文章编号:1003 - 2053(2015)02 - 0242 - 12

CiteSpace 知识图谱的方法论功能

陈 悦[1],陈超美[2],刘则渊[1],胡志刚[1],王贤文[1]
(1. 大连理工大学(中国)—德雷塞尔大学(美国)知识可视化与科学发现联合研究所,WISE 实验室,大连 116085;
2. 德雷塞尔大学计算与信息学院,美国)

摘 要:科学知识图谱的概念和 CiteSpace 工具自引入国内学术界,就迅速得到了大量关注,相关文献犹如雨后春笋般见诸国内情报学、科学学和管理学等各种期刊。但我们通过阅读国内 500 多篇应用 CiteSpace 工具的论文,发现存在知识可视化工具"滥用"和"误用"的现象,其缘由在于使用者对该工具的方法论功能认识不足。为此,本文从四个方面阐释 CiteSpace 知识图谱的方法论功能:从 CiteSpace 工具的设计理念入手阐发其改变看世界方式的核心功能;从 CiteSpace 的理论基础阐述其对研究领域解释与预见上的理论功能;从 CiteSpace 使用流程阐明其方法论功能的实现;从 CiteSpace 的新近技术介绍其应用功能的扩展。我们期望 CiteSpace 知识图谱在探测学科前沿、选择科研方向、开展知识管理和辅助科技决策诸方面能够更好地发挥方法论的功能。

关键词:科学知识图谱;方法论;CiteSpace
中图分类号:G301 文献标识码:A
DOI:10.16192/j.cnki.1003-2053.2015.02.009

自 2005 年我们率先在中国命名和引入科学知识图谱(mapping knowledge domains)[1]以来,科学知识图谱或知识图谱作为科学计量学的新方法和新领域在我国勃然兴起并获得长足的发展。科学知识图谱是以知识域(knowledge domain)为对象,显示科学知识的发展进程与结构关系的一种图像。它具有"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、演化或衍生等诸多隐含的复杂关系,而这些复杂的知识关系正孕育着新的知识的产生。

科学知识图谱的概念源于 2003 年美国国家科学院组织的一次研讨会,随着信息可视化的发展,绘制科学知识图谱的各种工具亦纷至沓来[2]。其中,CiteSpace 知识可视化软件如异军突起,成为目前最为流行的知识图谱绘制工具之一,阐释其基本原理的 CiteSpace Ⅱ: Detecting and visualizing emerging trends and transient patterns in scientific literature 一文[3],迄今(截至 2014 年 8 月 8 日)在谷歌学术搜索(Google Scholar,GS)上已被引 855 次,其中文版本[4]也被引 196 次(GS)。由于这种多元、分时、动态的引文分析可视化技术所绘制的 CiteSpace 知识图谱,能够将一个知识领域来龙去脉的演进历程集中展现在一幅引文网络图谱上,并把图谱上作为知识基础的引文节点文献和共引聚类所表征的研究前沿自动标识出来,因此我们将 CiteSpace 知识图谱的这两大基本特征概括为 "一图谱春秋,一览无余;一图胜万言,一目了然"[5]。

正是 CiteSpace 知识图谱的鲜明特征而导致 CiteSpace 迅速得到广泛的应用,随之出现了一批关于应用 CiteSpace 及其知识图谱的文献综述。国内较早开始应用 CiteSpace 的侯剑华和胡志刚[6]分析了收录在 WoS 和 CNKI 中应用 CiteSpace 的论文的学科分布和使用功能。中国科学技术信息研究所的

收稿日期:2014 - 04 - 27;修回日期:2014 - 10 - 21
基金项目:大连市科技计划软科学研究项目(2012D12ZC180)
作者简介:陈 悦(1975 -),女,辽宁大连人,副教授、博士生导师,研究方向为科学学、科学计量学。E-mail:chenyuedlut@163.com。
陈超美(1960 -),男,北京人,终身教授,博士,研究方向为信息可视化,知识图谱与科学计量学。
刘则渊(1940 -),男,土家族,湖北恩施人,教授、博士生导师,研究方向为科学学理论、科学计量学与科技管理。
胡志刚(1984 -),男,山东济宁人,博士生,研究方向为科学计量学与科技管理。
王贤文(1982 -),男,博士,湖南双峰人,副教授,研究方向为科学计量学与科技管理。

胡泽文等[7]以在综述了国内知识图谱应用现状之后惊呼“CiteSpace 及知识图谱绘制方法引入中国后,国内学者对该主题的研究呈井喷之势。”北京大学的赵丹群[8]在对国内基于 CiteSpace 的知识图谱应用现状调研的基础上,从领域文献的查找、突变词语的探测、时区分割与相关参数的阈值设置和图谱解读四个方面较为深入地探讨了应用 CiteSpace 中存在的重要问题。值得关注的是,我国不仅产生了一批以 CiteSpace 为知识图谱绘制工具的硕士博士学位论文,而且在学位论文中能够剖析使用 CiteSpace 过程中存在的诸多问题,如北京大学王钦炜[9]在其学位论文中提出国内研究者普遍缺乏对 CiteSpace 软件功能及使用方法的深入了解,由此造成了一系列科学知识图谱绘制中的诸多问题: 图谱绘制缺乏规范,图谱质量参差不齐,图谱解读不当,单张图谱信息量过载而导致图谱可视化直观程度下降等。

人们对待新鲜事物的态度往往是经历观望、追随、狂热、冷静、再回归理智的过程,我国学术界对于 CiteSpace 和知识图谱的态度也显示出了这种趋势。图 1 为以“CiteSpace OR 科学知识图谱”为检索式在 CNKI 中“全文检索”(2005.01 – 2013.12) 所检索到的 1352 篇学术论文年度分布。2005 年为我国关于科学知识图谱文献的起始年,《科学学研究》发表推出了国内第一篇科学知识图谱论文《悄然兴起的科学知识图谱》(被引 229 次,检索时间: 2014 年 8 月 13 日),同期发表的还有刘林青的《作品共被引分析与科学地图的绘制》,用多维尺度分析方法来绘制“科学地图”,也就是我们所说的科学知识图谱。其后,论文数在经历 2009 年到 2012 年的急剧增长后,2013 年增长趋于平缓。从 1352 篇论文中抽取出的 555 篇应用 CiteSpace 的论文,其应用目的和研究领域的分布非常广泛,但主要集中在管理学领域,其中图书情报与档案管理占 42.12%,管理科学与工程、公共管理和工商管理共占 22.72%,教育学、社会学、体育学共占 17.41%,其余大都为人文社科领域,自然科学领域仅基础医学和生物学只占 4.7%。值得注意的是,通过基于 CiteSpace 的专利文献知识图谱分析,它正在工程技术领域迅速扩散与应用。

