

CiteSpace:

Text Mining and
Visualization in Scientific Literature

CiteSpace: 科技文本挖掘及可视化

李 杰 陈超美 © 著

千百年来，人类的学习以记诵方式为主，
听觉器官发挥着很大的作用。

随着信息技术的飞速进步，**可视化**应用越来越普及，
今后的学习越来越多地借助各种**可视化**手段，
视觉器官将发挥前所未有的作用。

可视化方式成为主流学习方式后，
人类的学习效率将大大提高，
有可能带来一场认知革命。

为了适应这样的进程，
知识组织方式也必将走向**可视化**之路。

 首都经济贸易大学出版社
Capital University of Economics and Business Press

CiteSpace:

科技文本挖掘及可视化

李 杰 陈超美 © 著

 首都经济贸易大学出版社
Capital University of Economics and Business Press

· 北 京 ·

图书在版编目(CIP)数据

CiteSpace: 科技文本挖掘及可视化 / 李杰, 陈超美著.
—北京: 首都经济贸易大学出版社, 2016.1
ISBN 978-7-5638-2464-9

I. ①C… II. ①李… ②陈… III. ①可视化软件
IV. ①TP31
中国版本图书馆CIP数据核字(2015)第304197号

CiteSpace: 科技文本挖掘及可视化

李杰 陈超美 著

出版发行 首都经济贸易大学出版社
地 址 北京市朝阳区红庙(邮编 100026)
电 话 (010) 65976483 65065761 65071505(传真)
网 址 <http://www.sjmcb.com>
E-mail publish@cueb.edu.cn
经 销 全国新华书店
照 排 首都经济贸易大学出版社激光照排服务部
印 刷 北京九州迅驰传媒文化有限公司
开 本 710毫米×1000毫米 1/16
字 数 356千字
印 张 20.25
版 次 2016年1月第1版 2016年1月第1次印刷
书 号 ISBN 978-7-5638-2464-9 / TP·41
定 价 120.00元

图书印装若有质量问题, 本社负责调换

版权所有 侵权必究

序一

人类文明的进展之路，就是工具不断替代和补充人力之路。一开始，人们用工具代替双手双脚，将自身从繁重的体力劳动中解放出来；近年来，随着人工智能研究、大数据情报学研究、认知科学研究等方面的进展，人的脑力劳动也有望被广义的工具（包括计算机软件）所部分地代替或增效。

千百年来，人类的学习以记诵方式为主，听觉器官发挥着很大的作用。随着信息技术的飞速进步，可视化应用越来越普及，今后的学习越来越多地借助各种可视化手段，视觉器官将发挥前所未有的作用。由于视觉器官在单位时间内的信息吸收能力大大强于听觉器官，可视化方式成为主流学习方式后，人类的学习效率将大大提高，有可能带来一场认知革命。为了适应这样的进程，知识组织方式也必将走向可视化之路，图书情报研究人员在知识可视化征程中将发挥非常重要的作用。在这样的大背景下，应该承认，美国德雷塞尔大学计算机与情报学学院陈超美教授开发出广受欢迎的信息可视化软件 CiteSpace，是符合时代潮流的一项重要成就。

在人类发展的任何阶段，人的技术水平主要表现在两个方面：一是不断出现的、体现着最新技术成果的新工具，二是对已有工具的熟悉程度和掌握利用程度。这两方面都非常重要！对于中国的古人来说，能锻冶出干将、莫邪这样的宝剑，是了不起的；能像庖丁解牛那样熟练地用刀，也是了不起的。您瞧，“今臣之刀十九年矣，所解数千牛矣，而刀刃若新发于硎”，刀用了十九年了，解牛有几千头了，刀刃仍旧不钝、不卷，像新的一样，这里面有多深的功夫啊！对于今人来说，像陈超美教授这样开发出深受用户欢迎的 CiteSpace 软件，是了不起的成就；像首都经济贸易大学李杰博士这样把 CiteSpace 钻深钻透，能够写出 CiteSpace 的使用教程，也是相当难能可贵的！

