

科技创新路径识别研究进展:方法与工具*

■ 祝娜 王效岳 白如江

[摘 要] 为支持科技知识创新,科技创新路径识别已经成为情报学研究的热点主题。在梳理科技创新路径识别研究的数据源、研究方法以及可视化工具研究现状的基础上,重点分析基于关键词或者主题词识别和基于引文全文的语义分析识别两种主要的科技创新路径识别方法,并对 Sci2、HistCite、CiteSpace 以及 VOSviewer 4 种可视化工具进行比较分析,进而指出基于引文分析的科技创新路径识别研究面临的挑战并展望其在数据集、研究方法、研究技术以及研究结果方面的未来发展趋势,以期为科技创新路径识别研究提供支持和帮助。

[关键词] 科技创新 科技创新路径 引文分析 可视化

[分类号] G250.2

DOI:10.13266/j.issn.0252-3116.2014.13.022

1 引言

进入 21 世纪以来,科技创新已成为国际竞争成败的主导因素,科技发展领域的竞争已经演变为国际竞争的一个焦点。1997 年 12 月,中国科学院提交了《迎接知识经济时代,建设国家创新体系》报告;2012 年 9 月 23 日,中共中央、国务院印发的《关于深化科技体制改革 加快国家创新体系建设的意见》中提出全面落实《国家中长期科学和技术发展规划纲要(2006-2020 年)》,加快国家创新体系建设的意见。在此背景下,为了支持科技创新和科技决策,运用情报分析方法识别科技创新路径,揭示科技创新发展过程,显得尤为重要。

科技创新一般是在创新基础上的再创新,科研人员通过引用他人科技创新成果来帮助实现新的创新。科技文献之间的引用关系构成了科技创新的动态发展路径,为科技创新路径识别提供了研究方向。由于科技创新具有继承性和累积性,因此科技创新路径识别是指以数据挖掘等相关技术做支撑,深入发现隐藏在图书、期刊、网页等大量文献资源数据背后的信息并以一定形式显示、存储或者推送这些信息,从而帮助用户简捷、快速地获得科技创新的相关知识。传统基于引文分析法识别科技创新路径的方法主要以科技文献之间的引用频次为基础,注重将科技文献内容进行量化

统计,忽略科技文献引用动机,致使科技创新动机模糊,未深入到科技文献中对科技创新点进行系统的分析研究。随着计算机信息处理技术以及文本挖掘技术的发展,科研人员开始利用语义分析等方法分析科技文献文本数据,构建包含具体科技创新内容的科技创新路径。本文在梳理科技创新路径相关理论与实践研究的基础上,通过对科技创新路径识别涉及数据源的选择,科技创新数据识别方法以及科技创新路径可视化 3 个方面的分析研究,指出目前科技创新路径识别研究面临的挑战并展望其未来发展趋势。

2 科技创新与科技创新路径

2.1 科技创新

1952 年美国学者巴伯在《科学与社会秩序》一书中提出“科学创新”即科学发现或发明^[1]。1998 年中国学者何传启提出科学知识创新是指通过科学研究获得新的自然科学、社会科学和技术科学知识^[2]。目前存在许多与“科技创新”相似或者相近的概念,如“科学前沿”、“创新知识”、“科学热点”等,本文将科技创新定义为科学知识创新与技术创新概念的综合体,是指开辟新的研究领域或者以新视角来重新认识已知事物并提出新观点的科技研究。科技创新包括新概念、新思想、新理论、新方法、新发现和新假设等原创性或者继承性的科技知识创新、科学创新发现、科学技术创

* 本文系国家社会科学基金项目“学术文献‘意抄’检测研究”(项目编号:12CTQ032)和山东理工大学人文社会科学发展基金研究成果之一。

[作者简介] 祝娜,山东理工大学科技信息研究所硕士研究生;王效岳,山东理工大学科技信息研究所教授,通讯作者,E-mail: wangxy1098@163.com;白如江,山东理工大学科技信息研究所讲师,馆员。

收稿日期:2014-04-21 修回日期:2014-06-19 本文起止页码:132-139 本文责任编辑:杜杏叶

新等,它们结合在一起组成科技创新,使科技创新知识系统不断丰富和完善。科技创新结构如图 1 所示:

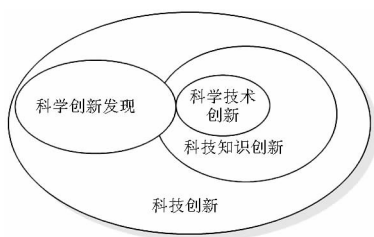


图 1 科技创新结构

2.2 科技创新路径

科技创新路径识别是以电子期刊、书籍或者专利数据为主要研究对象,综合使用引文分析、数据挖掘、信息处理、机器学习等计算机技术进行自动化、智能化的深入分析,得出某项科技创新的发展演化历史,帮助用户准确高效地获取科技创新相关的具体内容。本文将科技创新路径定义为将科技创新按照时间顺序串起来组合形成的,科技创新路径勾勒了某项科技从首次出现到不断创新发展的演化过程,揭示了科技创新知识的原理、结构、功能、性质、方法、过程等方面显著性变化的过程。