考察和分析这 555 篇应用 CiteSpace 的论文,我们深刻感受到国内学术界对科学研究新方法和新工具的渴求,CiteSpace 凭借其使用操作简单、适用源于多种数据库格式的数据、可以绘制多种图谱、可视化效果好、提供信息量大和自动标识易于图谱解读等强大功能优势吸引了各个专业学科的研究人员,从目前大多数使用 CiteSpace 的目的主要在于探测学科知识领域发展及其研究热点、前沿和趋势,大体符合开发 CiteSpace 的初衷。但从现有应用研究的后果来看,仍然存在一些问题:

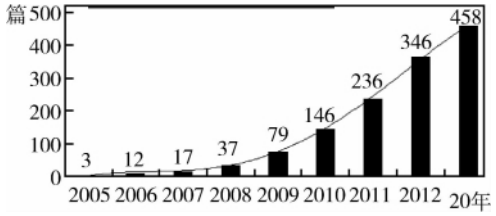


图 1　CNKI 中关于 CiteSpace 和科学知识图谱的学术论文年度分布: 2005 – 2013

(1) 知识领域数据下载策略不当,达不到数据集的完整性和准确性。

(2) 对 CiteSpace 中功能使用的选择与所要解决的问题不匹配。

(3) 图谱不美观。主要表现在结构过于拥挤、节点和标签的大小不匹配。

(4) 图谱信息缺失。主要表现在 CiteSpace 使用版本、节点和连线数量不清、阈值选择不明了。

(5) 图谱解读偏颇。大部分图谱对高频节点都进行了解读,一半左右的文章会对聚类解读,接近一半的文章对高中心性节点进行了解读,42% 的文章含有图例说明,时间趋势和 burst 的应用较少。

(6) 图谱绘制效果缺乏评估。我们在 555 篇论文中仅找到 1 篇论文利用聚类模块性指数 Q 值和聚类轮廓性指数 S 值来评估图谱聚类效果。

(7) CiteSpace 提供了很多深入分析的功能和解读信息,但目前对其应用还都处于较为简单的层次。

这些问题导致知识可视化工具的“滥用”和“误用”,损害了知识图谱的声誉,甚至威胁到知识图谱的命运。究其根源主要是使用者对 CiteSpace 工具的认识不足,尤其对其方法论功能上的理解还有所欠缺。因为方法论功能并非只是各种方法及其作用的集合,而主要是基于哲学理念和学科理论的观察世界、认识世界与变革世界的方式。正是基于这一点,本文作者作为 CiteSpace 开发者和主要合作者及优先使用者,试图把近几年对话交流所达成的共识,汇集为 CiteSpace 知识图谱的方法论功能,拟分别从如下四个方面加以探讨: 从 CiteSpace 工具的设计理念入手阐发其改变看世界方式的核心功能; 从

CiteSpace 的理论基础阐述其对研究领域解释与预见上的理论功能; 从 CiteSpace 使用流程阐明其方法论功能的实现; 以及从 CiteSpace 的新近技术介绍其应用功能的扩展。其中若干关键内容系在开发和改进 CiteSpace 工具的背后所坚守的宏观哲学观念和相关学科理论,在此首次坦诚地较为完整地披露出来,与国内学术同行分享,以期 CiteSpace 知识图谱保持旺盛的生命力,在探测学科前沿、选择科研方向、开展知识管理和辅助科技决策诸方面能够更好地发挥方法论的功能。

1 CiteSpace 的核心功能: 改变看世界的方式

CiteSpace 是应用 Java 语言开发的一款信息可视化软件,它主要基于共引分析理论(co – ciation) 和寻径网络算法(pathFinder) 等,对特定领域文献(集合) 进行计量,以探寻出学科领域演化的关键路径及其知识拐点,并通过一系列可视化图谱的绘制来形成对学科演化潜在动力机制的分析和学科发展前沿的探测。不仅如此,作为 CiteSpace 的开发者,陈超美特别强调[11]: 更重要是在于让使用者通过对知识图谱的绘制、生成和解读,看到知识图谱将会如何改变看世界的方式; 并明确袒露 "CiteSpace 的背后需要有对库恩或类似的宏观哲学思想体系的了解,才能明白 CiteSpace 到底在帮用户找什么。"[12] 这里,我们引入著名科学哲学家卡尔·波普尔关于三个世界的宏观哲学理论[13],来说明 CiteSpace 的设计理念,阐释其如何改变看世界方式的核心功能(图 2) 。

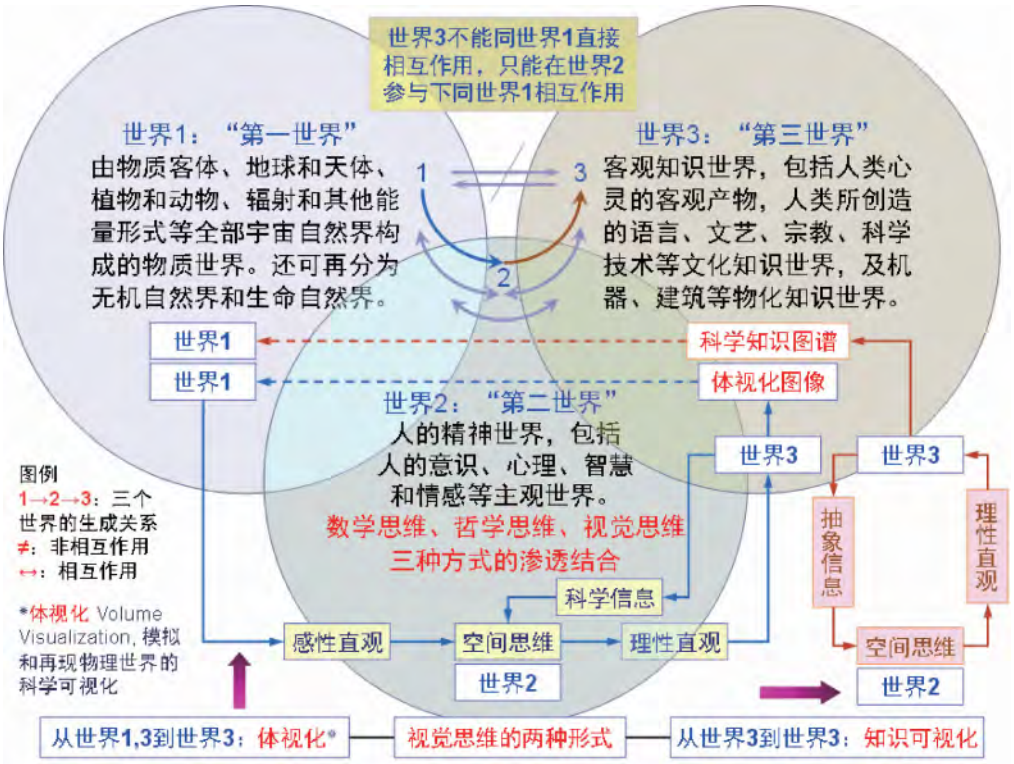


图 2 CiteSpace 的核心功能: 基于三个世界理论的看世界方式

在波普尔的三个世界理论中,存在着物理世界(世界 1) 、精神世界(世界 2) 和客观知识世界(世界 3) 。与划分为客观世界和主观世界的经典哲学认识论不同,波普尔的独到见解在于世界 1 和世界 2 相互作用所形成的世界 3,是人类创造的知识世界,一旦形成便具有客观性; 其结构非常复杂,可分为两部分: 文化知识世界和物化知识世界。从科学计量学之父普赖斯[13] 的科学学理论(1963 年) 看,世界 3 可分为两个层次: 由世界 2 认识自然界所获得的科学知识,属于 "一阶科学或一阶主题"(first – order–subject) ,而科学的科学,包括对科学的认识和计量,属于 "二阶科学或二阶主题"(second – order subject) 。后来,1999 年瑞典学者伍特斯[14] 提出科学表征的概念,将其分为三个层次,意味着世界 3 也可