本书两位作者都是学术园地的勤奋耕耘者。在完成本书时，李杰还是一名在读博士生，但已经发表了数十篇论文和两本著作。据李杰对 CiteSpace 软件更新手记的分析，自 CiteSpace 于 2003 年问世以来，至 2015 年 6 月 6 日，软件累计更新次数达 274 次。为便于计算，我们假定以 2003 年年中作为 CiteSpace 问世的起点，则 12 年来，该软件大约每 16 天就更新一次！一方面，这表明了陈超美教授的勤奋；另一方面也可以看出，由于 CiteSpace 深受广大用户欢迎，用户对它的期望值也越来越高，从而对陈教授产生了与时俱进、精益求精的推动力。

国内不知有多少人使用过 CiteSpace 软件，并根据该软件的分析结果发表了论文，但可能没有几人读过陈教授的四本著作。我呼吁，热爱 CiteSpace 的学人都应该好好读读这四本书，从而对陈教授的学术思想有更完整的把握：

1. (2011) *Turning Points: The Nature of Creativity* (转折点：创造力之性质). Springer and Higher Education Press.

2. (2004) *Information Visualization: Beyond the Horizon* (信息可视化：走出地平线). (2nd Edition). London: Springer. (Paperback: 2006)

3. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer, 该书中译本《科学前沿图谱：知识可视化的探索》于 2014 年 7 月由科学出版社推出。

4. (1999) *Information Visualisation and Virtual Environments* (信息可视化与虚拟环境). London: Springer-Verlag London.

笔者作为情报学领域的一名老兵，阅读、浏览过很多借助 CiteSpace 工具写出的论文，我一方面为该工具在中国的火爆而高兴；另一方面，也为其中相当一部分作者的懒惰而悲哀，因为他们的论文缺乏思想闪光点，只是通过 CiteSpace 的处理，简单地将有关数据展现得更漂亮而已。我相信，陈超美教授也不希望自己的软件只起到化妆品式的作用。今后如何杜绝这一类论文呢？首先，作者们应该知道，软件工具的设计者是有思想的，我们应该努力学习、把握他们的思想，如果自己不肯动脑筋，随便拽一个软

件就用，也许论文是得以发表了，但对自己的学术进步并没有多大的助力。其次，CiteSpace 具有非常丰富的功能，而我们多数利用 CiteSpace 发表文章者，只涉猎了该软件功能的一点皮毛。因此，认真阅读此书，更全面地掌握这个软件，今后一定能使我们的研究如虎添翼。

我从 2015 年 2 月起被调到中国科学技术发展战略研究院工作，依依不舍地离开了情报学界。但本书两位作者仍然热情地邀请我作序，我感到，却之不恭，应允下来却惴惴然。草成数言，希望没有耽误读者的时间。

是为序。

中国科学技术发展战略研究院研究员

武夷山

2015 年 10 月 1 日

序二

在科学探索中，无论是对于初出茅庐的年轻学者，还是对于训练有素的行家里手，最关注的莫过于在自己从事的知识领域，从海量的文献数据中了解到最感兴趣的主题及其科学文献，找到其中最为重要、关键的有效信息，弄清其过去与现在的发展历程，识别最活跃的研究前沿和发展趋势。

这些都是科学探索面临的首要难题。进入 21 世纪以来，一些信息可视化技术相继应运而生，为尝试解决这些难题进行了可贵的探测，提供了有益的线索。其中，国际著名的信息可视化专家陈超美教授用 Java 语言开发的、基于引文分析理论的信息可视化软件 CiteSpace，就是可以解决上述一系列难题的一种工具与技术。其突出特征是把一个知识领域浩如烟海的文献数据，以一种多元、分时、动态的引文分析可视化语言，通过巧妙的空间布局，将该领域的演进历程集中展现在一幅引文网络的知识图谱上；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，显示出图谱本身的可解读性。这两大基本特征就是我对 CiteSpace 知识图谱形态的概括：“一图展春秋，一览无余；一图胜万言，一目了然。”

因此，该软件一经问世，就以其神奇的魅力征服了科学计量学界，受到广大学术界的青睐，迅速传播到世界各地，被广泛应用于各个知识领域的可视化分析。如今，基于 CiteSpace 的知识图谱，如山花浪漫，技压群芳，异彩纷呈，成为知识世界百花园中盛开的一朵朵奇葩。

