● 赵丹群（北京大学　信息管理系，北京　100871）

基于 CiteSpace 的科学知识图谱绘制若干问题探讨

摘　要： 科学知识图谱绘制是近年来科学计量学的一个新兴而活跃的研究方向。文章在对基于 CiteSpace 软件的国内科学知识图谱研究现状调研分析的基础上，从 4 个方面就图谱绘制问题进行分析论述。

关键词： 科学知识图谱；科学计量学；分析

Abstract: Mapping Knowledge Domains (MKD) is a newly-rising and active research direction in scientometrics. Based on the investigation and analysis of the status quo of the CiteSpace software-based research on MKD at home, this paper analyzes and discusses the problems relating to KMD form 4 aspects.

Keywords: mapping knowledge domains; scientific metrology; analysis

科学知识图谱（Mapping Knowledge Domains）是用于显示科学知识的发展进程与结构关系的一种图形，通常具有“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，可对知识单元或知识群体之间存在（或形成）的网络结构及其互动、交叉、衍化等诸多复杂关系进行表达和描述[1]。

CiteSpace 是美国著名华裔学者陈超美应用 Java 语言开发的一个可视化软件，它主要基于共引分析理论和寻径网络算法等，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（以关键论文为代表），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。作为一个专用的科学知识图谱绘制工具，自 2004 年 9 月推出以来，CiteSpace 已得到国际科学计量学界相关研究机构和人员的广泛使用[2]。

国内的科学知识图谱研究始于 2005 年。根据对 CNKI 期刊文献数据库的检索结果，2005—2011 年间国内以“科学知识图谱”或“知识图谱”为主题发表的论文达 323 篇，其中有 96 篇（约占 30%）涉及对可视化图谱绘制软件 CiteSpace 的讨论。论文按年发表数量及逐年快速上升趋势详见表 1。

表 1　国内有关“科学知识图谱”研究的期刊论文发表数量（单位：篇）

年份	2005	2006	2007	2008	2009	2010	2011	合计
每年论文总数	2	14	15	32	52	100	104	323
每年含 CiteSpace 的论文数	0	0	1	7	10	32	46	96

经对这些文献的阅读调查发现，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，并由此造成了一系列的科学知识图谱绘制及解读问题[3]。因此，笔者拟结合国内研究实际，从 CiteSpace 的科学知识图谱绘制流程入手，对其中的关键环节及其研究对策进行分析探讨，以助于国内相关研究工作的深入开展。

1　领域文献的查找

瑞典学者 P. Wouters 曾将科学的“表示”（Representation）问题概括抽象为 3 个不同层次，其中，科学文献（Scientific Literature）被认为是科学的“第一级表示”（First Order Representation），而对科学文献（集合）的引文分析，则形成了对科学的“第二级表示”（Second Order Representation）[4]。因此可以说，几乎所有的科学知识图谱绘制都首先要建立在对相关领域研究文献获取的基础上，CiteSpace 当然也不例外。

领域文献查找是 CiteSpace 软件使用的第一个关键步骤。目前 CiteSpace 可用的文献数据源主要包括 Web of Science、PubMed、Derwent 专利数据库、CSCD、CSSCI 等，通过文献检索方法获得的相关文献记录（包括题目、摘要、主题词、关键词和参考文献等），将成为随后图谱绘制的数据基础。

调查发现，CiteSpace 所需的领域文献主要通过以下两种策略获得：①基于关键词检索；②基于领域核心期刊定位。策略 1 的运用，首先需要确定一组可以表征某学科领域研究内容的关键词，然后通过对相关引文数据库的检索，即可获得一批在题目、关键词、摘要等字段中包含这组关键词的领域学术文献（以及它们的参考文献）。策略 2 的使用，则需要通过对领域专家的咨询或借助于某些期刊数据库产品（例如 JCR），先筛选出目标领域的核心期

刊，然后再以这些核心期刊中发表的所有学术文献作为领域图谱绘制的文献源。

从理论上讲，以上两种文献查找策略各有所长。如果目标领域较为明确且专指，策略1较为适宜；而当目标领域较为宽泛时，则策略2更为合适些。例如，陈超美等人2006年绘制“生物灭绝”（Mass Extinction，1981—2003年）和“恐怖主义”（Terrorism，1991—2003年）两个研究领域知识图谱时，采用的是策略1[5]，而他们2010年对“情报学”（Information Science）学科结构及其演变的可视化分析，则是基于策略2来完成的，从情报学领域12种核心期刊获取的10 853篇文献（1996—2008年）是他们此项研究的文献来源[6]。