分为三个层次: 一阶表征(first order representation) 为科学文献; 二阶表征(second order representation) 为引文分析; 而他研究的引证文化(citation culture) 则属于三阶表征(third order representation) 。这就是说,世界 3 存在抽象程度不同的多阶科学。但抽象程度更高的高阶科学却能够更深刻地反映世界 1 中客体的本质。这就是所谓“思维中的具体”。不过这种思维中的具体毕竟远离直观的物理世界,人们难以理解,于是直观形象的可视化技术应运而生。

目前可视化技术有两种形式: 图 2 左侧表示模拟和再现物理世界的科学可视化,亦称体视化(volume visualization) 。它与一般科学研究的看世界方式一样,是通过人的视觉,即世界 2 来认识世界 1,形成世界 3,虽然属于“一阶科学”的范畴,但不同之处是以体视化的图像表征世界 3,直观再现世界 1。图 2 右侧为知识可视化,是世界 2 从世界 3 中获取抽象信息与知识,通过绘制知识图谱来认识世界 1。而制作知识图谱是复杂的认识与思维过程,需要“觉悟”、“感悟”,主要是“视觉顿悟”、“视觉思维”。CiteSpace 的设计理念正是由世界 2 以视觉思维方式,分析和加工世界 3 中“一阶科学”的一个领域文献,通过绘制知识可视化图谱来透视世界 1 的一个现实领域,从而“改变看世界的方式”。

在这种以科学知识图谱的方式来认识世界中,视觉思维、数学思维和哲学思维三种思维方式得以统一。首先,绘制出的图谱必须具有映射性、美观性和易读性,而品质优良的图谱是通过各种算法才得以实现,图谱的整体设计、算法选择及解读依赖的是哲学思维。法国数学家波尔达斯·德莫林斯曾说过,“没有哲学,固然难以得知数学的深度,然而没有数学,也同样无法探知哲学的深度,两者互相依存。还应特别指出,如果既无哲学也无数学,则就不能认识任何事物”[15]。因而,不从科学哲学的角度去把握 CiteSpace,不理解其中各种算法的选择,就难以绘制出令人满意的图谱,更难以去科学地解读图谱。

总之,我们以基于三个世界理论的看世界方式,诠释了 CiteSpace 的核心功能: 借助一个知识领域演进的可视化图谱,以更高抽象程度的“二阶科学”范畴和更为生动直观的形象化图像,从整体上更加深刻地反映和逼近物理世界一个具体领域的科学发展规律,不仅有助于解释现有科学发现,而且有利于建立在世界 3 基础上的新发现,即基于文献的科学发现。基于 CiteSpace 的可解释性与可计算性科学发现理论[16],就是这方面的一个范例。

2　CiteSpace 的理论功能: 对研究领域的解释与预见

包括 CiteSpace 的所有信息可视化工具都是旨在改变人类看世界的方式,在科学图谱中,“看”包括“搜索”和“解读”两个步骤。如何“搜索”和“解读”才更有效率和效用呢? 人们意料之中的信息实际上远不如意料之外的信息更有价值,因为后者意味着变化,很可能预示着新事物的出现。因而,寻找可视化图谱中那些不同寻常的点并分析这些不同寻常点之间的关联是非常重要的。针对于科学知识图谱的 CiteSpace 工具的设计主要基于库恩的科学发展模式理论、普赖斯的科学前沿理论、社会网络分析的结构洞理论、科学传播的信息觅食理论和知识单元离散与重组理论。这些理论基础的意义在于强化图谱的可解读性、解读的合理性和正确性,通过图谱解读,实现理论两大功能,即领域现状的解释功能与领域未来前景的预见功能。

库恩的科学发展模式理论。库恩把科学发展看成科学革命的历史过程。科学在未形成统一范式之前处于前科学时期; 范式形成之后,进入常规科学时期,人们在科学共同体中按范式解题,是范式积累期; 发展一定阶段,出现反常和危机,人们寻求新的范式取代旧范式,导致科学革命的发生; 之后,迈进新范式下的新的常规科学期。因此,科学发展本质上是常规科学与科学革命、积累范式与变革范式的交替运动过程。这个科学发展模式可以更深刻地阐释 CiteSpace 知识图谱上一个学科领域引文聚类的形成、积累、扩散、转换进程,揭示一个知识领域研究前沿的突现与演变进程。库恩理论关于发现的涌现、经典名著是科学的转折点等观点,仿佛预见到 CiteSpace 共引网络图谱中关键节点论著的被引突现性和转折点特征。

普赖斯的科学前沿理论。普赖斯受贝尔纳关于“科学发展总的模式与其说像树,更像网”思想的启发,在加菲尔德发明的科学引文索引(SCI) 基础上,预言“论文会因为引证关系而形成网络,人们可以借助于图论和矩阵的方法来加以研究。……论文一定会聚集成团,而形成几乎绘制成地图的(显示出拥有高地和不可逾越的沼泽地) ‘陆地’和‘国

家'。"[13] 紧接着在著名的《科学论文的网络》(1965)一文中,把它变成了现实,由此形成普赖斯的"参考文献的模式标志科学研究前沿的本质"的前沿理论。这个前沿理论是贝尔纳的创意、加菲尔德的发明和普赖斯的破解三者的结晶[18]。CiteSpace 在此基础上,创造性地将引证分析(历时性)和共引分析(结构性)综合起来,创建了从"知识基础"映射到"研究前沿"的理论模型,即"如果我们把研究前沿定义为一个研究领域的发展状况(如研究思路),那么研究前沿的引文就形成了相应的知识基础。一个研究领域可以被概念化成一个从研究前沿 $\Psi(t)$ 到知识基础 $\Omega(t)$ 的时间映射 $\Phi(t)$,即 $\Phi(t):\Psi(t)\rightarrow\Omega(t)$"[3]。