现在呈现在读者面前的《CiteSpace：科技文本挖掘及可视化》一书，不独可以引领初学者步入 CiteSpace 之门，而且可以帮助有兴趣者进一步训练，熟练地掌握它，绘制出合格满意的知识图谱。本书作者是年轻的学者李杰和 CiteSpace 的开创者陈超美。本书在依据陈超美教授的 CiteSpace

英文版手册的基础上,借鉴和吸收了陈悦、陈超美等所著《引文空间分析原理与应用: CiteSpace 实用指南》(以下简称《指南》)的成果,也包含了第一作者本人使用 CiteSpace 等信息可视化软件著述《安全科学知识图谱导论》的研究经验。这里不妨对中外三部 CiteSpace 手册性、普及性读物略加比较,以阐释这本著作出版的价值与必要性。

本书的主要内容,源自陈超美的 CiteSpace 英文手册。2015 年 11 月 26 日由陈超美本人将手册内容和 CiteSpace101 网站的资料,整理成电子书《How to Use CiteSpace》。该电子书反映了作者开发 CiteSpace 的初衷,分 10 章全面介绍了 CiteSpace 的各项功能、基本流程和操作细节,以及其他可视化软件要点,并用了 180 多幅图谱和若干经典案例,娓娓道出了如何使用 CiteSpace 来绘制满意的知识图谱。手册和该书的内容,处处体现出作者着眼于用户的特点、使用和需求。作者明确表示:这本电子书的内容将不断更新完善,并与 CiteSpace 新版软件保持同步。一个月之后的 12 月 26 日,《How To Use CiteSpace》修订版上网,新增了 4.0.R5 SE 版本的介绍与实例。我以为英语熟练的初学者可以直接阅读这本电子书,并时时关注 CiteSpace 及电子书的版本更新。当然,如果对照本书阅览电子书,既可加深对此书有关操作内涵的理解,又可认识电子书有关功能扩展的意义和作用。

本书参考了《指南》一书,吸收了其中有关理论基础的论述。《指南》是陈超美作为大连理工大学长江学者讲座教授,率领 WISE 实验室团队率先在中国应用和推广 CiteSpace 知识可视化技术的经验总结。该书原先拟在 2009 年编著出版,但在著述过程中发现 CiteSpace 的传播应用非常迅速,并了解到部分期刊文献出现信息可视化工具“滥用”“误用”的情况,CiteSpace 知识图谱良莠不齐,甚至不合格,严重损害了知识图谱的声誉。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和偏颇。因此,《指南》一书首先将开发和改进 CiteSpace 工具的背后所坚守的宏观哲学理念和相关理论基础向读者坦诚地披露出来,并从 CiteSpace 使用流程阐明其方法论功能的实现,最后

专用一章针对 555 篇国内运用 CiteSpace 工具的调查情况，归纳出 39 个常见问题，一一解答如何纠偏与处理。从软件蕴涵的理论基础和运用中的问题症结，来阐述其使用流程，构成《指南》的特色。

与 CiteSpace 英文手册或电子书和《指南》一书相比，《CiteSpace：科技文本挖掘及可视化》突出了 CiteSpace 区别于其他信息可视化软件的特色与优势，这在很大程度上得益于第一作者李杰在其专著《安全科学知识图谱导论》（以下简称《导论》）撰写过程中，奠定了厚实的科学计量学及知识图谱理论基础。而这得到了合作者陈超美对《导论》的高度评价，从而形成两位作者的共识。陈超美在《导论》一书的序言中指出：“李杰在本书中详细地展示了如何巧妙地运用一组最常用的科学图谱工具，包括加菲尔德的 HistCite、印地安纳大学的 SCI2、荷兰莱顿大学的 VOSViewer 和我的 CiteSpace，以及通用网络可视化软件 Pajek 和 Gephi，通过对中外相关文献的分析来了解安全科学的各个方面，为读者展示了灵活运用现有工具的能力。”无疑，多种工具在实际运用中的比较，显露出 CiteSpace 的独特功能与优势。

