

## 科学家学术谱系的内涵、构建与测度研究述评\*

■ 盛怡瑾<sup>1,2</sup> 赵勇<sup>1,2</sup><sup>1</sup> 中国农业大学图书馆 北京 100083<sup>2</sup> 中国农业大学情报研究中心 北京 100193

**摘 要:** [目的/意义] 学术谱系与知识传承密切相关, 具有历时性、系统性和评价功能, 蕴含大量可挖掘的信息和价值。系统梳理科学家学术谱系的相关研究, 揭示学术谱系的价值及潜力, 为该主题研究及实践发展提供参考。[方法/过程] 对学术谱系的内涵进行辨析, 从数据采集、导学关系识别、结果可视化等方面总结学术谱系的构建方法, 并将目前学术谱系的测度研究按照测度指标及方法和测度应用两个方面进行分析。[结果/结论] 学术谱系在包括情报学在内的多个研究领域具有重要研究价值及潜力, 未来应重点关注数据来源、平台建设和主题拓展方面的问题。

**关键词:** 学术谱系 谱系内涵 谱系构建 谱系测度

**分类号:** G250

**DOI:** 10.13266/j.issn.0252-3116.2023.14.011

## 1 引言

科技创新活动中存在“优势积累”现象, 师从名师对科技人才成长具有重要影响。朱克曼在《科学界的精英——美国的诺贝尔奖金获得者》<sup>[1]</sup>中最早提出了“科学上的师承关系”。书中提到, 1972 年以前在美国进行其获奖研究的 92 位诺贝尔奖获得者中, 一半以上曾是前辈诺贝尔奖获得者的学生。朱克曼发现, 师傅是徒弟的榜样、是徒弟杰出成就的诱发者、也是徒弟科学工作的严格批评者, 这对于培养高水平徒弟至关重要。高徒在未来也会成为名师, 一个学者的生命是有限的, 但他的贡献会通过一代又一代学生而放大、增强和延伸<sup>[2]</sup>, 形成了绵延不断的学术谱系, 推动了科学的蓬勃演进。从这个意义上说, 学术谱系实际上是学术史的另一种表现形式<sup>[3]</sup>, 蕴含着大量可挖掘的信息和价值。国外相关研究和实践起步较早, 可追溯到 20 世纪 30 年代科学家对自己学术谱系的书写<sup>[4]</sup>, 现已陆续建成一批成熟的学术谱系数据库; 2010 年 5 月起, 中国科协先后在数学、物理、化学、天文学等学科领域启动当代中国科学家学术谱系研究, 并编成了“当代中国科学家学术谱系丛书”<sup>[5-6]</sup>。2014 年前后, 学界对学术谱系的研究进入新阶段, 有学者开始将学术谱系作为研究内容和研究工具进

行量化研究, 学术谱系的功能和应用场景得到进一步拓展, 图书情报学、计算机科学与技术、科学技术史、教育学领域的学者都对其展现出了浓厚兴趣。

学术谱系与知识传承密切相关, 作为跨学科研究领域, 本身具有历时性和系统性特征以及评价功能, 因此正在引起学界尤其是国内学者更大的关注。全面系统地梳理学术谱系相关研究, 对了解学术谱系本质、把握研究进展和发现研究问题具有重要的理论及实践意义。本文以中国知网、万方数据、Web of Science 和 Emerald 管理学数据库中的论文作为主要数据来源, 使用中文主题词“学术谱系”“师承关系”“导学关系”和英文主题词“academic genealogy”“genealogy”“advisor-advisee”“mentor-mentee”“protege”进行检索, 查看检索结果, 剔除无关文献, 如仅讨论知识转移而不关注“师生关系”的文献。同时, 根据引文索引策略圈定已有文章中引用的重要文献, 并配合网络搜索发现, 对结果进行补充, 共得到相关文献 51 篇, 包括期刊论文、学位论文、会议论文和专著书籍。通过逐篇研读, 本文从学术谱系的内涵、构建、测度及应用这几个方面依次展开, 按照“理论研究→方法研究→应用研究”的逻辑顺序梳理学术谱系相关内容, 以期充分挖掘学术谱系研究的价值, 支持学界的进一步拓展和创新工作。

\* 本文系国家社会科学基金一般项目“学术谱系视角下科学家的知识传承及贡献评价研究”(项目编号: 22BTQ099)研究成果之一。

作者简介: 盛怡瑾, 副研究馆员, 博士; 赵勇, 研究馆员, 博士, 通信作者, E-mail: zhaoyong@cau.edu.cn。

收稿日期: 2022-12-14 修回日期: 2023-04-12 本文起止页码: 109-118 本文责任编辑: 杜杏叶

## 2 科学家学术谱系的内涵

目前,学界对学术谱系尚未形成统一的定义。《韦氏词典》中对“谱系”的定义是“对个人、家族或群体血统的描述”或“对事物起源或历史发展的描述”<sup>[7]</sup>,《应用汉语词典》对“谱系”的定义为“家族间的遗传系统”或“物种变化的系统”<sup>[8]</sup>,从社会学角度看,谱系主要是宗族世系或具有同一来源的同类事物的历代传承,而在进化论框架下,谱系是指物种变化和延续的系统<sup>[9]</sup>。最常见的谱系就是家谱,学术谱系的形式与家谱类似,但是谱系上的人物之间不再以血缘关系相互关联,而是以师承关系进行连接。

### 2.1 概念界定

国内外学界从不同学科视角对学术谱系进行了概念界定<sup>[4]</sup>,有学者认为科学家的学术谱系是学术“家谱”,反映一个学科或学术群体中主要成员的学缘关系和传承关系<sup>[3]</sup>。也有学者认为,学术谱系是指由学术传承关系(以实质性的师承关系为主)关联在一起的、不同代际的科学家所组成的、动态发展的、开放的学术群体,在深层意义上,学术谱系是学科学术联合体的重要组成单元,是各种各具特色的学术传统或亚学术传统的载体<sup>[10]</sup>。还有学者认为,学术谱系指在一定专业研究领域内的知识和技能的历代传承关系<sup>[9]</sup>。2014年,美国学者C. R. Sugimoto首次将学术谱系定义为“一项定量研究”<sup>[2]</sup>,她认为,学术谱系是通过导师与学生构成的链条,对知识传承开展的定量研究。在她看来,学术谱系已经不再只是学者用于追根溯源的专属工具,它还具有更多功能和价值,主要包括5种用途:分别是纪念型(honorific)、自我型(egotistical)、历史型(historical)、范式型(paradigmatic)与分析型(analytic)。

