱分析的原理，给出科学知识图谱的潜在分析价值。

0.3 科学知识图谱分析的基本步骤

科学知识图谱分析的基本步骤，如图0.6。对于以实践应用为主的研究者而言，其中的多个步骤在分析中会由软件代劳。

图0.6 科学知识图谱分析的基本步骤

（1）确定研究目的

研究者明确自己的研究目的，并根据研究目的制订可行的科学知识图谱研究计划，筛选将要使用的数据库。

（2）数据检索

在确定研究目的后，就需要进行数据检索。数据检索是整个分析的关键环节。在具体的实践过程中，我们发现很多科学知识图谱的研究者，由于数据获取存在问题，直接导致了最后结果的错误呈现和错误解读。

常见的可以用于进行科学知识图谱分析的英文科技文献数据库有WoS、Scopus以及德温特专利数据（Derwent Innovations Index）等，中文数据库有中国知网（CNKI）等。从数据格式上来看，主要有文本文档、Endnote以及Bibtex等。

（3）数据预处理

数据的预处理包含对原始数据进行的除重、消歧、格式转换以及排序等处理。现有的技术已经能很好地进行施引文献的消歧处理。相对而言，作者和参考文献的消歧还存在比较大的问题。

（4）提取知识单元

科技文献的题录数据由不同的知识单元构成，包含标题、作者、机构（地址）、摘要、关键词以及参考文献等。WoS的核心合集数据格式可以参见附录A，在实际研究中，就是从这些已有的知识单元中提取知识单元的共现矩阵，并将其可视化。

（5）提取共现矩阵

在科技文献知识单元共现分析中，常用的步骤是提取“知识单元—文献”矩阵，然后使用矩阵的乘法来获取相应的共现矩阵。例如，为了获得作者合作矩阵，首先可以从科技文献中获取“作者—文档”的隶属矩阵，在该矩阵中若作者属于某个文档，则对应矩阵的元素为1，否则为0。要得到“作者—作者”合作矩阵，只要使用该矩阵与矩阵的转置相乘即可。

（6）共现矩阵标准化

矩阵的标准化有多种方法，常见的软件中嵌套的多为基于集合论的矩阵标准化方法，例如Cosine、Jaccard和Dice方法。这三种方法可以统一用下式表示：

$$S_{ij}\left(c_{ij},c_i,c_j;p\right)=\frac{2^{1/p}c_{ij}}{\left(c_i^p+c_j^p\right)^{1/p}}。$$

当该式中p=0时，那么得到的公式就为Cosine的标准化公式；当p=1时，那么得到的标准化公式为Dice，与Jaccard方法的关系为下式[①]，

① van Eck N J, Waltman L. How to normalize co-occurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, 2009, 60(8): 1635−1651.