社会网络分析及结构洞理论。在社会网络分析理论的形成中,英国社会学家格兰诺维特(Mark Granovetter)提出社会网络"弱连接优势"的重要观点,认为信息在强关系的群体中高速传播,每个人知道的,其他人也多半会知道,新观点和新信息一定来自于与其他不同群体中的个体间的弱关系[19]。博特[20]在此基础上提出结构洞理论。2012 年 5 月基于 CiteSpace 的再生医学领域综述[21],正是利用结构洞理论分析和把握了其知识图谱上关于"诱导多能干细胞(iPSC)"的前沿聚类中,日本生物学家山中伸弥(Shinya Yamanaka)首创"iPSCs"的高被引、高突现性论文的关键基础作用(参见图 5-c 左下聚类 7),预言该领域这一研究前沿将会摘取诺贝尔奖。果然,山中伸弥和英国科学家格登(John Gurdon)因在此方面的贡献而获得 2012 年度此项殊荣。处于结构洞未知的个体透过信息过滤获得更多竞争优势与创新能力。CiteSpace 基于此理论开发出知识网络中关键节点及关键位置的发现技术,即发现知识转折点(turning point)[22]。

信息觅食理论。该理论主要用来解释和模拟人们在网络环境中的信息搜寻行为,通过模型的简历,模拟用户的信息搜寻过程,并对获取信息的效率进行计算,以其最小搜索成本获取最大利益。CiteSpace 将该理论融入科学发现中,揭示科学网络中的结构与时间属性,从发现知识转折点及其连接的角度,开发了一套探寻知识传播(或知识演变)路径的独特方法和技术。

知识单元的离散与重组理论。我国科学计量学家赵红州首先提出"任何一种科学创造过程,都是先把结晶的知识单元游离出来,然后再在全新的思维势场上重新结晶的过程。这种过程不是简单的重复,而是在重组中产生全新的知识系统,全新的知识单元。"[23] 在此基础上,刘则渊等[24]提出知识单元(knowledge unit)就是表征知识领域文献内容或信息内容的概念及陈述、语词及词组、术语及定律等可计量的基本单位。它是知识计量学的核心概念和基本计量单位。在一定条件下,某个关键的知识单元可能扮演"知识基因"(knowledge gene)的角色,决定着特定领域知识的进化与突变。因而,基于知识单元的特定知识领域所构成的复杂自组织知识系统,就能够在 CiteSpace 知识图谱上展示知识的产生、传播和应用,知识的基础、中介和前沿,知识的结构、演化和重组,知识的涌现、断层和变革,等等。因此,可以用关于凝聚游离的知识单元阐释科学发现的宏观和微观机制,这跟上述以网络结构(结构洞)和信息变化(概念假设突变)为基础的科学发现机制,可谓异曲同工。

3 CiteSpace 的应用流程:方法论功能的实现

一般说来,CiteSpace 知识图谱的合格满意标准主要是:数据完整、程序正确、图谱美观、解读合理,并在图谱制作中能够贯穿和体现 CiteSpace 的核心功能与和理论功能。这两方面是 CiteSpace 知识图谱方法论功能中的关键与基础。包括这两方面在内的方法论功能要得以实现,必须通过 CiteSpace 的一系列应用流程来保证。为此,这里汇集了 CiteSpace 当前版本使用中,能够达到知识图谱合格满意标准的主要流程,包括软件安装、数据采集、数据处理、参数功能选择、可视化和解读(图 3)。

在安装和启动 CiteSpace 软件之前,首先应确保电脑装有相匹配的 Java Runtime (JRE),如果电脑系统是 32 位的,需安装 Windows x86 的 JRE,如电脑系统是 64 位的,需安装 Windows x64 的 JRE。当前版本(CiteSpace3.8.R3)最优化的是用于装有 Java7 的 64 位 Windows 系统。当 CiteSpace 运行速度非常慢时,除了考虑数据量的原因外,也应该考虑计算机的系统配置。

CiteSpace 软件对数据格式的要求是以 Web of Science 数据库的文本数据格式为标准,并随着 ISI 数据库中数据格式的变化而不断更新。该软件可直接导入 Web of Science 和 arXiv 数据库中的数据,直接进行可视化分析,并对于来源于 CNKI、CSSCI,