2009 年,“大数据时代的预言家”维克托·迈尔-舍恩伯格指出:大数据时代,信息分析与处理也将发生改变,即我们需要分析更多的数据,甚至需要处理和某个数据相关的所有数据^[3]。庞大的数据资源使得各个科学领域开始注重量化研究,对于科技创新路径识别研究而言,不光需要注重量化统计研究,更需要提供准确的科技创新内容以及它们之间的前后承接关系。科技创新路径识别不是简单随机分析几篇科技文献样本数据,也不是单独分析一篇科技文献的文本内容,而是把握全局、整合资源、全面分析某项科学领域全部的数据资源。因此,本文认为科技路径识别的主要研究内容应该聚焦在以下两个方面:

- 资源的有序整合。科技创新路径识别处理的文献信息资源不再是单一文献类型,而是一个庞大的包括众多零散、割裂信息的“数据海洋”,最后整合形成一条条融会贯通的科技创新路径。
- 面向科技创新内容的路径揭示。通过相关方法以及技术手段,深入分析科技文献文本内容,并充分识别科技创新点之间的承接关系。

3 科技创新路径识别数据源

3.1 学术文献数据库

目前进行科技创新路径识别的数据源分为文献全

文数据库和引文索引数据库。国内全文数据库主要有维普、万方、CNKI、超星、书生之家、中国数字图书馆,国外主要有 SpringerLink、IEEE/IEE、Engineering Village、ProQuest、EBSCO、Science Direct 等。这些在线全文数据库具有很高的学术权威性,提供期刊文献的全文、索引和文摘等信息下载,学科覆盖面广,专业性强,是进行科技创新路径识别的首选资源。国内引文索引数据库主要有中国科学引文数据库(CSCD)、中国科技论文与引文数据库(CSTPCD)、中国社会科学引文数据库(CSSCI);国外主要有工程索引(EI)、科学引文索引(SSCI)、科技回忆录索引(ISTP)、科学评论索引(ISR)等。

3.2 新兴网络学术数据源

陈悦等在《悄然兴起的科学知识图谱》一文^[4]中详细论述了目前传统数据源之外的数据库及其数据格式和存取内容;康奈尔大学的保罗·金斯帕格介绍了 <http://arXiv.org> 网站内部的细分学科,它可以从一个大的资源中摘录出新出现的研究领域;而 J. Kleinberg 则介绍了网址内容随时间呈现信息爆发流的规律,提出用网址数据分级用法来组织这种即时变化的信息;此外,还有 Google Scholar、Scirus、NLM、MEDLINE、LWW、HMC 等新兴的进行科学创新路径识别研究所需要的数据来源,arXiv、CiteSeerX、DBPL、ADS 等科技文献计量数据源以及 USPTO、EPTO、NSF 等专利数据源和基金数据源。

相比于传统学术文献数据库数据源,这些新兴数据源在科技创新路径研究中具有很强的学术影响性,其引文类型多样,引文覆盖面更广,包含学术期刊、图书、预印本库、机构库和非联机出版物等多种文献类型,内容涵盖自然科学、人文科学、社会科学等多个学科,可以弥补专业数据库数据源在费用、效率和速度等方面的劣势。

4 科技创新路径识别方法

1965 年,D. Price 提出“科学创新前沿”的概念^[5],定义了科学发展路径节点的创新知识点。此后,J. Law、J. Nicolaisen 等^[6-7]通过研究证明了科研人员进行科学写作时,需要参考前人研究成果来支持个人的科技创新,因此引文分析法被广泛运用于揭示某一科学知识领域在一定时期的科技发展趋势与动向,然而传统基于引用频次识别科技创新路径时,注重进行量化的引文分析法难以从复杂的引文网络中有效地识别出科技创新路径。在此背景下,为了识别出科技创新路

径中的科技创新内容,笔者提出基于关键词或主题词的引文分析以及基于引文全文的语义分析等识别科技创新路径的方法,以为面向科技创新的科技决策和科技评价提供重要支持。

4.1 基于关键词或主题词的识别方法

关键词或主题词是概括科技文献研究内容的规范化的、高度浓缩的文本,通过分析不同时间段关键词或主题词的演化,可以把握某一科技研究领域的科技创新路径。基于关键词或者主题词的识别就是从科技文献标题、摘要、关键词、文本中抽取关键词或者主题词,根据科技文献之间的引用关系,进行共词分析或聚类分析,以较为准确地表达概念之间的关系,将科技文献引用网络变为由科技知识单元直接联接的概念网,使科学创新知识体系从外部宏观结构的研究转变为内部微观内容的研究,并生成具有实际内容的科技创新路径。早期科研人员是从标题、摘要抽取关键词或主题词来识别科技创新路径,如表 1 所示:

表 1 基于关键词或主题词的科技创新路径识别方法

年份	研究者	具体内容
1964	E. Garfield ^[8]	提出科学引文索引(SCI),可以分析科学文献之间的引文关系,从而揭示科学技术发展的历史脉络
1978	H. G. Small ^[9]	提出将引用内容作为表达观点的概念符号,通过共被引聚类,利用主题词将引文标签化来识别科技创新路径
1993	Z. Chen 等 ^[10]	尝试在科技文献之间建立一种逻辑关系,将分散于科技文献中的科技创新知识整合起来,进而发现和创造新的科技知识
2003	陈超美 ^[11]	开发了 CiteSpace 软件,基于关键词或主题词识别科技研究领域的演变,并可可视化显示
2004	E. Garfield ^[12]	开发了 HistCite 软件,利用 WoS 数据来分析和判定某一研究领域的文献以及相互引用情况,把握其科技创新研究进展,产生可视化谱

为了进一步识别科技创新点,科研人员由从标题、摘要抽取关键词或主题词发展到从整个文本中抽取关键词和主题词。如侯跃芳等^[13]通过对高频科技文献的重要来源文献中关于妊娠糖尿病科技创新内容的主
题词分析来揭示其科技创新内容;通过运用引文共被引聚类分析方法,揭示了妊娠糖尿病的发展历史。祝清松等^[14]通过 HistCite 构建碳纳米管纤维领域被引文献集和施引文献集,基于规则的引文内容抽取技术从施引文献集中抽取被引文献集的引文内容,实验结果证明:在科技文献引文全文内容分析基础上进行关键词或主题词分析与基于标题或者摘要中的关键词或主题词识别科技创新主题相比,前者能够更加有效地揭示被引科技文献的科技创新内容。

基于关键词或主题词识别有效解决了单纯依赖引

用频次识别科技创新路径的缺陷,一定程度上可以揭示出科技创新路径中的科技创新内容。郑彦宁、化柏林^[15]比较了句子级知识抽取与词语级知识抽取的差异性,分析了句子级知识抽取在情报学中的意义,指出有些科技知识并不能用词或短语来完整地表达,句子是完整地表达科技知识信息的基本单位。但是基于关键词或主题识别科技创新路径时,科技创新动机识别不清晰,容易造成路径不完善、虚假路径链接等,因而对科技创新内容表达不具体,未能深入理解科技创新的语义内容,不能在语义级别把握科技文献中的创新知识。

4.2 基于引文全文语义分析的识别方法

基于引文全文语义分析的识别方法就是将引文分析法与自然语言处理技术结合起来,一方面进行定量分析,形成科技创新路径,另一方面进行定性分析,以揭示科技创新之间的语义承接关系,并对科技创新路径的关键节点进行语义解读,最后构建包含具体内容的科技创新演化路径。引文全文语义分析主要包括数据预处理、科技文献引文内容抽取和科技文献引文全文深度分析 3 个处理步骤,其目的是为了识别科技创新引用动机、科技文献中的科技创新点的语义内容和语义承接关系等。

在科技创新引用动机方面,自从 1964 年 E. Gafield 提出 15 种引用动机^[16]以来,涌现了大量关于识别科技创新引用类型的研究,但是它们并没有应用到实际的科技文献语义分析中来。2011 年杨思洛^[17]在研究中指出了传统引文分析理论(基础理论和引用动机)的不完善,即两篇论文引用同一篇早期文献,由于引用目的不同,一篇是引用其方法,另一篇则可能是引用其结果,从而揭示了基于的引文全文语义分析对识别科技创新路径的必要性。2013 年齐燕^[18]提出语义引用的概念,定义了科技文献的语义引用方向,将科技文献的语义引用表示为 $SC_i = \sum SC_{ij} (i, j = 1, 2 \cdots n)$; $SC = (C, RR, RA, A, F)$,最后生成了语义引文网络,见图 2。

图 2 中,语义引用 SC 设置为 n 元组,包含了引用过程中的科技创新知识内容 C、该创新知识与被引文献的关系 RR、该创新知识与施引文献的关系 RA、作者对科技创新知识的感情色彩 A、科技创新知识的作用 F。通过使用推理规则,可以识别科技创新内容以及科技创新的演化路径。

在引文语义分析方面,20 世纪 80 年代以后,在信息理解研讨会(MUC)、自动内容抽取(ACE)以及文本

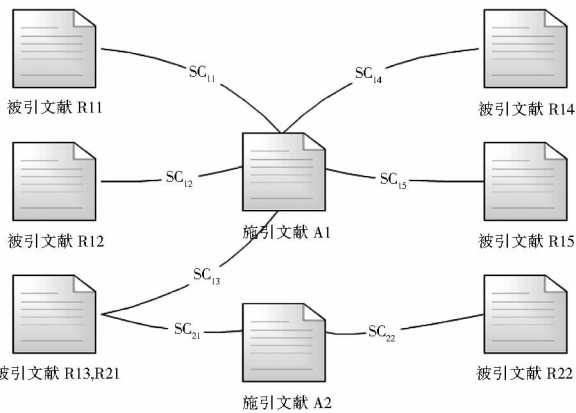


图 2 语义引文网络

分析(TAC)等会议的推动下,增强了文本语义分析能力。2009 年,赵火军、温有奎^[19]提出利用科学文献引文链提取文献特征句子和参考文献特征句子,表示成本体,准确表达科技文献中的科技创新知识单元并串联成路径。2012 年,贺德方等^[20]提出以引文链接为基础,结合主题链接、行为关联链接、本体链接等关联方式实现基于引文的科技创新知识链接。2013 年,高劲松^[21]等在分析传统引文知识链接模式的基础上,提出基于关联数据的引文知识链接模式,构建语义 Web 环境下的科技创新知识路径。同年,Ding. Y 等^[22]基于语法和语义两个层面(见表 2、表 3),提出一个二维的 CCA 理论框架来分析研究引文全文内容,更加详细地描述引用内容及引用关系,帮助准确识别引文内容之间的语义贡献,进一步促进科技创新点的语义关联识别。为了更加准确有效地识别科技创新路径,可以增强对文本内容的语义理解,将引用关系语义化。