与“学术谱系”相近的概念主要有“师徒关系”“师承关系”和“学术传承”。其中,“师徒关系”是学术谱系的核心要素,学术谱系由一代一代的师徒关系构成基本链条;“师承关系”较“师徒关系”加入了“继承”的意味,学术谱系不仅包含一代代师徒间的知识继承,还包括学术观点的变异、颠覆和创新;“学术传承”是学术谱系的重要组成部分,但不是学术谱系独有的要素,非师生之间也可以发生学术传承。

### 2.2 概念剖析

以上定义显现的差异表明学界对学术谱系的理解源于不同视角,这些视角在一定程度上暗含了定义者对于学术谱系某种用途的认可,体现了学术谱系的跨学科性和多用途性。综合以上定义可以看出,

学术谱系具备3个要素,分别是师徒关系、历代延续和知识转移。首先,师徒关系是学术谱系的核心要素,是学术谱系中独立的个体相互链接的依据,也是学术谱系区别于其他类型谱系的根本原因。其次,学术谱系由不同代际的科学家组成,强调师徒关系的代代延续,勾勒了各代师徒关系发展变化的清晰脉络。最后,学术谱系中必然存在知识转移。显性和隐性知识在导师与学生的互动中发生传递和继承,这一点很容易得到经验和数据上的证明,而且转移的知识不仅限于学科知识,还包括导师的治学态度、社会经验、学生培养模式等。师徒关系是学术谱系的明线,知识转移是学术谱系的暗线,二者相辅相成。

此外,学术谱系还具有一个重要特征,即系统性。学术谱系符合系统的定义,也具有系统的特性,如集合性、相关性、目的性、层次性、环境适应性和动态性。学术谱系由历代师徒构成,师徒之间以导学关系相互作用,实现了多用途性;历代师徒关系呈现出明显的层次性,随外部环境变化而发展和延续,并随时间发生着生长、分化、聚合、衰落等动态变化。因此,本文认为,学术谱系是以师徒关系的历代传承为连接进行知识转移的系统。

## 3 科学家学术谱系构建

学术谱系测度和使用的前提是谱系的构建,C. R. Sugimoto认为,学术谱系的构建方法包括初始化搜索、操作化联系、确定数据源和可视化结果4个关键环节<sup>[2]</sup>。W. Dorés等<sup>[11]</sup>构建学术谱系的流程主要是:收集数据、抽取特定字段并标准化和姓名消歧。Madeira等<sup>[12-13]</sup>开发了名为The Gold Tree系统,从多个来源提取和集成元数据以创建学术谱系树,并对谱系树进行可视化。该系统使用的构建流程是:数据源定义、数据收集与预处理、数据建模及标引、网络信息系统构建。综上,构建学术谱系一般要经历的步骤有:确定数据源、构建数据集、数据清洗、进行连接和可视化(见图1)。

确定数据源是构建学术谱系时面临的首要问题。在实际中,为确保数据全面性,构建者需从多个数据源收集数据,会用到网络爬虫、信息抽取、访谈等方法。多源数据集要在格式转换及去重后进行合并,并进行数据清洗,这关系到谱系构建的准确度。数据清洗需要关注的问题是:姓名消歧、字段值标准化及异常值处理等。在进行链接或建模时,要确定清晰明确的标准,保证构建逻辑的合理和一致性。

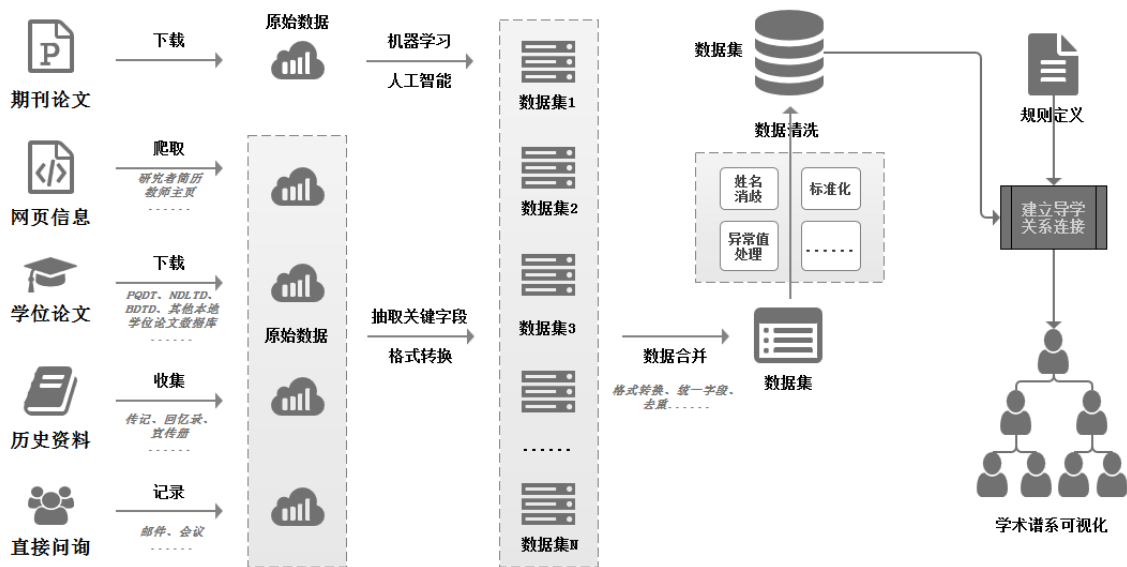


图 1 学术谱系构建流程

### 3.1 数据采集

构建学术谱系时面临的重点及难点问题是收集数据。学术谱系包含了同一根基下多代学者的多代师生关系, 这些数据年代跨度大, 涉及机构多, 不会被有意识地公开, 而且其中一些已经无法追踪, 给学术谱系的构建带来了挑战。目前, 构建学术谱系常用的数据来源有以下 4 类:

#### 3.1.1 直接问询

获取学者的师承关系和师门人员构成信息, 直接问询当事人或知情人是最简单的做法。直接问询的实现方式有两类, 一类是面对面交流, 如访谈、会议交流等; 另一类是非面对面交流, 如通过电话、传真、信件、电子邮件、网络社交媒体等。E. A. Kelley 等<sup>[14]</sup>在编纂美国野外灵长类动物学家的学术谱系时, 采用的三个数据来源中两个都是直接问询, 分别是电子邮件调查和在美国体质人类学协会 (American Association of Physical Anthropology) 会议等论坛进行口头交流。直接问询得到的数据比较准确, 而且能挖掘到公开渠道难以获取的信息, 但缺点是被询问者响应度不高, 调查难度大。如数学领域谱系项目 (MGP) 创始人 Harry Coonce 在 1996 年给几百个数学系写信问询博士姓名、论文题目以及导师姓名时, 收到的回应不到 30%。在今天, 直接问询法也存在类似的问题, 学者或知情人在不了解谱系构建者目的和意图时, 容易因隐私保护而存有警惕心。