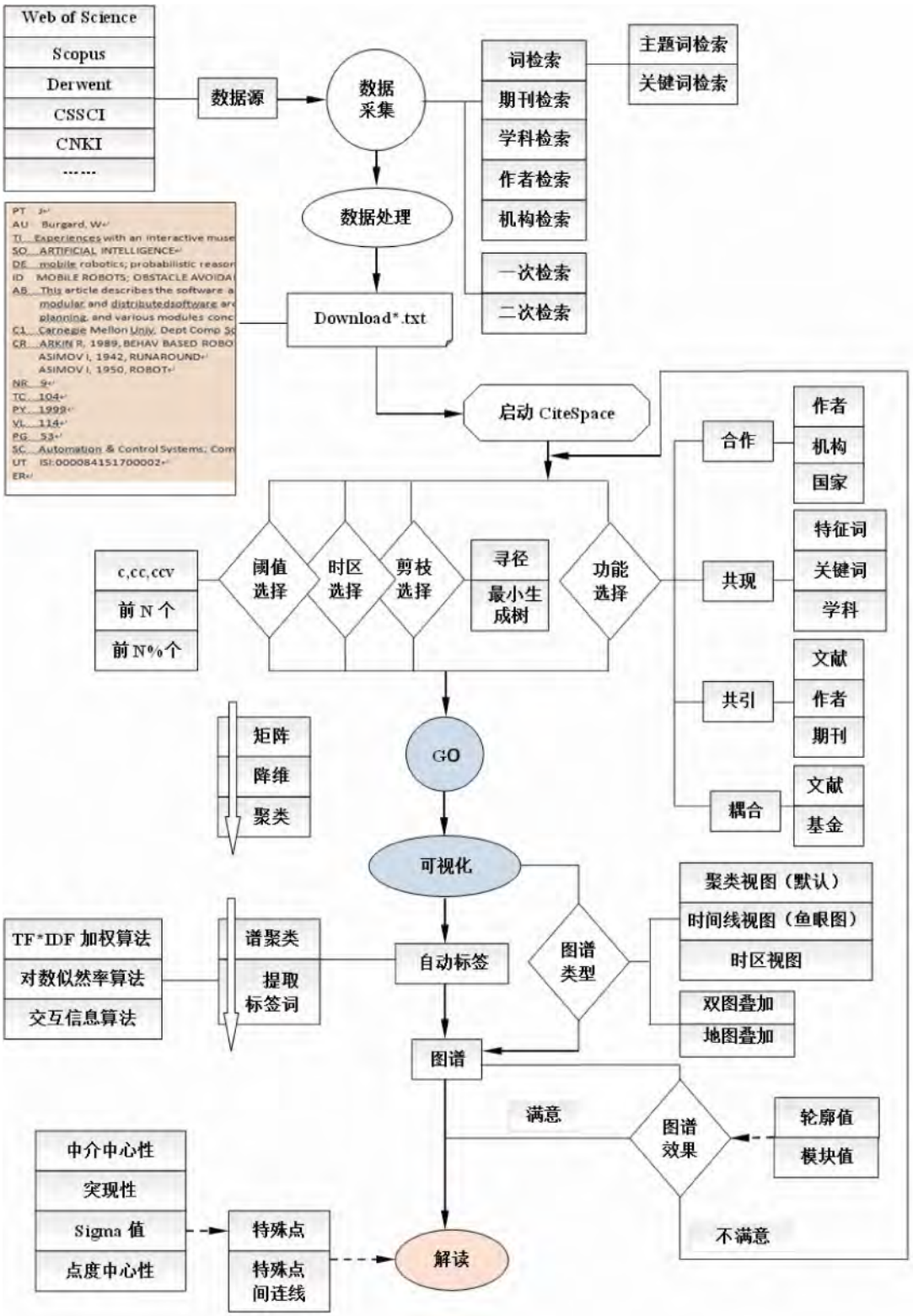


图 3　CiteSpace 的应用流程

Derwent，NSF，SCOPUS，SDSS 和 Project DX 的数据提供了数据格式转换器。CiteSpace 更适用于研究某个主题的演进，所以有针对性的主题检索效果相对更好，由于该工具基于数据的聚类而形成可视化图谱，因而数据量应达到一定的规模，如果一次检索的数据量较少，可以以此为“种子”，进行引文的二次

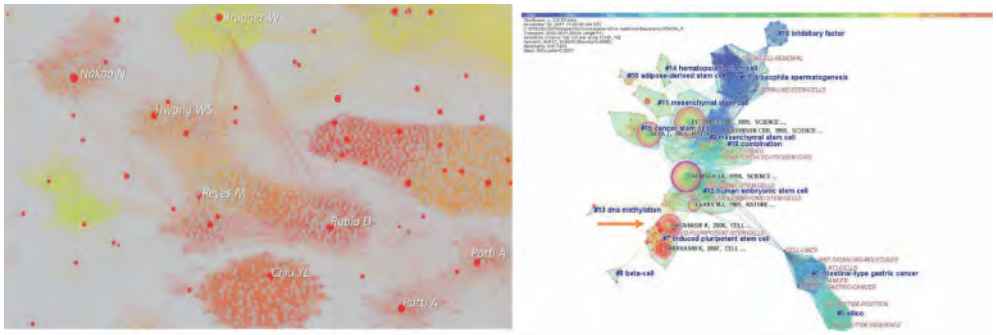
检索，数据将会更完整，这可以在一定程度上提高可视化效果，“再生医学”前沿研究[21]就采用了这种数据检索方式。任何知识图谱绘制的科学性都根源于数据基础，即如何精准全面地检索到所要研究主题的全部文献是关键的问题，作者应该重视数据检索方式，并在研究论文中有明确表述。除了利用已有数据库的数据之外，我们也应该试图根据所要研究的问题自己搜集挖掘数据，武夷山说“对于从事科学计量学研究的，如果不肯花力气去搜集、挖掘待分析的数据，那就趁早离开得了”[25]。CiteSpace是一个开源软件，它有强大的数据处理功能，我们可以在数据的搜集和检索方面做更多的努力。

数据准备好之后，进入CiteSpace运行阶段，该阶段包括一系列的选择，即时区选择、阈值选择、剪枝选择和功能选择。时区选择是CiteSpace工具的一大特色，但当研究内容并不在于反映“演化”时，就可以灵活地将数据划为一个时区。阈值选择提供了多种数据筛选的策略。

数据准备好之后，进入CiteSpace运行阶段，该阶段包括一系列的选择，即时区选择、阈值选择、剪枝选择和功能选择。时区选择是CiteSpace工具的一大特色，但当研究内容并不在于反映“演化”时，就可以灵活地将数据划为一个时区。阈值选择提供了多种数据筛选的策略。首推最简单的Top N选择，即在每个时区中选择前N个高频出现的节点；次推Top N%选择，即在每个时区中选择前N%个高频出现的节点；第三种比较复杂，通过前、中、后三个时间段的(c，cc，ccv)，即(被引或出现的频次，共被引或共现频次，共被引率或共现率)的设置来筛选数据的方式，具体运行过程中通过线性插值的方法对各个时间段进行阈值控制。(c，cc，ccv)的前两项是绝对值控制，实现对点的控制，ccv是相对值控制，实现对线的控制，经验值为15或20，这意味着我们对出现频率较高的两点的共现频率的要求也相应提高；第四种选择是要与上述三种选择策略配合使用，选择出现频率在某个区间的文献(或词等)，这使得我们可以根据研究的具体内容，方便地删除掉可能无太大意义的高频文献或低频文献。在CiteSpace运行过程中，后台的数据处理状况都能够显示出来，我们可以根据数据运行状况进行阈值调整。如果可视化初期结果杂乱难以解读，CiteSpace提供了寻径(PathFinder)和最小生成树(Minimum Spanning Tree，MST)两种剪枝方式的选择，Path-Finder的作用是简化网络并突出其重要的结构特征，它的优点是具有完备性(唯一解)，MST的优点是运算简捷，能很快出结果。CiteSpace提供了11种功能选择，针对于施引文献的合作图谱(作者合作、国家合作和机构合作)和共现图谱(特征词、关键词、学科类别)，以及针对于被引文献的共引图谱(文献共被引、作者共被引和期刊共被引)。这些图谱都可以用来揭示科学结构的发展现状乃至变化情况，并进而用于前沿分析、领域分析、科研评价等，但针对于具体的研究问题，应根据不同图谱的绘制原理来进行选择。如使用最频繁的是文献共被引图谱，可以帮助人们通过图谱中的关键节点、聚类及色彩来分析某个研究主题的演变；合作图谱可以发现某个研究领域学者、国家或研究机构之间的社会关系，为评价科研人员、国家或机构的学术影响力提供一个新的视角，有利于我们发现那些值得关注的科研人员、国家或机构；共词(特征词或关键词)图谱更有利于人们分析研究热点及热点的演变，尤其配合突现词(burst term)功能的使用；学科类别贡献图谱往往用来分析学科知识结构及其演变；作者共被引图谱可以用于分析某个领域内的科学共同体及其演变；期刊共被引可用于研究领域的学科基础及其演变的分析。完成这一系列选择，按下运行按钮，CiteSpace将在后台进行创建矩阵、降维和聚类的过程，数据筛选和运行情况会显示在运行窗口的左侧。随后进入可视化阶段。