在对国内研究文献调查时发现，对领域文献查找不重视、查找策略运用失当等现象较为普遍。具体表现为：仅选取某一种核心期刊作为领域文献来源；仅使用单一检索词进行文献查找（甚至仅在标题中匹配）；使用的检索词与图谱绘制领域语义失配，例如用“图书馆”指代“图书馆学”，“h指数”指代“h类指数”等；文献检索的时间跨度缺乏标准，随意性较强，等等。有的甚至没有明确提及具体的检索用词和检索策略，仅模糊地以“在Web of Science用主题词检索”一笔略过。所有这些问题，无不直接影响到领域文献检索的查全率和查准率，所获取的文献样本也无法完整反映（或代表）某学科的研究实际。以如此质量的文献源进行知识图谱绘制，即使花费再多的精力去详细解读，其意义和价值究竟又有几何？

2 突变词语的侦测

CiteSpace软件的核心功能之一是探测和分析学科研究前沿的历时性变化趋势以及研究前沿与其知识基础之间的关系。为此，其设计者陈超美博士使用了以下3个基本概念对CiteSpace的工作原理进行阐释。这3个基本概念分别是“研究领域”（Specialty）、“研究前沿”（Research Fronts）和“知识基础”（Intellectual Bases），它们之间的关系则定义为 $\Phi(t):\Psi(t)\rightarrow\Omega(t)$。其中，$\Psi(t)$ 表示某研究领域的“研究前沿”，它由一组在 t 时刻与研究前沿新趋势（或动态）密切相关的专业术语或短语组成；而 $\Omega(t)$ 代表该研究领域的“知识基础”，它主要由包含前沿术语的研究论文所引用的大量参考文献组成（包括引文信息和共引信息）；而“研究领域”$\Phi(t)$ 则被概念化为一个从“研究前沿”到“知识基础”的时间映射。对上述基本概念及其关系的一个更为形式化的描述如下[4]：

$$\Psi(t)=\{\text{term}\mid\text{term}\in S_{\text{title}}\cup S_{\text{abstract}}\cup S_{\text{descriptor}}\cup S_{\text{idertifier}}\wedge\text{IsHotTopic}(\text{term},t)\}$$

$$\Omega(t)=\{\text{article}\mid\text{term}\in\Psi(t)\wedge\text{term}\in\text{article}_0\wedge\text{article}_0\rightarrow\text{article}\}$$

其中，S_{title}，S_{abstract}，$S_{\text{descriptor}}$和 $S_{\text{ldermifier}}$分别表示来自论文题目、摘要、主题词和关键词字段的一系列专业术语；IsHotTopic（term，t）是一个布尔函数，表示 t 时刻term是否为一个热点术语；而 $\text{article}_0\rightarrow\text{article}$ 表示论文 article_0 引用了论文article。

具体来说，CiteSpace对研究前沿的侦测分析主要通过对“突变词语”的提取来实现，其内嵌的“Find Burst Phrases”算法功能，会将出现频次快速增加的专业术语（即突变词语）确定为研究前沿术语，而算法思想则主要源于2003年J. Kleinberg提出的“突变侦测算法”（Burst Detection Algorithm）[7]。

目前，CiteSpace的“Find Burst Phrases”算法允许从论文题目、关键词和摘要等字段提取候选专业术语，通过跟踪分析它们在不同时间区间内出现频率的突然变化（激增），识别出代表研究前沿的若干名词术语。进一步地，对前沿术语所在论文的参考文献集合进行共引分析，构建形成包含参考文献和前沿术语及其相互关系的混合性网络，以辨识研究前沿的结构及发展演化。

因此，对于CiteSpace来说，突变词语的侦测是完成图谱绘制的关键步骤之一。不过，CiteSpace目前并不允许使用者自行修改“Find Burst Phrases”算法中的相关参数。这种做法虽简化了操作使用难度，但也在一定程度上限制了软件应用的自由度和个性化设置水平。而从已有的研究文献看，具体涉及突变词语侦测的文献比例也很少，对CiteSpace这一功能的关注及讨论存在较多空白。

3 时区分割与相关参数的阈值设置

“时区分割”是CiteSpace软件的另一个重要特色功能，它在使用过程中要求用户进行时区分割（Time Slicing），以确定单个时间片的长度，以便对某个知识领域进行时序“抓拍”，然后将这些分时抓拍的图片连接起来，完成历时性研究视角下的学科知识图谱绘制任务。目前，CiteSpace允许用户自定义时间间隔的大小，以增强软件应用的灵活性，但设置多大的时间间隔是适宜和合理的并不能一概而论。一般来说，时间间隔越小，图谱绘制对所在领域知识演进趋势揭示的时间敏感性就越大。但过小的时间间隔设置，又会导致一系列被设计用来侦测突变的观测项目因在相邻时区内的变化幅度小而不易被发现，进而影响到对关键节点的识别。