#### 3.1.2 网页信息

互联网时代, 导师及其学生的部分信息也可能公布在网页上, 目前学术谱系研究中常用的网页

信息是研究者简历, 其中尤以巴西的 Lattes 平台为代表<sup>[15-16]</sup>。Lattes 平台由巴西科学技术发展委员会 (CNPq) 主管<sup>[17]</sup>, 包括履历表、机构名录、研究团队名录和展示分析板块 4 部分, 履历表中包含人员介绍、工作单位、研究方向、教学内容、科研项目、期刊 (会议) 文章、指导学生等信息。该平台强制参加研究生课程及请求财政支持的研究者提供简历, 实现了数据的实时更新和开放获取, 并能有效控制数据质量。截至 2013 年底, 该平台履历数量已达 276.5 万份<sup>[18]</sup>。Lattes 平台的履历信息能用于学术谱系研究的原因是: ①记录并公开师承关系数据, 大多研究者简历平台并不设置师承关系字段, 因此无法用于学术谱系构建; ②规模大, 经过 20 多年发展, Lattes 平台积累了大量履历, 能从中提炼出师承关系的复杂网络; ③数据可靠, Lattes 平台建设之初, 科技部等部门的下属基金委将履历填写与 CNPq 基金项目申请资格挂钩, 进行了一定的强制推广, 随着系统不断强大, 研究人员已对其产生了粘性, 填写和更新都更为及时, 平台后端已实现电子认证等网络安全措施, 也使用了数据挖掘技术识别和过滤不实信息<sup>[18]</sup>。其他网页如新闻报道、学校公布的入学名单等也可能包含师承信息, 但因其规模不足、数据静态及质量参差等问题, 只能作为补充或印证数据。

#### 3.1.3 学位论文

学位论文封面会注明学生及导师姓名且信息真实可靠, 也是构建学术谱系的重要数据来源。目前, 谱系构建时常用的学位论文数据库有 Proquest 硕士论文数据库 (ProQuest Dissertations & Theses Glob-



al, PQDT)、网络学位论文数字图书馆(Networked Digital Library of Theses and Dissertations, NDLTD)<sup>[11]</sup>、巴西学位论文数字图书馆(Brazilian Digital Library of Theses and Dissertations, BDTD)<sup>[12]</sup>。PQDT是全球最大的多学科博硕士论文精选数据库,收录了来自欧美、加拿大等100多个国家顶级研究机构的500万篇学位论文,并以每年20万篇的速度增长,被美国国会图书馆指定为官方论文库<sup>[19]</sup>。C. R. Sugimoto对该数据来源十分青睐,认为该库的优点是可靠性强,但不足在于,只有近几十年的论文才包含导师信息,这些信息的提取需要人工操作。NDLTD是一个致力于促进电子论文(ETD)采用、创建、使用、传播和保存的国际组织<sup>[20]</sup>,支持电子出版和开放获取,以增强全球知识共享。NDLTD由来自世界各地数百个学术机构的电子论文(ETD)集合组成。其存储库通过OAI-PMH协议从其他来源收集单个ETD记录。BDTD由巴西科学技术信息研究所(IBICT)开发和管理<sup>[21]</sup>,整合了巴西教学和研究机构现有的学位论文信息系统,鼓励以电子形式注册和出版学位论文。目前,该平台包含127个机构,513 715篇博士论文和195 176篇硕士论文。

### 3.1.4 历史资料

学术谱系具有历史性,因此在历史资料中也能获得师承信息。领域发展史和机构历史记录着领域和机构发展中的关键人物,若人物之间存在关联,会被当作重要线索提及。名人传记、回忆录和纪念册等具有重现、传承和见证等价值,也会记录一些师承信息及细节。历史资料一般存于图书馆、档案馆、机构资料室等地,优点是可靠性高,拥有其他渠道难以获知的信息和细节,不足是这些资料中师承信息往往不是主角,相对零散、随机,无法快速定位,信息获取效率很低,而且也缺乏动态性。

### 3.2 导学关系识别

从传统渠道获取师承关系数据不仅工作量大,还面临大量信息遗漏和缺失的问题,因此,学者开始探索以期刊论文为数据来源自动识别导学关系的方法。大多数学生都会与导师合作发表期刊论文,期刊论文中蕴含了大量可挖掘的师承关系信息。而且相比学位论文,期刊论文体量较大、容易获取、包含更多信息,能够最大限度地扩展谱系范围。但过去没有将其作为数据来源的主要原因是,期刊论文只记录作者姓名和机构,不直接体现师承关系。随着技术的发展,学者尝试引入数据挖掘、深度学习等智能算法识别期

刊论文中的师徒关系,取得了一定的进展。C. Wang等<sup>[22]</sup>提出了一种时间约束概率因子图模型(TPFG),该模型以出版网络为输入,使用联合似然目标函数对导学关系挖掘问题建模,其还设计了学习算法优化目标函数,该模型可为作者匹配几位可能的导师,并对导师人选按照概率大小排名。李勇军等<sup>[23]</sup>根据学生与导师共同署名现象,提出基于最大熵模型的导师—学生关系识别算法。W. Wang等<sup>[24]</sup>提出了一种基于深度学习的导学关系识别方法,名为Shifu,该方法同时考虑了本地属性和网络特征。导学关系识别方法能够大大减轻人力,是未来发展的趋势,但目前在准确性上还有待进一步提高和验证。Z. Zhao等<sup>[25]</sup>提出了一种时间感知的Advisor-advisee关系挖掘模型(tARMM),这是一个深度模型,配备了改进的更新门循环单元(Refresh Gate Recurrent Units, RGRU),能更好识别导学关系。Y. Gao等<sup>[26]</sup>认为以上方法中合作网络都是静态的,因此引入了动态网络,用有监督的机器学习方法识别导学关系。虽然自动识别方法有种种好处,但准确率较其他数据源还有一定差距,在实际识别中可能出现错误,需要一定的人工辅助。

### 3.3 结果可视化

学术谱系构建流程中的最后一步是进行结果可视化,可视化的目的是为了更好呈现谱系中包含的大量关系,方便查询和使用。学术谱系的结果呈现方式包括文字叙述、表格和图。文字叙述可以包含更多信息,但不直观,容易淡化和模糊学术谱系中存在的连接;表格相较文字更为结构化,能包含充足信息,也能做到代际之间链接跳转,但对于多代关系不直观;图可以清楚地呈现谱系规模及演变脉络,有利于在较大时间跨度内发现规律,但图的构建和呈现需要较好的支撑技术。