表 2 CCA 语法层面理论框架^[22]

类别	具体内容
被引	文献类型 期刊论文、会议论文、书籍/书籍章节、报告、链接/个人博客、其他
	作者类型 一个作者、合著作者
	引用关系 自引、平行、等级
	引用位置 摘要、引言、文献综述、方法、结果/讨论、结论、其他
	文章出现次数 1 次、2-4 次、大于等于 5 次
	引用方式 没有特别提及、特别提及并解释说明、直接引用
施引	文献类型 期刊论文、会议论文、书籍/书籍章节、报告、链接/个人博客、其他
	作者类型 单一作者、合著作者

J. Swales 曾指出在引文研究中将逐渐产生两个相对独立的方向:定量地论述参考文献引用情况(以引用频次为主,分析大量参考文献的引用关系)和定性地解释在引用中的系统性含义(以引文内容分析为主,识别

表 3 CCA 语义层面理论框架^[22]

类别	具体内容
被引	引用作用 提供背景信息、构建理论框架、提供之前的经验/实验证据、描述面临的挑战和局限性
	引用感情色彩 积极、消极、复杂、中立
施引	研究领域类型 社会科学、人文科学、自然科学、应用科学和工程
	研究重点类型 理论研究、经验研究、实验研究、其他(特殊)

科技文献语义内容)^[23]。通过以上对目前科技创新路径识别方法的分析可以看出,传统引文网络分析方法与自然语言处理技术相结合的基于引文全文的语义分析将成为科技创新路径识别方法的未来发展趋势:一方面基于引文分析的方法和工具分析科技创新点之间的引用动机、引用关系;另一方面结合自然语义处理技术对科技文献引文全文进行语义分析,识别科技创新的具体内容。

5 科技创新路径识别可视化技术及工具

5.1 科技创新路径可视化工具

将科技创新路径可视化是为了直观地显示科技创新路径的演变过程,一般包括数据的采集、预处理、知识单元构建、数据分析、可视化与解读 5 个步骤,可视化科学研究领域的发展历史,预测科学创新的未来发展趋势。目前科技创新路径可视化工具主要有:Sci2^[24]、CiteSpace^[25]、HistCite^[26]以及 VOSviewer^[27]等,见表 4。

5.2 科技创新路径可视化图谱分析

科技创新路径可视化工具的功能是直观显示科技创新知识发展进程与结构关系的一系列不同的图形,科技创新路径可视化图谱分析如下:

5.2.1 Sci2 Sci2 绘制科学创新路径时,从时间、空间、主题、网络分析和可视化等多角度将数据进行分段处理,以直观地显示每个时间段最重要的关键词。Sci2 通过构建知识单元来可视化地形成科技创新关系网络,将不同级别的关键词通过不同大小的节点和连线形象地表示出来,让读者可以一目了然地看出其中的层级关系,构建科技创新演化路径。邱小花等专门研究了 Sci2 工具^[28]及其可视化的主要步骤,发现它是利用 Gnuplot 图以及可视化插件 GUESS 或 Cytoscape 来执行科技创新数据可视化操作,并构建情报学科技创新知识图谱^[29-30]。K. Börner 不断升级 Sci2 的分类、文本挖掘等功能,在语义理解方面也有研究,并绘制了医学、化学等领域的科技创新路径图谱^[31-34]。

5.2.2 CiteSpace CiteSpace 以树形图及连线等表示各个主题关系强弱,绘制共被引图谱、关键词图谱和时

表 4 4 种科技创新路径可视化工具简介

工具简介	Sci2	CiteSpace	HistCite	VOSviewer
开发设计者	Katy Börner	陈超美	Garfield	Van Eck 等
处理的数据格式	bibtex 格式、WOS 中的 TXT、xml、.net、.csv、.bib、.enw 等	WOS 中的 TXT 格式、软件转化的 CSSCI 格式	WOS 中的 TXT	WOS 中的 TXT 格式
数据预处理	去重处理,时间切片,数据和网络的缩减	时间切片,数据和网络的缩减	无	无
标准与简化	用户定义	Salton 余弦、Dice 强度和 Jaccard 强度	无	关联强度
分析方法	突发检测、地理空间分析、构建网络、时序分析	突发检测、地理空间分析、构建网络、时序分析	时序分析	社区检测、构建网络
研究结果	共词网络图、共被引网络图、耦合网络图等	科学计量分析、可视化图谱	引文编年图、引文矩阵	二维引文密度图谱、学科分布图、科研协作图等