与“时区分割”同时存在的一个伴生问题是图谱网络节点显示阈值的设置。CiteSpace软件内置有多种不同的图谱网络节点显示阈值设置方案，其中效果较好且复杂度较高的一个方案是“c-cc-ccv”，即引文数量（citation，

c)、共被引频次（co-citation，cc）和共被引系数（co-citation coefficient，ccv）。需要注意的是，CiteSpace 软件要求对不同时间分区内的节点显示阈值分别进行设置，为简化设置操作，用户可仅对前、中、后 3 个时间分区中的阈值进行设置，其余时间分区的阈值设置可由软件利用插值算法自动生成。鉴于时间间隔设置的不确定性及其对随后的相关参数（例如“c-cc-ccv”）阈值设置影响的难以掌握和控制，可以说，时区分割及相关参数阈值的合理设置问题已成为 CiteSpace 软件应用中的一大难点，应引起研究人员和使用者的高度关注。

目前，国内的研究现状是很多研究人员在使用 CiteSpace 绘制各类学科知识图谱时，很少关注或几乎不论及自己的时区分割及相关参数阈值的设置依据，而热衷于把更多的精力集中在对绘制图谱的分析解读上。这不能不说是国内科学知识图谱研究中存在的一个大“误区”。

笔者以为，对于 CiteSpace 软件使用过程中不可避免的时区分割和参数阈值设置问题，以后不妨采取一种更为谨慎的态度：选择多套不同的设置方案，分别绘制图谱并进行比较分析，以便从中选取一个更为贴近学科发展与研究实际的知识图谱绘制结果。此外，在相关参数阈值设置时，还可将一些其他因素考虑进来，例如目标领域的研究规模、引文增长情况、可视化观测需要等。

4 图谱解读

任何一项科学知识图谱研究工作都离不开对所绘制图谱的解读分析，而准确、充分的图谱解读也是科学知识图谱研究发挥效用的重要体现。不可否认，即使拥有了先进的图谱绘制工具，图谱解读仍是有一定难度的，因为它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。

很多研究人员在解读图谱时往往受限于个人的主观判断和感受，忽视了解读工作所要求的科学性和规范性，以致出现诸如错误解读、遗漏解读、过度解读等一系列问题，这一点在国内 CiteSpace 软件的应用研究中尤为突出。CiteSpace 绘制流程的最后一步要求对图谱中的关键节点进行验证，验证方式主要采取同行验证法。但从已发表的国内研究论文看，几乎很难发现验证分析的文字。验证缺失不仅造成图谱解读的不规范，而且也在客观上阻碍了图谱绘制效用的发挥，进而影响到对图谱绘制工具的改进、优化以及在更高层次上对图谱绘制理论的不断创新。

追溯科学知识图谱的研究历史，不论是 20 世纪 60 年代 E. Garfield 等人手工绘制完成的 DNA 研究领域的知识演进图谱[8]，还是 20 世纪 70 年代 H. G. Small 和 B. C. Griffith 等人对自然科学领域学科（专业）结构的共引聚类分析[9-10]，直至近期陈超美等人基于 CiteSpace 对“生物灭绝”和“恐怖主义”两个研究实例的图谱绘制[5]，无不都在图谱解读过程中严格地执行了同行验证步骤。这种严谨而富有科学精神的学术传统，亟待回归并切实贯彻于当前的国内科学知识图谱研究中。

5 结束语

20 世纪 80 年代初，当引文分析方法首次被介绍到我国时，迅即在国内学术界掀起一场研究热潮。如今 30 年过去了，自 2005 年科学知识图谱研究在国内兴起以来，其研究进展状况及存在问题竟与当年的引文分析研究存在惊人的相似之处，这种相似可借用如下的一句比喻予以形象地描述和表达 “Give a small boy a hamper，and he will find that everything he encounters needs pounding”。本文对 CiteSpace 及其科学知识图谱绘制若干问题的上述讨论分析，希望能对国内相关研究的深入开展有所助益。□

参考文献

[1] 刘则渊，陈悦，侯海燕. 科学知识图谱：方法与应用[M]. 北京：人民出版社，2008：3-5.

[2] CiteSpace：visualizing patterns and trends in scientific literature[N/OL]. [2011-11-20] http://cluster.cis.drexel.edu/~cchen/citespace.

[3] 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.

[4] WOUTER P. The citation culture [D]. Amsterdam：University of Amsterdam，1999：5-9.

[5] CHEN Chaomei. 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.

[6] CHEN Chaomei，IBEKWE-SANJUAN F，HOU Jianhua. 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.

[7] KLEINBERG J. Bursty and hierarchical structure in streams [J]. Data Mining and Knowledge Discovery，2003（7）：373-397.

[8] GARFIELD E，SHER I，TORPIE R J. The use of citation data in writing the history of science [R]. Philadelphia：Institute for Scientific Information，1964：86.

[9] SMALL H. G，GRIFFITH B C. The structure of scientific literature Ⅰ：identifying and graphing specialties [J]. Science Studies，1974，4（1）：17-40.

[10] GRIFFITH B C，SMALL H G. The structure of scientific literature Ⅱ：toward a macro-and microstructure for science [J]. Science Studies，1974，4（4）：339-365.

作者简介：赵丹群，女，1966 年生，副教授，博士。

收稿日期：2012－04－23