目前,国外已经建成许多不同规模的学术谱系数据库,其中Mathematics Genealogy Project(MGP)是数学领域知名的学术谱系数据库,也是最大的单学科谱系库,记录全球数学领域博士及其导师信息<sup>[27]</sup>。至2021年8月,该数据库中已有超过27万条记录<sup>[28]</sup>,是很多学者研究学术谱系的数据来源<sup>[29]</sup>。Neurotree是神经科学领域的学术谱系在线数据库,建立初衷是想通过谱系了解高度跨学科性的神经科学领域<sup>[30]</sup>。Neurotree中的数据由用户提供,可公开编辑,因此准确性难以保证,但会有报告系统和志愿编辑对内容进行适当把关。Academic Family Tree(AFT)是由

用户内容驱动的网络数据库,旨在准确记录和公开共享学术界所有领域研究人员的学术谱系。Neurotree 成立后不久,建立者意识到神经科学的指导关系大量来自其他领域,领域间的联系具有很大价值,因此开发了该系统,允许多个不同学术领域的谱系(包括 Neurotree)连接到同一个中央数据库。目前,该数据库囊括 69 个领域,包含大约 800 700 名研究者和他们之间的 758 200 个连接,并且以每周约 560 人的速度增长<sup>[31]</sup>。RePEc Genealogy 是经济领域学术谱系数据库,其数据以众包模式收集,任何人都可以对数据进行添加和修改,目前,该数据库中已经包含 14 000 名经济学家。此外,还有规模虽不大但更为详细和聚焦的贸易经济学家谱系 Family Tree of Trade Economists<sup>[32]</sup> 也经常被用到。

以上谱系数据库的可视化方式各不相同,MGP 为每个科学家建立单独页面,在页面上以文字形式描述该数学家名字、博士毕业年份、学位授予单位、国家、论文题目、博士生导师信息,并使用表格罗列其学生信息,页面上每个人名都能链接跳转至其本人页面。Neurotree 和 AFT 中导学关系以直观的家谱图形式呈现,可实现数据库的直接可视化和导航。RePEc 以网页文字形式记录了经济学家获得最高学位的时间、地点、导师及学生信息,但姓名中不包含跳转链接。谱系数据库可视化为用户提供了许多便利,但当前结果可视化仍存在一些挑战,如在规则和标准定义方面,科学家成长过程中未必都是单一导师指导,如何定义导学关系并将其制定为统一的规范仍有争议,也导致不同数据库可视化形式和所含信息存在差异;科学家可能在多个机构指导学生或经历单位变动,如何将不同机构表示在导学关系中,进而进行可视化,也是较为复杂的问题。此外,在数据来源方面,现有学术谱系库的构建都过于依赖人工,因数据来源困难,谱系数据库往往要依靠公开编辑完善信息,但这种做法难以保证质量和准确性,仍会出现大量数据缺失的现象,造成可视化结果并不完整。

## 4 科学家学术谱系测度

测度学术谱系的基础是学术谱系图(Academic Genealogy Graphs, AGG)。学术谱系图是对学术谱系结构化结果的表示,实际上类似于有向树中的根树(在严格限定条件后),也就是恰好有一个顶点的入度为 0,其余顶点的入度均为 1 的非平凡有向树,其中,入度为 0 的顶点称为树根,出度为 0 的顶点称为树叶,

从根到顶点  $v$  的距离称为  $v$  的层数,所有顶点的最大层数称为该树的高。根树中点与点的关系可以用家族关系表达。若点  $u$  不等于  $v$  且  $u \rightarrow v$ ,则称  $u$  为  $v$  的祖先, $v$  为  $u$  的后代;若  $(u, v)$  是根树中的有向边,则  $u$  称为  $v$  的父亲, $v$  为  $u$  的儿子,若某  $n$  个顶点是同一个父亲的儿子,则这  $n$  个顶点称为兄弟<sup>[33]</sup>。对应到导学关系上,树根即为整个谱系的初代导师(研究目标),树的每一层代表一代学生,树高代表师门延续的代数,父亲节点代表导师,儿子节点代表学生,兄弟节点代表同门师兄弟。对某一研究者学术谱系的量化研究和测度,一般都需要基于学术谱系图进行计算,了解树的基本理论和特征,对理解学术谱系图有重要意义。

### 4.1 测度指标及方法

学术谱系的测度指标主要分为两大类,一是描述性指标,即描述谱系基本特征的指标,该类一般基于谱系图中各结点进行简单计数,如描述某谱系绵延几代(图的层数)、某导师有几个学生(子辈节点数量)等;二是评价性指标,相较于描述性指标,评价性指标具有更深层次的衍生内涵,如对学者学术繁殖能力的测度指标。如表 1 所示:

表 1 学术谱系的测度指标

指标类型	具体指标	特点
描述性指标	A、C、A+C、T、G、W、TD、TA、繁殖力、生育力、后代、表兄弟、世代、关系、宽度、深度以及其他图(根树等)的相关描述性指标	直观、计数简单,不同学者对其赋予不同名称
评价性指标	繁殖力、谱系指数、 $g_m$ 指数	逐渐抽象,具有数学的论证逻辑,有明确的测度意义

就描述性指标而言,T. G. Russell 等<sup>[34]</sup>首次提出了计算和量化学术谱系的技术或方法,他们设计了 8 种不同指标:A 为个人担任学位论文导师的次数总和;C 为个人在学位论文中担任(非导师)委员会成员的次数总和;A+C 为个人以任何身份(导师或委员会成员)参与学位论文的次数总和;T 为个体谱系中后代总数;G 为个体后代的代数总和;W 为个体谱系中最大一代的后代总和;TD 计算个人谱系中后代的衰减影响;TA 分数为个人谱系中后来成为导师的后代数量的总和。以上指标可以量化学术谱系的产出,虽然通过文字叙述和表格形式的谱系也能计算出上述指标,但学术谱系图显然更直观和便于计数。L. Rossi 等<sup>[35]</sup>提出了 12 个衡量学术谱系图拓扑结构的指标,包括 6 个度量指标及其镜像对称指标,6 个度量指标分别是:繁殖力(Fecundity)、生育力(Fertility)、后代