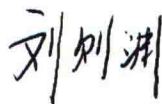
正是基于以上达成的共识，本书全面系统地陈述了正确使用 CiteSpace 软件的基本流程与操作程序，从数据来源与科技文本挖掘，到软件的界面功能与功能模块，并结合实际案例讲解 CiteSpace 的文献共被引分析与耦合分析、科研合作网络分析、共词分析与领域共现网络分析、网络叠加与双图叠加功能拓展，以及基于 CiteSpace 的火灾科学研究。全书还有一个显著特点，就是几乎每一个重要步骤和关键环节，都独具匠心——加注“小提示”，实现了整个使用流程的可操作性。全书分为八讲，每一讲末尾都列出一系列“思考题”，供读者自己复习、回味和总结，推进了知识可视化技术的广谱性。值得赞叹的是，全书除了穿插有大量的统计图表和软件界面截图外，还匹配了大量形态各异的 CiteSpace 知识图谱，令人信服，展现出又一个又一个知识领域演进的“一图展春秋”意境，蕴涵着知识图谱的可解释性与可预见性。

我相信这部著作定会在 CiteSpace 知识可视化技术的传播普及中发挥

巨大的作用。当然,在我看来,中外三本 CiteSpace 普及读本各有所长,本书突出软件全流程的可操作性,《指南》强调软件蕴藏的理论性和运行的针对性,电子书的原创性与软件功能拓展的同步性,均可在 CiteSpace 的传播普及中发挥各自优势、彼此配合、相得益彰、并行成长、升级再版;三本书的最大公约数是都包含共同作者陈超美,显然其独著的《How to Use CiteSpace》通过软件版本升级以保持原创性,始终扮演着主导引领的角色。

我曾经说过:“视觉思维乃是 CiteSpace 系统不言而喻的主要思维方式。视觉在人类感知外部信息中起绝对主导的作用,图像又是视觉信息的第一要素。不能把视觉思维误解为传统的感性认识。视觉思维既可以从感性视觉,到抽象思维,再到理性直观的螺旋式上升过程,也可以跨越感性视觉,直接把抽象信息与数据变换为可视化的空间结构与知识图谱。”^①

我们欣喜地看到,在大约 14 年间,基于知识单元的 CiteSpace 可视化软件从 1.0 版升级到 4.0 版,知识可视化技术正是以独到的视觉思维方式发展而不断更新换代。人们可以期待,随着视觉思维方式向深度和广度的变革,知识可视化技术必将进一步迈向新的发展阶段。



大连理工大学科学学与科技管理研究所
暨 WISE 实验室教授、博士生导师
2015 年 12 月 28 日于大连新新园

^① 刘则渊.《科学前沿图谱:知识可视化探索》序.北京:科学出版社,2014.

第 1 讲	CiteSpace 应用总述	1
1.1	软件作者及 CiteSpace 简介	2
1.2	CiteSpace 应用现状的总结	8
1.3	CiteSpace 应用需要注意的问题	25
1.4	CiteSpace 操作问题解答	27
1.5	本书的逻辑结构及安排	28
第 2 讲	CiteSpace 分析数据采集	33
2.1	数据库及数据概述	34
2.2	中文数据采集	36
2.3	外文数据采集	44
第 3 讲	CiteSpace 安装及界面功能	61
3.1	CiteSpace 的下载与安装	62
3.2	CiteSpace 初次启动及案例运行	65
3.3	CiteSpace 界面及功能	68
3.4	CiteSpace 对数据的预处理	114
3.5	CiteSpace 分析项目的建立	129
3.6	CiteSpace 操作关键及结果解读	136

第 4 讲	CiteSpace 文献共被引和耦合分析	141
4.1	认识共被引与耦合	142
4.2	科技文献共被引分析	146
4.3	施引文献的耦合分析	174
第 5 讲	CiteSpace 科研合作网络分析	179
5.1	认识科学合作分析	180
5.2	作者 - 机构 - 国家 / 地区合作网络	181
5.3	科学合作网络的地理可视化	186
第 6 讲	CiteSpace 主题和领域共现分析	193
6.1	认识词频和共词分析	194
6.2	CiteSpace 关键词共现网络	196
6.3	CiteSpace 名词性术语共现网络	199
6.4	科学领域的共现分析	203
第 7 讲	CiteSpace 高级功能补充	207
7.1	网络图层的叠加分析	208
7.2	期刊的双图叠加分析	212
7.3	文本挖掘及可视化模块	218
7.4	CiteSpace 与数据库的结合	225
7.5	CiteSpace 与外部软件结合	228
第 8 讲	基于 CiteSpace 的火灾科学研究	245
8.1	中国火灾科学研究	246
8.2	国际火灾科学研究	253