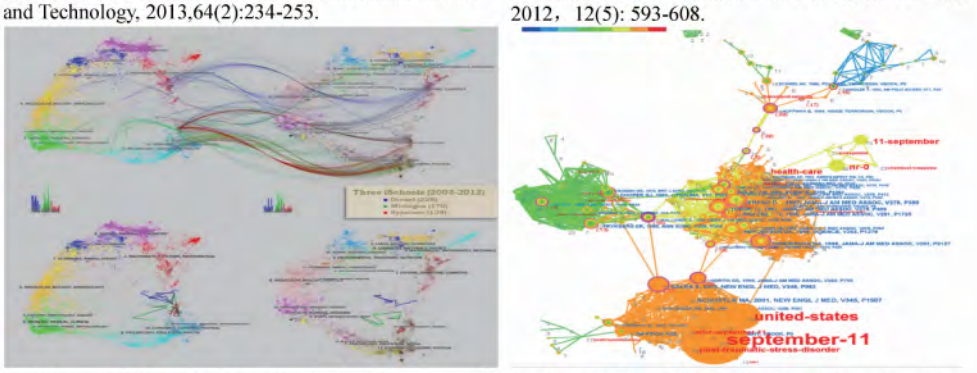
CiteSpace提供了三种可视化方式的选择，其中默认的是聚类视图(cluster)，它侧重于体现聚类间的结构特征，突出关键节点及重要连接，时间线视图(Timeline)侧重于勾画聚类之间的关系和某个聚类中文献的历史跨度，时区视图(timezone)是另一种侧重于从时间维度上来表示知识演进的视图，它可以清晰地展示出文献的更新和相互影响。在聚类视图的基础上我们还可以选择双图叠加以寻求两个图谱之间的关联，或是以Googlemap为基础图，绘制一幅空间知识图谱。CiteSpace依据谱聚类算法提供了自动聚类的功能，并提供了从聚类施引文献中提取聚类主题词的三种算法，默认的自动标签词是依据TF* IDF加权算法而给出的。绘制图谱的要求之一是要美观并易解读。“美观”就是指看上去舒服，对于一副知识图谱而言，如果显示出结构过于拥挤、节点大小和标签大小不协调、色彩混乱，则称不上“美观”，但若结构布局清晰、节点大小和标签大

小适度、色彩层次化、干净利索的图谱会让人舒服，甚至有艺术的享受，即“美观”(图 4)。



a. Chen, C., Hu, Z., Milbank, J., Schultz, T. A visual analytic study of retracted articles in scientific literature. Journal of the American Society for Information Science and Technology, 2013,64(2):234-253.

b. Chen, C., Hu, Z., Liu, S., Tseng, H.Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. Expert Opinions on Biological Therapy, 2012, 12(5): 593-608.



c. Chaomei Chen & Loet Leydesdorff,Patterns of Connections and Movements in Dual-Map Overlays: A New Method of Publication Portfolio Analysis,Journal of the Association for Information Science and Technology 2014,65(2) : 334-351.

d. Chen, C. Searching for intellectual turning points: Progressive Knowledge Domain Visualization. Proceedings of the National Academy of Sciences of the United States of America(PNAS),2004, 101(Suppl. 1); 5303-5310.

图 4　美观的知识图谱示例

CiteSpace 依据网络结构和聚类的清晰度，提供了模块值(Q 值)和平均轮廓值(S 值)两个指标，它可以作为我们评判图谱绘制效果的一个依据。一般而言，Q 值一般在 [0,1) 区间内，$Q>0.3$ 就意味着划分出来的社团结构是显著的，当 S 值在 0.7 时，聚类是高效率令人信服的，若在 0.5 以上，聚类一般认为是合理的。知识图谱的绘制是需要选取不同的阈值多次绘制，依据 Q 值和 S 值选取较理想的图谱作为最终的结果。另外，值得一提的是，为了便于读者对图谱的认识和理解，我们应该尽可能保留软件生成图谱的坐上方信息栏，其中提供了各种阈值设置、节点数、连线数、网络密度、轮廓值及模块值等数值。

绘制知识图谱的目的是更好地理解科学发展的状态和机制，因而解读是关键。图谱解读是一项兼具科学性和建构性的工作，建构性必然会带来图谱解读的因人而异，无法强求一致，而科学性则要求图谱解读的规范和严谨，需遵循一定的规则和程序。专家解读固然能提高图谱解读的科学性，但随着科学的交叉、融合、纵深的快速发展，新兴研究领域和主题不断涌现，所谓的专家也未必能对科学的局部与整体把握得十分准确，实际上从某种角度而言，科学知识图谱工具的使用有助于改善人们的这种认识不足。关于 CiteSpace 的三篇重要文献[3][22][26]，除了对形成的文献结构进行分析外，都经过了专家的认证解读。这从一定程度上可以证明 CiteSpace 是可以用来反映科学发展的客观情况的。CiteSpace 是通过多种阈值选择而形成的一种独特的多个文献共被引网络组合而成的知识网络，并提供了一些自动生成的信息，可以利用这些信息从网络的整体结构、形成的聚类、聚类之间的关系(包括结构的关系和时间的关系)来入手，解读过程中应参照各种自动生成的指标信息(右键弹出菜单提供很多功能)。另外，自动聚类和自动提取出的聚类标签词极大地帮助我们理解网络的内容，在理解网络结构和内容

时，寻找特殊点和连接线是很重要的，这些特殊点占据着知识网络中的一些重要位置，在知识结构演变中扮演着特定的角色，这些特殊点的寻找可以依据中介中心性（betweenness centrality）、突现性（burst）、综合考虑中介中心性和突现性的 Sigma 值等来灵活判断。

4 CiteSpace 的功能拓展：从地理图谱到双图叠加

CiteSpace 知识图谱问世之初仅限于展示知识领域研究前沿演进的基本功能，其后技术不断改进，功能不断拓展。鉴于国内大多数应用 CiteSpace 的论文都是使用了该工具较为初级的功能，本文在此推介几种较为高级的功能，以便国内学者能更有效地应用该工具。

（1）基于 Google Map 的知识图谱。CiteSpace 可以按图 3 应用流程引入 Google 地图，自动生成合作网络的地理分布图谱，它可以从空间位置上直观地显示出作者和合作作者之间的关系（图 5）。

（2）CiteSpace 的数据处理功能。CiteSpace 软件内置了 MySQL 数据库，可以导入 WoS 格式的 txt 数据。通过菜单按钮或直接输入 SQL 语句，可以对生成的数据库进行查询和更新，实现对数据的统计、过滤和清洗。CiteSpace 软件 3.7.R7 版本中的内置数据库采用的是 MySQL 数据库技术，要求本地机器可以成功运行 MySQL 数据库，并且需要在在“C:\\Documents and Settings\\Administrator\\.citespace”文件夹下创建一个名为“mysql.ini”的数据库文件，而后导入的数据文件信息都将存储于在此数据库中。图 6 为内置数据库操作界面，此界面可以划分为三部分，最上部分是功能菜单，中间部分为工程信息、最下面部分为 SQL 语句查询。

图 5 合作网络的地理图谱[27]