(Descendants)、表兄弟(Cousins)、世代(Generations)、关系(Relationships)。“繁殖力”指的是个体后代的数量;“生育力”指的是能够繁殖的个体数量,即能够变成导师的学生数量;“后代”指的是与所研究学者有直接或间接指导关系的学生;“表兄弟”指的是两个人有不同的导师,但两位导师均是同一学者的学生。“关系”表示目标学者后代之间的联系数量。W. Dorés 等<sup>[11]</sup>用宽度(width)和深度(depth)两个指标来描述谱系树的结构,其中,宽度指的是学者指导的学生数量,深度代表谱系大小,也就是上文提到的树高。作者通过数据证明宽度和深度之间有较大相关性,并通过改良的 h 指数理解这种相关性。

值得注意的是,“繁殖力”是一个特殊而重要的指标维度,该指标既是描述性指标,也是评价性指标,

不同学者先后对其给出了定义和计算公式,差别在于是否考虑后代的多产性。后来,L.Rossi 等<sup>[36]</sup>基于文献计量学中的 h 指数提出谱系指数(Genealogical Index),D. K. Sanyal 等<sup>[37]</sup>基于文献计量学中的 g 指数提出了  $g_m$  指数,解决了谱系指数对多产后代关注不足的问题。谱系指数、 $g_m$  指数都包含了繁殖力指标的内涵。目前,评价性指标的数量相对较少,但伴随网络科学的不断发展,学者对学术谱系量化研究不断深入,评价性指标也将受到更多的关注。表 2 对比了目前部分学术谱系评价性指标的内涵、算法与特点。近年来,研究者选择性地使用描述性或评价性指标,结合对比分析法、文献计量法、履历分析法、网络分析法等方法,测度了各种应用场景下学术谱系中包含的特征和规律。

表 2 评价性指标对比

指标	指标内涵	计算公式	特点
繁殖力	导师在其整个学术生涯培训的学生数量	繁殖力分布 $p(k \Theta)=\pi_k p(k \kappa_k)+(1-\pi_k)p(k \kappa_{km})$	R. D. Malmgren 等 <sup>[39]</sup> 提出,首次给出了繁殖力的概率公式
	个人在培养多产后代中的多产程度	迭代统计(衡量长期影响)公式为 $l=n1+\gamma n2+\dots+\gamma m-lnm+\dots$ $\gamma=0$ 时繁殖力指的是直接后代	S. V. David 等 <sup>[30]</sup> 提出,考虑后代多产程度
谱系指数	学者拥有的学术后代(直接后代),即学者的学生数量	数据集 $v$ 的直接后代表示为: $F^+(v)=\{u \in V: (v,u) \in E\}$ 则繁殖力 $f^+(v)= F^+(v) $	L. Rossi 等 <sup>[35]</sup> 提出,只考虑学者直接培养的学生数量,不考虑所培养学生在指导学生中的贡献
	学者至少有 $g$ 个后代,且每个后代至少有其他 $g$ 个后代(可迭代多代)	对任意 $d \in N, m \in N \cup \{0\}, N$ 是正整数, $d$ 是整个谱系的代数,谱系指数 $g_{(d)}(v)=\max\{k \in N: l(v) \geq k \text{ 且 }  A_{(d)}^{(k)}(v)  \geq k\}$ , 其中 $A_{(d)}^{(m)}(v)=\{u \in D(v): g_{(d-1)}(u) \geq m\}$	可迭代计算第 1 至 $d-1$ 代范围内学者谱系指数,能衡量长期影响力。对多产后代关注不足,若学者学术后代中少数人执教,但执教后代的后代较多,影响力会被忽略
$g_m$ 指数	学者至少有 $g$ 个直接后代,而这些后代共同拥有至少 $g^2$ 个直接后代	令 $v$ 的 $l(v)$ 个后代的后代之和为 $S$ , $g_m(v)=\min(l(v), \sqrt{S})$	关注多产后代的贡献;谱系指数相等时, $g_m$ 指数能更好说明学者在塑造学术社区中的影响力

注:令学术谱系图  $G^*$  是有向图  $(V, E)$ ,  $V$  是顶点的有限集合,  $E$  是边的集合,  $v$  的直系后代为  $D(v)=\{u \in V: (v,u) \in E\}$ , 数量为  $l(v)=|D(v)|$

## 4.2 测度应用

### 4.2.1 评价学术影响力

目前,学界在评价学者的学术影响力时往往更关注其学术成果,如论文、专著以及引用等。但实际上,很多科学贡献并不反映在基于研究成果的计量指标上,研究者认为,学者最持久和重要的贡献其实是对下一代学者的培养<sup>[38]</sup>。学术谱系提供了新的视角和指标,使我们能够有效衡量学者在科学知识传播中的价值和在人才培养中的贡献,从而更为全面地评价学术影响力。

2010 年, R. D. Malmgren 等<sup>[39]</sup>在 *Nature* 上发文对比分析了 MGP 中的导师数据、学者出版物清单及入选美国国家科学院的学者名单,证明导师的学术繁殖力与其他学术成功指标密切相关,并认为这个结论提供了判断学术影响的方法。杨波<sup>[40]</sup>研究了学术谱系结构与学者影响力之间的关系,认为谱系内合作比例与谱系外引用比例都能显著增加学者影响力。

2017 年, L. Rossi 等<sup>[36]</sup>使用谱系指数量化了研究人员对其多代后代的指导贡献和影响,并认为谱系指数与文献指标相关性不高,具有独立性,能够作为科学出版物的补充,为科学影响力分析增加新的维度;而且该指数可迭代计算学者在第 1 至  $d-1$  代范围内的贡献( $d$  为谱系的总代数),因此能检验长期的学术影响力。D. K. Sanyal 等<sup>[37]</sup>使用  $g_m$  指数研究了导师在塑造研究社区方面的贡献,认为即使学者后代中只有少数人执教,但执教人群的指导贡献也不能被忽略。D. Kumar 等<sup>[41]</sup>提出 3 个端到端深度学习模型 ResIP-M1、ResIP-M2、ResIP-M3,通过对学术谱系网络的预测分析来预测研究人员的学术影响力。

### 4.2.2 探索人才培养机理

科研人员薪火相传、奋飞不辍,推动科学巨轮不断向前。如何培养出优秀的年轻力量,如何复制甚至超越前代科学家的辉煌,是各国都在积极关注的战略问题。学术谱系为这一问题的解答提供了新的视角和