区视图,图谱中点的颜色年轮表示不同时间段该文献的被引情况,节点的大小与被引用频次成正比,可以识别科技创新内容起始时间,每个时区条块同步生成匹配的专业术语,并通过关键词聚类 and 突变词探测来确定科技创新路径演化过程中的关键节点、研究热点和发展趋势,最终按照时间顺序得到可视化科学路径演进图。陈超美将科学创新定义为“潜在研究领域及发展趋势”^[34]、可视化“恐怖主义”以及“物种灭绝”的科学演进过程^[11]。CiteSpace 选择“keyword”作为节点并选定突变短语作为“termtype”,选择“detect bursts”,探测突变的科技专业术语,设定阈值选择 Pathfinder(关键路径算法)构建科技创新路径,在信息科学、生命科学、专利数据文献等领域得到广泛应用^[35-39]。

5.2.3 HistCite HistCite 按年代顺序显示科技创新点之间的前后承接关系,快速定位最新文献以及重要文献,分析各自科技研究领域的科技文献及其相互引用情况,展示某一领域不同科技文献之间的关系,将各自科学领域的研究进展按年代排列生成编年科学知识图谱。HistCite 的主路径分析是根据中心度计算出链接通度,画出具有最高中心度的节点来构建科技创新主路径,中心度测度针对科技文献引文网络,既考虑了一篇文献的被引次数(入度),又考虑了参考文献的数量(出度)。目前 HistCite 在领域可视化方面的研究较多^[40-42],例如,E. Garfield 分析研究了 DNA 分子结构研究领域、“小世界”等知识领域的科学引文编年图^[8],D. Lucio-Arias 等^[43]结合主路径和路径依赖两种方法,利用相对熵(公式 1)测量文本之间相似度,构建了富勒烯与富勒烯样结构纳米管科技创新的编年路径图。

$$I(q:p)=\sum_{i=1}^nq_i\log_2\left(\frac{q_i}{p_i}\right)\tag{1}$$

其中, $p_i=(p_1,p_2,\cdots,p_n)$ 代表参考文献的先验分布(在第一个文本中), $q_i=(q_1,q_2,\cdots,q_n)$ 是后验分布(在接下来考虑的文本中)。

5.2.4 VOSviewer VOSviewer 是基于 VOS 以及 Vx-Ord 可视化技术开发的,通过标签、密度、聚类密度、分散 4 个视图展示科技创新路径图谱。图谱中字体大小反映节点权重,节点间的连线数量可以在 0 至 ∞ 之间微调,通过不同的颜色、大小、年轮的不同厚度和视角等来揭示科技创新信息。简单密度视图中,项目密度表示科技创新项目之间的距离,即用最小化所有科技创新项目对的欧几里得的平方距离的加权和(公式 2)来表示科技创新点相似性,向量 x_i 代表创新项目 i 在二维科技创新知识图谱上的位置, s_{ij} 为项目 i 和项目 j 之间相似度的权重, $\parallel*\parallel$ 代表欧几里德距离。简单密度视图中点 x 项目密度为 $D(x)$ (公式 3), $K\in[0,\infty)$ 为核函数, $h>0$ 为调用内核宽度参数, w_i 为 i 的权重, d 为项目间的距离。通过图谱上点的项目密度可以计算出该点的颜色强度来控制科技创新知识密度,点 x 颜色强度为 $C(x)$ (公式 4), $\alpha>0$ 为色彩的变换参数, $D_{max}=\max D(x)$ 为最大密度项。N. J. VanEck 等利用 VOS 图谱技术可视化了计算智能领域的科技研究现状及发展历程^[44-45],国内赵勇、王晴、宗乾进等基于 VOS 方法进行了相关领域的可视化分析研究^[48-50]。

$$E(x_1,\cdots,x_n)=\sum_{i<j}s_{ij}\parallel x_i-x_j\parallel^2\tag{2}$$

$$D(x)=\sum_{i=1}^nw_iK\left(\frac{\parallel x-x_i\parallel}{dh}\right)\tag{3}$$

$$C(x)=1-\left(1-\left(\frac{D(x)}{D_{max}}\right)^\alpha\right)^{1/\alpha}\tag{4}$$

5.3 可视化工具比较评估

由于对海量科技文献全文的语义分析技术尚未发展成熟,科技创新路径可视化工具对科技创新具体语义内容无法进行详细描述,也无法反映科技创新的语义承接类型。通过梳理相关文献发现,国内使用 Sci2 和 VOSviewer 的研究较少,CiteSpace 和 HistCite 应用较多;另外,Sci2 的数据处理分析能力强;CiteSpace 具有突变词探测功能;HistCite 操作简单,运行迅速;VOSviewer

具有文本挖掘功能,可视化效果好。它们各有特点经过 对比分析,发现它们各自的优缺点如表 5 所示:

表 5 科技创新路径可视化工具比较

可视化工具	优点	缺点
Sci2	可以加载各种格式科技创新数据;支持多种网络插件;能构建多种矩阵,具有强大的数据分析统计和数据处理能力	运行对电脑内存要求较高,进行大量科技文献数据分析时容易出现死机现象 ^[29] ;处理中文数据,必须进行转换
CiteSpace	具有突变词探测功能;可以导入 CSSCI 的数据;可以构建共现矩阵,进行文献共被引分析	缺少相应的规范和标准来评价其科技创新可视化结果;针对科技文献引文的原则、规范、标准等方面没有自动处理功能
HistCite	可以直接导入 SCI 文献,而不用进行数据处理;操作方法简单易学;可以迅速分析出数据的最基本的信息	只能识别 WOS 中的数据;不具备共被引分析功能;阈值选择具有一定的随机性
VOSviewer	可读性好,视觉效果好;具有文本挖掘功能;应用范围广泛,可信度较高;图形展示能力强,适合大规模样本数据 ^[49]	数据源较多时 VOSviewer 无法响应;不能从文献数据中抽取和构建共现矩阵;数据预处理需要外部程序执行

6 科技创新路径识别面临的挑战和发展趋势

6.1 面临的挑战

6.1.1 科技创新类型标引不清晰 引用动机反映了科技创新的动机,引用动机研究起源于 E. Garfield 提出的 15 种引用原因^[16]。传统引文分析方法识别科技创新路径时,将引文内容同等对待,不能从本质上揭示科技创新点之间的语义关联;对引文的上下文语义内容未进行深入分析,造成科技创新潜在语义关联识别模糊。科技创新类型识别与引文类型识别紧密关联,需要从科技创新的引用动机和引用内容出发进行语义分析,揭示出科技创新点之间的语义关联,分析被引文献对施引文献的语义贡献,准确标引科技创新的语义类型,弥补传统科技创新路径识别方法不足之处。

6.1.2 科技创新点的具体语义内容需进一步识别 利用关键词或者主题词代表科技创新文本内容特征,忽略了对科技文献具体内容的语义理解,不能深入揭示科技创新路径关键节点及科技创新的具体语义内容,对于与科技创新内容相关的信息也没有进行详细的语义分析研究。目前基于引文全文语义分析来识别科技创新路径的研究已经出现,但仍需进一步深入研究,而要识别科技创新具体语义内容,就必须对科技文献篇章结构分布和语篇内容进行全面语义分析,标引出科技创新的语义内容和语义承接关系,进而构建科技创新的语义路径。

6.1.3 科技创新路径识别研究结果的评价有待规范 目前存在的规则库和训练语料库数量有限,大部分科技创新路径识别研究是采用人工或小规模样本标注进行实验研究,没有统一的评价标准和测评语料,以供对科技创新路径识别结果进行统一的测试和评价。另外,科技创新路径识别的研究结果都仅是通过与之前科技创新的识别方法进行比较,因此构建科技创新路径识别的

评价标准以及测评语料库迫在眉睫。

6.1.4 科技创新路径识别研究结果的应用水平有待提高 科技创新路径识别研究的目的是揭示科技创新发展的具体演变过程并用来支持科技决策,为科研评价提供依据。目前单纯依赖科技文献被引频次计算的影响因子进行评价,科研评价结果具有片面性,因此必须实现科技创新内容数量和质量测度的结合,才能揭示科技创新的真正价值,达到完善科研评价体系的目的。另外,科技创新路径需要构建一个完善的识别系统模型为广大科研人员提供知识服务,应该识别科技创新具体语义内容,实现数量和质量识别的有机结合,才能为科技决策和科研评价提供借鉴和支持。

6.2 未来发展趋势

6.2.1 研究数据集来源更广,面向多语言数据处理 目前科技创新路径识别研究涉及的数据源一般是对 WoS 数据集进行处理,也开始使用其他数据集(基金数据、专利数据、网络出版数据等),多为引文索引数据。随着 Google Scholar 以及 Citebase Search 等科技文献搜索引擎的应用,一些科学博客、学者主页等包含科技创新内容的专家观点网页数据集都会变成科技创新路径识别的数据来源。目前研究数据主要针对英文,针对中文以及其他语言的科技文献或者多种语言的混合科技文献数据也将是未来科技创新路径识别的研究对象。

6.2.2 引文分析法与语义分析方法相结合 这是未来科技创新路径识别的发展趋势。科技创新路径从最早的基于引用频次发展到基于关键词或主题词、基于引文全文语义分析的识别方法,其研究目的一方面是为了利用科技文献之间的引用关系挖掘出科技创新传承路径,利用语义分析方法识别科技创新类型,判别路径前后承接关系,剔除虚假链接,生成科技创新路径;另一方面,句子是表达语义含义的基本单位^[15],科技创新路径识别需要对科技文献文本进行句子级别的语义分析,对科技创新语义内容进行理解、总结、归类、