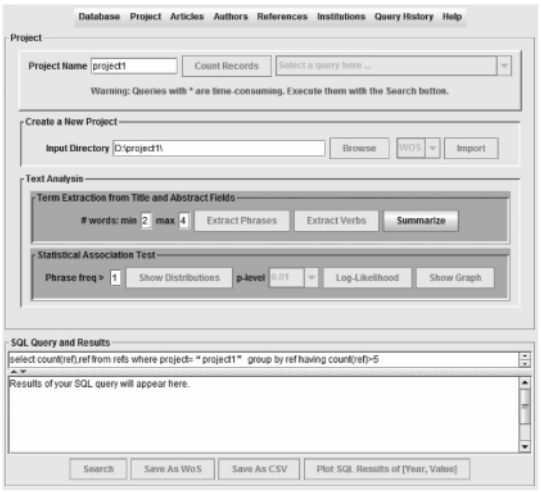


图 6 CiteSpace 内置数据库操作界面

例如，想要查询数据库文件“project1”（需事先创建并导入数据）中被引频次大于 5 的引文，可以在查询栏中输入“select count（ref），ref from refs where project = “project1” group by ref having count（ref）>5”进行查询，查询结果可以保存为 WoS 格式或 CSV 格式。

通过“update”SQL 语句，还可以对数据表中的数据进行修改。例如，输入“UPDATErefsSETref = Kuhn，T. S.，STRUCTURESCIREVOLU，1962’ WHEREproject = ’project1’ andref = ’Kuhn，T.，STRUCTURESCIREVOLU，1962’”，可以将所有写成“Kuhn，T.，STRUCTURESCIREVOLU，1962”的引文统一成“Kuhn，T. S.，STRUCTURESCIREVOLU，1962”。修改后的数据可以利用数据库的导出功能

重新生成 WoS 格式，以便在 CiteSpace 中重新运行和可视化。

(3) 鱼眼图。鱼眼视图技术(fish - eye)，一方面把人们感兴趣的研究区域放大显示，另一方面使焦点周围的信息内容逐渐缩小，而且保持着整体视图的可见性，这是一种 Focus + Context 技术。CiteSpace 为便于用户的分析，提供了基于时间线图的鱼眼图功能，图 7 显示的是一般时间线图和鱼眼图的比较。

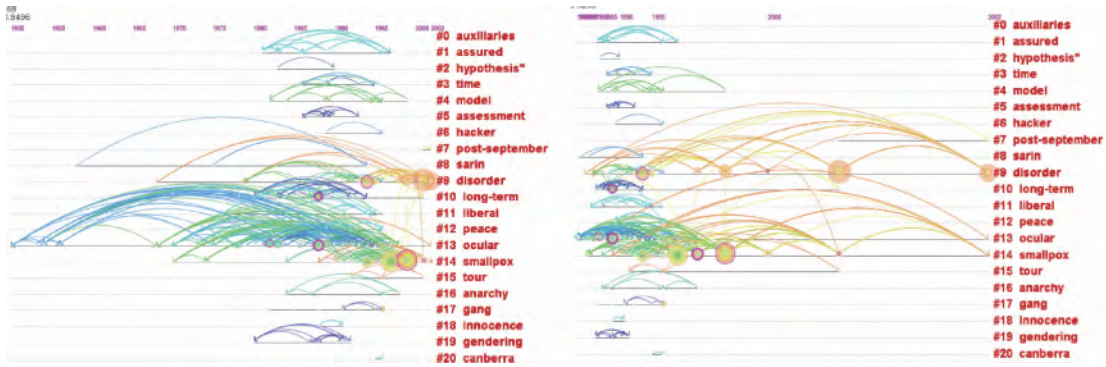


图 7　一般时间线图和鱼眼图的比较

(4) 双图叠加。双图叠加(overlay)功能是将一幅 CiteSpace 图谱上叠加到另一幅图谱之上，前者称为叠加图，后者称为底图(base map)。通过双图叠加功能，可以展现一张图谱所代表的知识领域在另一张图谱所代表的知识领域中的分布和地位。图 5(c)就是一幅双图叠加图谱，但目前的 CiteSpace 版本还无法实现。我们用现有的版本绘制了另一双图叠加图谱(图 8)，底图展现了 Scientometric 期刊论文中的共被引图谱，可以看出，该期刊主要分成了 7 个子领域; 另外再做一个引用普赖斯《小科学、大科学》一书的文献共被引图谱，并将其叠加到前者的底图上，这样就可以展现普赖斯的影响力主要体现在 Scientometrics 的 7 个子领域中的哪些方面。比如，图中可以看出，普赖斯的影响力主要体现在“科学合作”、“科学评价”、基于科学论文网络的“科学知识图谱”和包含洛特卡定律或普赖斯定律的“科学生产率”等领域中。

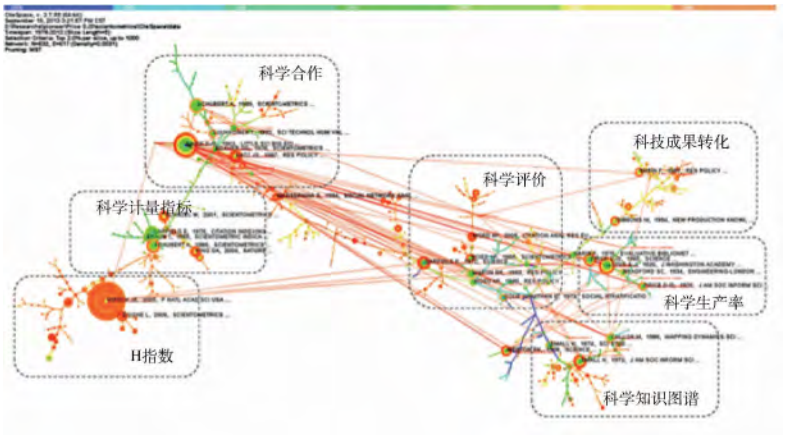


图 8　引用普赖斯的文献在整个 Scientometrics 期刊论文中的分布

5　结　论

本文从四个方面，即 CiteSpace 改变看世界方式的核心功能、对研究领域解释与预见上的理论功能、方法论功能的实现、及其应用功能的扩展，较为全面深入地论述了 CiteSpace 知识图谱的方法论功能。

(1) CiteSpace 知识图谱整合了视觉思维、数学思维和哲学思维，改变了人们认识世界的方式，即以世界 2 对世界 3 中“一阶科学”的一个知识领域文

献,进行可视化分析,以抽象程度更高的“二阶科学”范畴和直观形象的科学知识图谱,得以更深刻地反映和揭示世界1中一个现实领域的本质与规律,也有利于建立在世界3基础上的科学新发现。

(2) CiteSpace的设计及功能实现有着极其深刻的理论基础,包括库恩的科学发展模式理论、普赖斯的科学前沿理论、社会网络分析的结构洞理论、科学传播的信息觅食理论和知识单元离散与重组理论。这些理论的意义在于,强化图谱的可解读性、解读的合理性和正确性,通过图谱解读,实现理论两大功能,即领域现状的解释功能与领域未来前景的预见功能。

(3) CiteSpace的应用流程,汇集了体现合格知识图谱满意标准的主要流程,包括数据采集、数据处理、参数功能选择、可视化和解读,从而在软件运行操作上保证了方法论功能的实现。