思路。

E. F. Tuesta 等<sup>[15]</sup>通过导学关系探究了巴西精密与地球科学 (Exact and Earth Sciences) 及其 8 个子领域研究人员的合作关系, 并基于 Kulczynski 指数分析了导师和学生间依赖关系的演变。张志强等<sup>[42]</sup>通过对诺贝尔物理学奖获得者中师承关系的量化研究, 证明杰出导师可显著缩短其研究生接受新知识、做出重大创新成果、获得学术认可时所需的时间。冯靖雯等<sup>[43]</sup>聚焦诺贝尔化学奖得主 Lipscomb 的学术谱系, 从学术指导关系、表征谱系网络、节点连接强度 3 个维度量化表征师承关系特征, 揭示了以优秀人才为核心的科学家师承关系特征, 为识别和培养高层次人才提供参考。刘俊婉等<sup>[44]</sup>对谱系间的科学合作网络及引文网络特征进行计量分析, 构建了谱系内研究人员科研合作的评价指标 PR, 揭示谱系成员合作规律及其产生的科学生产力。王双等<sup>[45]</sup>采用图灵奖人工智能领域获奖者履历和学术产出数据, 运用理论分析和社会网络分析方法研究了科技人才成长的一般特征和规律。

#### 4.2.3 分析领域发展过程

科学既有创新性又有传承性, 师承关系是学派生成和发展的重要动力, 也是推动科学领域建制化、体系化、学科化的关键力量, 凸显着历史发展的重要信息<sup>[46]</sup>。研究者在很早以前就开始基于学术谱系对某一学派、领域和学科的发展演进进行梳理, 但这种研究大多使用定性分析方法, 遵从历史学研究的范式。目前, 已经有学者开始引入量化方法, 基于学术谱系分析领域发展脉络, 这种方法能更为直观和深入地理解领域发展。韩天琪等<sup>[47]</sup>对比分析了唐敖庆谱系和福井谦一谱系, 通过统计数据对比了国内外理论化学发展状况和特征。进入 20 世纪后, 科学在高度分化的基础上呈现出高度综合的特征, 跨学科研究正在不断增加, 如何监测领域的跨学科发展变化和规律是学界面临的一个难题。学术谱系有利于监测跨学科演化, C. R. Sugimoto 等<sup>[48]</sup>利用 80 年 (1930-2009 年) 内 3 038 篇图书馆与信息科学 (LIS) 博士论文的学术谱系网络数据来描述该学科的跨学科变化, 并认为使用学术谱系研究跨学科性是一种新的机会, 学术谱系的丰富性能够很好地促进这项研究。

#### 4.2.4 挖掘知识转移脉络

师徒传承是知识传播的重要手段之一, 故而导学关系下隐藏着知识流动的线索, 清晰完整的导学关系数据库可以作为研究思想生命周期的工具<sup>[30]</sup>。识别

几代研究人员之间的知识流动是了解最先进知识如何传播的重要途径, 但因缺乏数据, 目前还少有相关研究。L. Rossi 等<sup>[49]</sup>通过学术谱系提供的框架映射了学术知识主题, 从而分析科学知识发展模式。R. G. Castanha 等<sup>[50]</sup>通过文献耦合方法分析谱系内研究者之间理论—方法的近似度, 并将此近似度用于衡量谱系人员间网络链接强度及在后代中传播强度的贡献。F. Albornoz 等<sup>[51]</sup>通过构建领域相似度公式, 探索了研究主题的代际传播模式。S. A. Lee<sup>[52]</sup>在博士论文中定量分析了导学网络对发表论文的影响, 并研究了学术谱系在经济领域研究主题传播中的作用, 认为研究主题的传播模式与导学网络结构相关。J. H. Chariker 等<sup>[53]</sup>通过 AFT 谱系数据库对博士论文导师关系进行网络分析, 发现诺贝尔奖指导关系模式不随机, 并识别了诺贝尔奖的社区。

## 5 结论与展望

### 5.1 结论

学术谱系是跨学科研究领域, 涉及图书情报学、计算机科学与技术、科学技术史、教育学等多学科知识。从情报学研究视角来看, 学术谱系具有以下几方面的研究潜力和价值: ①在学术谱系构建上, 基于情报学的数据分析思维, 结合计算机科学与技术, 可以在一定程度上解决导学关系数据获取问题。同时, 通过多源数据融合方法, 将科学家的谱系数据、学术论文数据等多类型数据综合使用, 可以深入揭示科学家的知识传承问题。②在学术谱系测度上, 近年来, 研究者开始探索学术谱系的量化研究方法, 伴随着谱系指数、 $g_m$  指数的出现, 评价性指标、预测性指标计量正在成为学术谱系研究的新趋势, 也需要评价型情报研究的关注。③在学术谱系应用上, 目前学术谱系在各场景下的作用还有待深入挖掘。尤其是, 当前, 以文献计量学为方法论的科技人才评价将复杂的科技人才贡献局限于计量指标和数学模型, 简化了科学研究过程, 忽略了科学知识生产和传播的社会条件约束以及更多元的个体差异, 难以反映科技创新活动的复杂性。而学术谱系则记录了学者学术思想的生产和传播痕迹, 还原了科学研究的真实复杂过程。情报学研究者可以利用学术谱系的评价参考系作用, 发现学者和学术评价所需的关系线索<sup>[54]</sup>, 有助于完善现有科技人才评价方法。

### 5.2 展望

为更好发挥学术谱系的功能与作用, 未来学界应

重点关注以下 3 个方面的研究问题：首先，在数据来源上，学术谱系构建的最大难题是收集师承数据。目前，巴西在学术谱系构建及量化研究方面比较领先，根本原因是巴西的 Lattes 平台为此类研究提供了扎实可靠、容易获取的数据，虽然该平台的建设并不是为了解决学术谱系研究问题，师承数据也仅是该平台的一小部分，但不可否认的是，该平台已为学术谱系研究贡献了巨大的力量。我国在记录及公开师承信息方面还存在一些问题：①意识不足，各界总是将论文、项目、专著等成果信息看作是学者履历的重要元素，并未意识到师承信息的价值，因此在学者履历的收集及统计中，常常忽略此类信息。归根到底，是对学者在人才培养中的贡献认可不足。②缺少官方公共平台，目前学者主页、机构主页等网页上都会公开学者信息，即使包含师承信息，也面临信息零散、格式不一、缺少动态更新等问题，对学术谱系的构建和研究作用不大。Lattes 平台在成立之初，强制全国研究者对履历相关信息进行填写，平台运营平稳后，又稳抓数据更新和质量，因此能为学术谱系构建和研究提供丰富可靠的数据。我国也应在合适的条件下考虑借鉴 Lattes 模式，记录并公开学者的师承关系信息，否则，很多领域的师承信息将会逐渐湮没在历史之中，更深层次的测度和应用也无法开展，造成整个科学界的重大损失。