参考文献·····	267
附录 1 CiteSpace 文献列表·····	277
附录 2 常见数据库数据格式·····	288
附录 3 20 种常用免费科技文本挖掘及可视化软件·····	304
附录 4 教程涉及数据分享列表·····	305
后 记·····	307

第 1 讲

CiteSpace 应用总述



1.1 软件作者及 CiteSpace 简介

陈超美(Chaomei Chen)教授是美国德雷赛尔大学计算机与情报学教授,从 2008 年开始担任大连理工大学长江学者讲座教授,同时也是 Drexel-DLUT 知识可视化与科学发现联合研究所(美方)所长。他被国内外同行专家评价为当代信息可视化与科学知识图谱学术领域中的国际顶尖级领军人物。2004 年在“作者共被引分析(Author Co-citation analysis, ACA)”的诞生地——美国德雷赛尔大学信息科学与技术学院,陈超美教授使用 Java 语言开发了 Information Visualization-CiteSpace 信息可视化软件。陈超美教授开发 CiteSpace 软件(最早称为 StarWalker 软件)的主要灵感来自库恩(Thomas Kuhn, 1962)的科学结构的演进,主要的观点认为“科学研究的重点随着时间变化,有些时候速度缓慢(incrementally),有些时候会比较剧烈(drastically)”,科学发展是可以通过其足迹从已经发表的文献中提取的。

陈超美和刘则渊教授及其在大连理工大学的 WISE 团队将 CiteSpace 的理论基础系统地总结为五个方面(陈悦等, 2015):

(1)库恩的科学发展模式理论:即科学发展是科学革命的历史过程(前科学→常规科学→科学危机→科学革命→新常规科学),科学发展的本质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。

(2)普赖斯(Derek John de Solla Price)的科学前沿理论:在贝尔纳“科学发展模式的网状思想”和加菲尔德(Eugene Garfield)发明的“引文数据库”基础上,普赖斯在其《科学论文网络》(de Solla Price, D. J., 1965)中形成“参考文献的模式标志科学研究前沿的本质”理论,并认为“研究前沿是基于新近研究成果,随着发展知识网络也会变得越来越密”。

(3)社会网络分析的结构洞理论(Structure hole):该理论来源于格兰诺维特(Granovetter, 1973)提出的“弱关系的强度”。在此基础上,博特(Ronald S. Burt, 1992)提出结构洞,并认为处于结构洞位置的个体

通过信息过滤而能获得更多的竞争优势和创新能力。在 CiteSpace 中，使用节点在网络中的中介中心性来测度结构洞（Freeman, 1979; Brandes U, 2001）以及转折点（Turning points）。

（4）科学传播的信息觅食理论（Information foraging）：该理论描述信息搜索就像人类和动物捕获食物，认为我们在信息搜索中倾向于能量消耗最小化。在最优信息觅食理论和隐马尔科夫模型（Hidden Markov Model, HMM）基础上，陈教授等提出了一种集成视觉导航策略研究方法来以最小搜索成本获取最大效益。

（5）知识单元离散与重组理论：该理论由我国科学计量学家赵红州提出，他认为，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。

CiteSpace 是 Citation Space 的简称，可译为“引文空间”。CiteSpace 是一款着眼于分析科学分析中蕴含的潜在知识，并在科学计量学（Scientometric）、数据和信息可视化（Data and information visualization）背景下逐渐发展起来的一款引文可视化分析软件。由于是通过可视化的手段来呈现科学知识的结构、规律和分布情况，因此也将通过此类方法分析得到的可视化图形称为“科学知识图谱”（Mapping knowledge domains, MKD），即有图和谱的双重特性。该软件最初专门针对文献的共引进行分析，并挖掘引文空间的知识聚类 and 分布。随着 CiteSpace 的不断更新，它已经不仅仅提供引文空间的挖掘，而且还提供其他知识单元之间的共现分析功能，如作者、机构、国家 / 地区的合作等。