(4) CiteSpace知识图谱问世之初仅限于展示知识领域研究前沿演进的基本功能,其后技术不断改进,功能不断拓展,包括地理地图、数据处理、鱼眼图和双图叠加等。

为充分发挥CiteSpace的方法论功能,避免知识图谱及其可视化工具的“滥用”与“误用”,本文对目前应用CiteSpace中存在的问题提出了相应的改进办法。

(1) 重视数据下载的准确性和完整性,可以灵活地采用多种下载数据的方式,如除了主题检索、关键词检索等还可以将文献作为检索依据;

(2) 明确科学前沿和研究热点的概念,研究前沿往往用是通过文献共引聚类的是施引文献的研究内容来表征,而研究热点往往通过引用突显来体现;

(3) 理解图谱美观的含义,灵活掌握点、线、色彩、标签的显示技巧,灵活使用各种图谱剪枝功能;

(4) 保证图谱的信息完整,即保留视图左上角的信息说明窗口;

(5) 深刻理解CiteSpace设计所基于的基本理论,从而领悟解读视图的视角,并灵活使用CiteSpace提供的各种节点、聚类信息表;

(6) CiteSpace知识图谱的绘制往往需根据图谱聚类效果进行多次绘制选择,主要依旧是聚类模块性指数Q值和聚类轮廓性指数S值;

(7) 随之CiteSpace日益广泛的应用,人们对它的功能也随之有了更多的需求,因而它的功能不断拓展,国内学界应不断关注并使用其新的功能。

参考文献:

[1] 陈悦,刘则渊. 悄然想起的科学知识图谱[J]. 科学学研究,2005,23(2):149-154.

[2] 陈悦,刘则渊. 科学知识图谱的发展历程[J]. 科学学研究,2008,26(3):449-460.

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

[4] 陈超美. CiteSpace Ⅱ: 科学文献中新趋势与新动态的识别与可视化[J]. 情报学报,2009,28(3):401-421.

[5] 刘则渊. 知识图谱的科学学源流[R]. 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PPT,2013-08-25.

[6] 侯剑华,胡志刚. CiteSpace软件应用研究的回顾与展望[J]. 现代情报,2013,33(4):99-103.

[7] 胡泽文,孙建军,武夷山. 国内知识图谱应用研究综述[J]. 图书情报工作,2013,57(3):131-137.

[8] 赵丹群. 基于CiteSpace的科学知识图谱绘制若干问题探讨[J]. 情报理论与实践,2012,35(10):56-58.

[9] 王钦炜. 基于CiteSpace Ⅱ的科学知识前沿图谱研究[D]. 北京: 北京大学,2011.8-13.

[10] 刘林青. 作品共被引分析与科学地图的绘制[J]. 科学学研究,2005,23(2):155-159.

[11] 陈超美. 序言二[A]. 刘则渊,陈悦,侯海燕. 科学知识图谱: 方法与应用[C]. 北京: 人民出版社,2008.4-5.

[12] 卡尔·波普尔. 客观知识—— 一个进化论的研究[M]. 上海: 上海译文出版社,1987.

[13] Price D J D. The science of science[A]. Goldsmith M, Mackay A L The Science of Science[C]. Souvenir Press,1964.

[14] Wouter P. The Citation Culture[D]. Amsterdam: University of Amsterdam,1999.5-9.

[15] 莫里兹. 数学家言行录[C]. 南京: 江苏教育出版社,1990:80.

[16] Chen C, Chen Y. Horowitz M, et al. Towards an explanatory and computational theory of scientific discovery[J]. Journal of Informetrics, 2009,3(3), 191-209.

[17] 库恩 T S. 科学革命的结构[M]. 上海: 上海科学技术出版社,1980.

[18] 刘则渊,陈悦,朱晓宇. 普赖斯对科学学理论的贡献——纪念科学计量学之父普赖斯逝世30周年[J]. 科学学研究, 2013, 31(12):1761-1772.

[19] Granovetter M. The strength of weak ties[J]. American

Journal of Sociology, 1973(5): 1360 – 1380.

[20] Burt R S. Structural Holes – The Social Structure of Competition [M]. Harvard University Press, 1992.

[21] Chen C, Hu Z, Liu S, et al. Emerging trends in regenerative medicine: Ascientometricanalysis inCiteSpace [J]. Expert Opinions on Biological Therapy, 2012, 12(5): 593 – 608.

[22] Chen C. Searching for Intellectual Turning Points: Progressive Knowledge Domain Visualization [C]. Proceedings of the National Academy of Sciences of the United States of America(PNAS), 2004, 101(Suppl. 1): 5303 – 5310.

[23] 赵红州,蒋国华. 知识单元与指数规律 [J]. 科学学与科学技术管理,1984,(9): 39 – 41.

[24] 刘则渊,侯海燕,陈超美,等. 知识计量学及其可视化技术的应用研究 [A]. 中国科学学与科学技术管理研究年鉴 2008/2009 年卷 [C]. 大连: 大连理工大学出版社.

[25] 武夷山. 做菜与科学计量学研究 [J]. 情报学报, 2013, 32(10): 编者的话.

[26] Chen C, Ibekwe – SanJuan F, Hou J. The structure and dynamics of co – citation clusters: A multiple – perspective co – citation analysis [J]. Journal of the American Society for Information Science and Technology, 2010, 61(7): 1386 – 1409.

[27] 陈超美. CiteSpace 3. 5. R7: Geospatial 功能已恢复. [EB/OL] http://blog. sciencenet. cn/blog – 496649 – 693498. html, 2014 – 10 – 20.

The methodology function of CiteSpace mapping knowledge domains

CHEN Yue[1], CHEN Chao – mei[2], LIU Ze – yuan[1], HU Zhi – gang[1], WANG Xian – wen[1]

(1. Joint – Institute for the Study of Knowledge Visualization and Science Discovery, Dalian University of Technology(China) – Drexel University(USA), WISE Lab, Dalian 116085, China;
2. College of Computing & Informatics, Drexel University, USA)

Abstract: The concept of Mapping Knowledge Domains and the CiteSpace (an information visualization tool) have been got a lot of attention quickly since they are introduced in domestic academia, the relevant literature spring up like mushroom in a variety of domestic academic journals in Information Science, Science of Science and Management. But we find the knowledge visualization tool is "abused" and "misused" seriously by reading more than 500 papers that applied CiteSpace, which shows that academic circles are lack of sufficient understanding of the methodology function of Mapping Knowledge Domains. The paper explains the methodology function of CiteSpace mapping from four following aspects, 1) elucidate its core function, changing the way we see the world, from the design concept, 2) elaborate on its theoretical function of interpretation and foresight to knowledge domain, 3) clarify the realization of the methodology function from applied process, and 4) introduce its application function from newly technology extension. We expect CiteSpace could better develop the function of methodology in detecting academic frontiers, choosing research directions, managing knowledge and making S&T decision.

Key words: Mapping Knowledge Doamains; methodology; CiteSpace