其次，在平台建设上，学术谱系数据库是学术谱系大规模使用和研究的重要前提。虽然我国已经为多个领域梳理了学术谱系并出版了丛书，但还缺乏如 MGP、AFT 这样的在线谱系数据库。在线谱系数据库有很多优势：①有助于探究人才成长规律、学者评价及学科知识转移，推动学术谱系的普遍使用和深入研究；②有利于谱系的动态更新，学术谱系的宽度和深度逐年变化，网页谱系数据库有利于数据的动态更新，促使整个谱系之树日益繁茂；③有利于信息的收集和补充，师承关系数据收集困难，在线谱系数据库能让用户成为内容编辑和提供者，集众人之力完善谱系。未来我国应考虑构建多学科学术谱系数据库，为科学发展记录宝贵信息，推进学术谱系的研究和应用。

最后，在主题拓展上，虽然学术谱系的量化研究已经受到重视，但目前还处于起步阶段，很多研究仅局限在概念梳理或提出相对简单的测度方法，仍有大量问题需要解决，如学术谱系图的测度可以引入更多树图计算方法，对谱系演化进行更深描述和预测；

可开发更多描述性和评价性指标，并深化各类场景的应用研究。未来，可以进一步聚焦知识、挖掘知识，探索更多方法测度知识遗产的转移规律；也可以通过学术谱系揭示的学科内部运行规律，在研究学派和学术传统，发现杰出科学家<sup>[55]</sup>，预测重大奖励、重大科技攻关，指导团队构建等方面有所突破。此外，谱系数据也可以与人口学数据融合使用，通过人口学特征，测度国际化教育、人才流动等研究问题。总之，作为跨学科研究领域，学术谱系研究需要多维视角，只有用更广阔的思维去看待学术谱系，才能让其焕发更大的生机和活力。

#### 参考文献：

- [1] 朱克曼. 科学界的精英：美国的诺贝尔奖金获得者 [M]. 周叶谦, 冯世则, 译. 北京：商务印书馆, 1979.
- [2] SUGIMOTO C R. Academic genealogy[M]// CRONIN B, SUGIMOTO C R. Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact. Cambridge: MIT Press, 2014: 365-382.
- [3] 科学家的学术家谱怎么修? [EB/OL]. [2023-06-07]. [https://epaper.gmw.cn/gmrb/html/2011-10/12/nw.D110000gmrb\\_20111012\\_1-06.htm?div=-1](https://epaper.gmw.cn/gmrb/html/2011-10/12/nw.D110000gmrb_20111012_1-06.htm?div=-1).
- [4] 佟艺辰. 科学家学术谱系的编史学研究 [D]. 北京：中国科学院大学, 2018.
- [5] 我国启动当代中国科学家学术谱系研究 [EB/OL]. [2023-06-07]. <http://paper.sciencenet.cn/htmlnews/2011/9/253046.shtml>.
- [6] 研究学术谱系 探究人才规律 [EB/OL]. [2023-06-07]. <http://news.sciencenet.cn/sbhtmlnews/2017/1/319478.shtml>.
- [7] genealogy[EB/OL]. [2023-06-07]. <https://www.merriam-webster.com/dictionary/genealogy>.
- [8] 商务印书馆辞书研究中心. 应用汉语词典 [M]. 北京：商务印书馆, 2002.
- [9] 冯永康, 田泓, 杨海燕, 等. 当代中国遗传学家学术谱系 [M]. 上海：上海交通大学出版社, 2016.
- [10] 袁江洋, 樊小龙, 苏湛, 等. 当代中国化学家学术谱系 [M]. 上海：上海交通大学出版社, 2016.
- [11] DORES W, BENEVENUTO F, LAENDER A. Extracting academic genealogy trees from the networked digital library of theses and dissertations[C]//Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries. New York: IEEE, 2016: 163-166.
- [12] MADEIRA G, BORGES E N, BARANANO M, et al. The Gold Tree: an information system for analyzing academic genealogy[C]// Proceedings of the 21st international conference on enterprise information systems (iceis). Portugal: SciTePress, 2019: 114-120.
- [13] MADEIRA G, BORGES E N, LUCCA G, et al. A tool for analyzing academic genealogy[C]//Lecture notes in business information processing. Germany: Springer Science and Business Media Deutschland GmbH, 2020:443-456.