从陈超美教授的 Google Scholar 论文列表中了解到，其论文被引用最多的也是关于 CiteSpace 原理及其应用案例的经典论文 CiteSpace II : Detecting and visualizing emerging trends and transient patterns in scientific literature（中文译为：《CiteSpace II : 科学文献中新趋势与新动态的识别与可视化》，下文称为 CiteSpace 经典文献），截至 2015 年 6 月 21 日已经引用达 1 083 次（见图 1.1）。该论文截至 2015 年 7 月 9 日在 Web of Science（下文简称为 WoS）中被引用了 244 次（见图 1.2）。

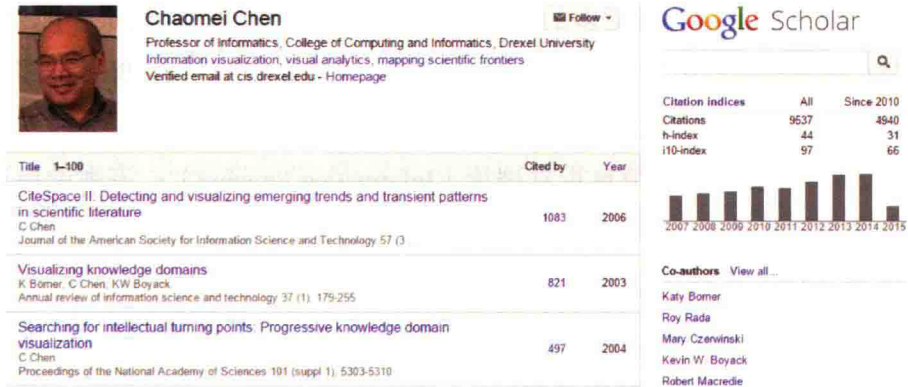


图 1.1 陈超美教授 Google Scholar 主页

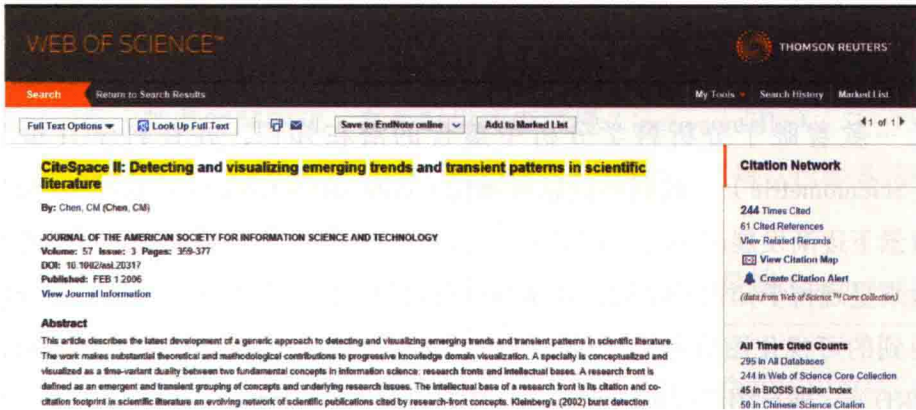


图 1.2 CiteSpace 经典文献在 WoS 中的引用情况

通过对 CiteSpace 使用过程状态的检测，得到了 2013 年 8 月至 2015 年 5 月的全球 CiteSpace 使用状态的分布，见图 1.3。从图中能够明显地得到 CiteSpace 使用的三大区域，分别为亚洲、欧洲以及美国。在亚洲，中国对 CiteSpace 的使用占了绝大部分。可以进一步对我国使用 CiteSpace 软件的城市分布视图进行了局部放大，参见图 1.4。

CiteSpace 记录了 88 个使用过程中的状态，统称为 events。比如，使用的版本，用户的 IP，聚类标签的选择和网络大小，等等。图点的大小代表 events 的数量，可以在线打开该版本 <http://cdb.io/1V940uI>。