推送,从而提高科技创新路径识别的准确率和召回率。

6.2.3 技术融合 在研究技术上,将融合自然语言处理技术、文本数据挖掘技术以及计算机处理技术,增强对文本内容的语义理解,深入分析科技创新内容之间的语义承接关系。目前针对单一文献进行的科技创新内容识别和标引已经发展成熟,利用自然语言处理技术(比如语义角色标注 SRL 等)对海量科技文献数据集中的科技创新研究方法、过程或相关研究内容进行深入的语义标引与识别,将会更加有利于准确有效地揭示科技创新内容之间的语义关联的具体内容和强度大小,便于充分解读科技创新路径。目前科技创新路径可视化工具(如 HistCite、CiteSpace 等)几乎都是基于关键词或主题词的方法将文本进行切割,例如将“科技创新路径”切分为“科技创新”和“路径”,失去了其原来要表达的完整语义,造成语义缺失,所以应该发展可视化技术,增加其对文本的语义挖掘功能,将研究结果形象、全面、直观地呈现出来。

6.2.4 对科技创新路径全面可视化语义解读 研究结果上从单一的科技创新关键路径可视化解读转变成对科技创新路径全面可视化语义解读,即注重对科技文献内容的语义研究,构建训练数据集,基于领域语料库,利用机器学习的方法,识别所有的科技创新点及其相关内容,其中包括:研究目的、研究方法、实验设备、研究结果等。另外,科技创新路径识别结果可以是某一领域整体复杂的路径网络,也可以是某项科技研究某一方面科技创新研究路径。比如“语义角色标注”,可以单独识别它的“方法”、“理论”、“工具”等科技创新演化路径,克服传统科技创新关键路径识别或者主路径识别的片面性,解决科技创新各类型内容混淆不清的问题。

参考文献:

- [1] 伯纳德·巴伯. 科学与社会秩序[J]. 上海:上海三联书店,1991.
- [2] He Chuanqi. National knowledge innovation systems: Evolution, definition and indicators[J]. 中国科学院院刊(英文版),1999(4):244-247.
- [3] Tsai W H, Tsaur T S, Chou Y W, et al. Evaluating the information systems success of ERP implementation in Taiwan's industries [C]//Proceedings of IEEE International Conference on Digital Object Identifier. Hong Kong:IEEE,2009:1815-1819.
- [4] 陈悦,刘则渊. 悄然兴起的科学知识图谱[J]. 科学学研究,2005(2):149-154.
- [5] Price D. Statistical studies of networks of scientific papers[C]//Proceedings of Statistical Association Methods for Mechanized Documentation Symposium. Washington: Government Printing Office,1965.

- [6] Law J, Bauin S, Courtial J P, et al. Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification[J]. Scientometrics, 1988, 14(3): 251-264.
- [7] Nicolaisen J. Citation analysis[J]. Annual Review of Information Science and Technology, 2007, 41(1): 609-641.
- [8] Garfield E, Sher I H, Torpie R J. The use of citation data in writing the history of science[M]. Philadelphia: Institute for Scientific Information, 1964.
- [9] Small H G. Cited documents as concept symbols[J]. Social Studies of Science, 1978, 8(3): 327-340.
- [10] Chen Z. Let documents talk to each other: A computer model for connection of short documents[J]. Journal of Documentation, 1993, 49(1): 44-54.
- [11] Chen Chaomei. Searching for intellectual turning points: Progressive knowledge domain visualization[J]. Proceedings of the National Academy of Sciences, 2004, 101(suppl 1): 5303-5310.
- [12] Garfield E. Historiographic mapping of knowledge domains literature[J]. Journal of Information Science, 2004, 30(2): 119-145.
- [13] 侯跃芳,崔雷,吴迪. 应用引文共引聚类——内容词分析法对学科发展的研究[J]. 情报学报,2007,26(2):309-314.
- [14] 祝清松,冷伏海. 基于引文内容分析的高被引论文主题识别研究[J]. 中国图书馆学报,2014(1):39-49.
- [15] 郑彦宁,化柏林. 句子级知识抽取在情报学中的应用分析[J]. 情报理论与实践,2011(12):1-4.
- [16] Garfield E. Can citation indexing be automated[C]//Proceedings of Statistical Association Methods for Mechanized Documentation, Symposium. Washington: National Bureau of Standards Miscellaneous Publication, 1965: 189-192.
- [17] 杨思洛. 引文分析存在的问题及其原因探究[J]. 中国图书馆学报,2011(3):108-117.
- [18] 齐燕. 引用语义化相关问题初探[J]. 情报理论与实践,2013(8):15-20.
- [19] 赵火军,温有奎. 基于引文链的知识元挖掘研究[J]. 情报杂志,2009(3):148-150.
- [20] 贺德方,曾建勋. 基于语义的馆藏资源深度聚合研究[J]. 中国图书馆学报,2012(4):79-87.
- [21] 高劲松,梁艳琪,马倩倩,等. 面向关联数据的引文知识链接模式研究[J]. 现代图书情报技术,2013(3):21-26.
- [22] Zhang Guo, Ding Ying, Milojević S. Citation content analysis: A framework for syntactic and semantic analysis of citation content[J]. Journal of the American Society for Information Science and Technology, 2013,64(7):1490-1503.
- [23] Swales J. Citation analysis and discourse analysis[J]. Applied Linguistics, 1986, 7(1): 39-56.
- [24] Sci2Tool[EB/OL]. [2014-04-19]. <http://Sci2.cns.iu.edu/user/index.php>.
- [25] CiteSpace[EB/OL]. [2014-04-19]. <http://cluster.cis.drexel.edu/~cchen/citespace/>.
- [26] HistCite[EB/OL]. [2014-04-19]. <http://interest.science>.

thomsonreuters.com/forms/HistCite/.