- [14] KELLEY E A, SUSSMAN R W. An academic genealogy on the history of American field primatologists[J]. American journal of physical anthropology, 2007, 132(3): 406-425.
- [15] TUESTA E F, DELGADO K V, MUGNAINI R, et al. Analysis of an advisor-advisee relationship: an exploratory study of the area of exact and earth sciences in Brazil[J]. Plos one, 2015, 10(5): e0129065.
- [16] DAMACENO R J P, ROSSI L, MUGNAINI R, et al. The Brazilian academic genealogy: evidence of advisor-advisee relationships through quantitative analysis[J]. Scientometrics, 2019, 119(1): 303-333.
- [17] Lattes [EB/OL]. [2023-06-08]. <https://lattes.cnpq.br/>.
- [18] 高晓培, 武夷山, 李伟钢. 巴西人才库 Lattes 平台在优化科研和教育管理中的作用及其借鉴意义 [J]. 全球科技经济瞭望, 2014, 29(7): 32-42.
- [19] ProQuest Dissertations & Theses Global[EB/OL]. [2023-06-08]. <https://about.proquest.com/globalassets/proquest/files/pdf-files/brochures/pqdt/pqdt-global.pdf>.
- [20] Networked Digital Library of Theses and Dissertations[EB/OL]. [2023-06-08]. <https://ndltd.org/mission-goals-and-history/>.
- [21] Biblioteca Digital Brasileira de Teses e Dissertações [EB/OL]. [2023-06-08]. <http://bdtd.ibict.br/vufind/>.
- [22] WANG C, HAN J, JIA Y, et al. Mining advisor-advisee relationships from research publication networks[C]//Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2010: 203-212.
- [23] LI Y J, LIU Z, YU H. Advisor-advisee relationship identification based on maximum entropy model[J]. Acta Physica Sinica, 2013, 62(16): 168902.
- [24] WANG W, LIU J, XIA F, et al. Shifu: deep learning based advisor-advisee relationship mining in scholarly big data[C]//Proceedings of the 26th international conference on World Wide Web companion. Perth: International World Wide Web Conferences Steering Committee, 2017: 303-310.
- [25] ZHAO Z, LIU W, QIAN Y, et al. Identifying advisor-advisee relationships from co-author networks via a novel deep model[J]. Information sciences, 2018, 466:258-269.
- [26] GAO Y, WU X, YAN W, et al. Dynamic network embedding enhanced advisor-advisee relationship identification based on internet of scholars[J]. Future generation computer systems, 2020, 108: 677-686.
- [27] Allyn Jackson. A Labor of Love: The Mathematics Genealogy Project[EB/OL]. [2023-06-07]. <http://www.ams.org/notices/200708/tx070801002p.pdf>.
- [28] Mathematicians in the Genealogy Project[EB/OL]. [2023-06-07]. [https://genealogy.math.ndsu.nodak.edu/growth\\_image.php](https://genealogy.math.ndsu.nodak.edu/growth_image.php).
- [29] Mission statement[EB/OL]. [2023-06-07]. <https://genealogy.math.ndsu.nodak.edu/mission.php>.
- [30] DAVID S V, HAYDEN B Y. Neurotree: a collaborative, graphical database of the academic genealogy of Neuroscience[J]. Plos one, 2012, 7(10): e46608.
- [31] About the Academic Family Tree[EB/OL]. [2023-06-07]. <https://academictree.org/about.php>.
- [32] Family tree of trade economists[EB/OL]. [2023-06-08]. <http://www-personal.umich.edu/~alandear/tree/INDEX.HTM>.
- [33] 张先迪, 李正良. 图论及其应用 [M]. 北京: 高等教育出版社, 2005.
- [34] RUSSELL T G, SUGIMOTO C R. MPACT Family Trees: quantifying academic genealogy in library and information science[J]. Journal of education for library & information science, 2009, 50(4): 248-262.
- [35] ROSSI L, DAMACENO R J P, FREIRE I L, et al. Topological metrics in academic genealogy graphs[J]. Journal of informetrics, 2018, 12(4): 1042-1058.
- [36] ROSSI L, FREIRE I L, MENA-CHALCO J P. Genealogical index: a metric to analyze advisor-advisee relationships[J]. Journal of informetrics, 2017, 11(2): 564-582.
- [37] SANYAL D K, DEY S, DAS P P. G(m)-index: a new mentorship index for researchers[J]. Scientometrics, 2020, 123(1): 71-102.
- [38] MARSH E J. Family matters: measuring impact through one's academic descendants[J]. Perspectives on psychological science, 2017, 12(6): 1130-1132.
- [39] MALMGREN R D, OTTINO J M, AMARAL L. The role of mentorship in protege performance[J]. Nature, 2010, 465(7298): 622-626.
- [40] 杨波. 基于社会网络与内容分析的学术谱系传承与发展研究 [D]. 北京: 北京工业大学, 2018.
- [41] KUMAR D, BHOWMICK P K, PAIK J H. Researcher influence prediction using academic genealogy network[J]. Journal of informetrics, 2023, 17(2): 101392.
- [42] 张志强, 门伟莉. 诺贝尔物理学奖获得者中师承效应量化研究 [J]. 情报学报, 2014, 33(9): 926-935.
- [43] 冯靖雯, 赵勇. 学术谱系视角下杰出科学家的师承关系特征研究——以诺贝尔化学奖得主 Lipscomb 为例 [J]. 情报工程, 2020, 6(6): 22-32.
- [44] 刘俊婉, 王瑞, 杨波. 基于合作与引文网络的学术谱系职业传承研究 [J]. 情报探索, 2021(5): 8-14.
- [45] 王双, 赵筱媛, 潘云涛, 等. 学术谱系视角下的科技人才成长研究——以图灵奖人工智能领域获奖者为例 [J]. 情报学报, 2018, 37(12): 1232-1240.
- [46] 仇鹏飞, 孙建军, 闵超. 科学研究中的师承关系评述与思考 [J]. 图书与情报, 2018(5): 50-55, 118.
- [47] 韩天琪, 樊小龙, 袁江洋. 唐敖庆谱系与福井谦一谱系比较研究 [J]. 科学与社会, 2013, 3(1): 110-123.
- [48] SUGIMOTO C R, NI C, RUSSELL T G, et al. Academic genealogy as an indicator of interdisciplinarity: an examination of dissertation networks in library and information science[J]. Journal of the Association for Information Science & Technology, 2014, 62(9): 1808-1828.
- [49] ROSSI L, MENA-CHALCO J P. Mapeamento do conhecimento

- científico: uma proposta de método baseado em Genealogia Acadêmica[J]. Em questão, 2018, 24(6): 172-192.
- [50] CASTANHA R G, GRÁCIO M C C. Bibliographic coupling indicators for the evaluation of theoretical-methodological proximity in academic genealogy networks[J]. Revista digital de biblioteconomia e ciência da informação, 2020, 18: e020039.
- [51] ALBORNOZ F, CABRALES A, HAUKE E, et al. Intergenerational field transitions in economics[J]. Economics letters, 2017, 154: 1-5.
- [52] LEE S A. Advisor-advisee networks in economics[D]. Baltimore: The Johns Hopkins University, 2016.
- [53] CHARIKER J H, ZHANG Y, PANI J R, et al. Identification of successful mentoring communities using network-based analysis of mentor-mentee relationships across Nobel laureates[J]. Scientometrics, 2017, 111: 1733-1749.
- [54] 索传军, 张璇, 李木子. 论评价参照系的内涵、作用与构建——兼论学术谱系的功能和作用[J]. 中国人民大学学报, 2022, 36(4): 180-190.
- [55] 吕瑞花, 常欢, 席仕佳. 基于文献计量学的科学家学术谱系研究[M]. 北京: 中国科学技术出版社, 2020.

#### 作者贡献说明:

盛怡瑾: 收集资料, 撰写、修改论文;

赵勇: 提出选题, 修改、审定论文。

### The Connotation, Construction and Measurement of Scientists' Academic Genealogy

Sheng Yijin<sup>1,2</sup> Zhao Yong<sup>1,2</sup>

<sup>1</sup> China Agricultural University Library, Beijing 100083

<sup>2</sup> Information Research Center, China Agricultural University, Beijing 100193

**Abstract:** [Purpose/Significance] Academic genealogy is closely related to knowledge inheritance, with diachronic, systematic, and evaluative functions, containing a large amount of exploitable information and value. This paper systematically combs the relevant research on academic genealogy of scientists, aiming to reveal the value and potential of academic genealogy and provide reference for relevant research and practice development. [Method/Process] This paper analyzed the connotation of academic genealogy, summarized the academic genealogy construction method from the aspects of data collection, recognition of advisor-advisee relationship and visualization of results, and analyzed the current academic genealogy measurement research from measurement indicators, methods and measurement applications. [Result/Conclusion] Academic genealogy has important research value and potential in multiple research fields, including information science. In the future, we should focus on data sources, platform construction and theme expansion.

**Keywords:** academic genealogy genealogy connotation genealogy construction genealogy measurement