- [27] VOSviewer[EB/OL]. [2014-04-19]. <http://www.vosviewer.com/>.
- [28] 邱小花, 李国俊, 肖明. Sci2——一款新的知识图谱分析软件介绍与评价[J]. 图书馆杂志, 2013(9): 79-87.
- [29] 李国俊, 刘恩涛, 邱小花, 等. 国外图书馆学情报学知识图谱研究——基于 Sci-2 的文献引文耦合分析[J]. 情报杂志, 2014(1): 78-90, 151.
- [30] 邱小花, 李国俊, 肖明. 基于 Sci-2 的国外图书馆学情报学研究主题演变分析——以共词分析为例[J]. 情报杂志, 2013(12): 110-118.
- [31] Börner K, Scharnhorst A. Visual conceptualizations and models of science[J]. Journal of Informetrics, 2009, 3(3): 161-172.
- [32] Boyack K W, Börner K, Klavans R. Mapping the structure and evolution of chemistry research[J]. Scientometrics, 2009, 79(1): 45-60.
- [33] Skupin A, Biberstine J R, Börner K. Visualizing the topical structure of the medical sciences: A self-organizing map approach[J]. PloS ONE, 2013, 8(3): e58779.
- [34] 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.
- [35] 陈超美, 陈悦, 侯剑华, 等. 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.
- [36] Morar M, Agachi P S. Review: Important contributions in development and improvement of the heat integration techniques[J]. Computers & Chemical Engineering, 2010, 34(8): 1171-1179.
- [37] Chen Kaihua, Guan Jiancheng. A bibliometric investigation of research performance in emerging nanobiopharmaceuticals[J]. Journal of Informetrics, 2011, 5(2): 233-247.
- [38] 邱均平, 温芳芳. 近五年来图书情报学研究热点与前沿的可视化分析——基于 13 种高影响力外文源刊的计量研究[J]. 中国图书馆学报, 2011(2): 51-60.
- [39] 张士靖, 杜建, 周志超. 信息素养领域演进路径、研究热点与前沿的可视化分析[J]. 大学图书馆学报, 2010(5): 101-106.
- [40] 董立平, 郭继军. 利用 Histcite 的人胚胎干细胞引文编年图主要路径分析[J]. 医学信息学杂志, 2010(11): 38-40, 49.
- [41] 闫雷, 关晶, 崔雷. 基于 HistCite 的抗疟药研究相关文献引文编年图和主要路径[J]. 医学信息学杂志, 2012(9): 51-54.
- [42] 刘丽. 基于 HistCite 的图书馆服务领域引文脉络分析[J]. 情报科学, 2014(5): 91-96, 101.
- [43] Lucio-Arias D, Leydesdorff L. Main-path analysis and path-dependent transitions in HistCite™-based historiograms[J]. Journal of the American Society for Information Science and Technology, 2008, 59(12): 1948-1962.
- [44] Van Eck N J, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping[J]. Scientometrics, 2010, 84(2): 523-538.
- [45] Van Eck N J, Waltman L. Text mining and visualization using VOSviewer[J]. arXiv preprint arXiv:1109.2058, 2011.
- [46] 赵勇, 李晨英, 韩明杰. 基于 VOS 方法的农业经济学研究主题的可视化分析[J]. 中国农业大学学报(社会科学版), 2012(3): 118-125.
- [47] 王晴. 基于 SSCI 的 2012 年国外档案学研究特征分布与热点探析[J]. 档案与建设, 2013(8): 14-18.
- [48] 宗乾进, 袁勤俭, 沈洪洲. 基于 VOSviewer 的 2010 年中国图书馆学研究热点分析[J]. 图书馆, 2012(4): 88-90.
- [49] 廖胜蛟. 科学知识图谱绘制工具 VOSviewer 与 Citespace 的比较研究[J]. 科技情报开发与经济, 2011, 21(7): 137-139.

Research Status and Development Trend of Science and Technology Innovation

Path Recognition: Methods and Tools

Zhu Na Wang Xiaoyue Bai Rujiang

Institute of Scientific & Technical Information, Shandong University of Technology, Zibo 255049

[Abstract] In recent years, science and technology innovation path identification has become a hot topic of Informatics in order to support scientific and technological knowledge innovation. This paper first reviews data sources, research methods and visualization tools in the field of recognition of science and technology innovation path. Then the authors focus on two main technological innovation path identification methods. One is based on keywords or subject words and the other is based on semantic analysis of full-text citations. Comparative analysis of visualization tools such as Sci2, HistCite, CiteSpace and VOSviewer are conducted. At last the authors point out the challenge of science and technology innovation path recognition based on citation analysis trends on data collections, research methods and research outcomes. This paper aims to provide support and assistance to the recognition of science and technology innovation study.

[Keywords] science and technology innovation science and technology innovation path citation analysis visualization