

●王建芳 冷伏海

共引分析理论与实践进展*

摘 要 共引分析是一种重要的情报研究方法。系统总结了共引分析理论与方法的发展过程;分析了共引分析在应用和实践层面,特别是科学结构揭示方面的进展。参考文献 15。

关键词 文献计量学 共引分析 科学结构

分类号 G257

ABSTRACT Co-citation analysis is an important method of information researches. In this paper, the authors systematically summarize the development path of theories and practices of co-citation, and analyze its developments in practical applications. 15 refs.

KEY WORDS Bibliometrics. Co-citation analysis. Scientific structure.

CLASS NUMBER G257

用于情报研究的文献计量方法中最具影响力的首推共引分析方法。它不仅可以用来揭示科学结构的发展现状乃至变化情况,还可以用来进行研究前沿分析、领域分析等,进而为宏观科技决策提供先行支持,为科技规划与评估提供基础。

1 共引分析理论的发展

共引,就是两篇文献同时被其他文献引用。一般认为同被引用的文献在主题上具有或多或少的相似性,因此同被引次数(即共引强度)可以测度文献在内容方面的相关度。由此,通过一组文献之间的共引关系可以形成共引网络,该网络内节点之间的远近便可以反映它们主题内容的亲疏关系。共引分析就是以此为原理,将一批文献(或著者、期刊)作为分析对象,利用聚类分析、多维标度等多元统计分析方法,借助计算机,把众多分析对象之间错综复杂的共引网关系简化为数目相对较少的若干类群之间的关系并直观地表示出来的过程^[1]。

共引分析方法始于 Small 于 1973 年提出的以文献为单位的共引分析,但共引概念可以推广到与文献相关的各种特征对象上,形成各种类型的共引概念,如:词的共引、文献共引、著者共引、期刊共引、主题共引和类的共引等。其中以文献共引分析研究最为突出。

1.1 对不同层次科学结构的揭示导引的共引理论发展

Small 等人的共引理论,是基于共引可以反映文献主题内容方面的相似性,及对共引关系的测度可以作为揭示科学结构的一种有效方法这样的假设而提出的。因此最初的理论是从科学结构图示的角度来验证方法的有效性,主要用于对学科及其相互关系的反映与印证。随着方法可靠性的确证,及科学技术发展的复杂化,共引理论作为揭示科学结构的实用方法,在科学结构逐步被从不同层面揭示的过程中,理论也不断完善。如在揭示了单个学科结构的基础上,为了综合揭示自然科学和社会科学的结构,发展了多次聚类方法来反映不同层次学科间的关系,并通过比较连续年度的聚类图来分析学科间关系的动态变化情况^[2]。Small 还利用共引聚类生成了多层次的科学结构图,即用一个总体图展示多学科宽度,然后逐层下钻,直到文献层次^[3]。

1.2 共引分析导引的知识发现探索

Small 等人的研究不局限于形式角度的科学结构揭示,还尝试用共引分析方法来展示知识结构。用高被引的文献集合代表关键概念,它们之间的共引关系表示概念间的关系,这样共引聚类形成的共引簇便转化为知识库,进一步在知识库中的知识搜寻可以导引新知识的发现。如 1986 年他试图从共引网络中利用准最小遍历树和深度优先检索方法来合成专业描述。另外,1999 年提出的书目数据库中的知识发现,提供了一种依据强大的共引链接在科学文献间创建路径的方法,得到了一条从经济学到天体物理学的包括

* 本文得到国家软科学项目“科学发展趋势预测的理论与方法及其实践研究”(2003DGQIB170)和国家自然科学基金项目“我国情报学学科发展、建设与前瞻性研究”(70373038)的资助。

331 篇文献的专门路径^[4],为新知识的发现提供了全新的线索。

1.3 其他理论进展

在文献共引分析理论发展的同时,其他角度的共引分析的出现与探索也说明了该理论的新进展,自 White 和 Griffith 在 1981 年提出著者共引分析理论以来,其理论发展也已经比较成熟,它除了可以反映科学的知识结构,还被用来研究科学交流模式,以及信息检索中知识结构的可视化。期刊共引的引入并运用于期刊及学科领域的研究^[5],主题和类的共引的引入并运用于领域分析^[6],乃至利用共引理论来探讨科学范式^[7]等等,都将共引分析理论研究提高到了新高度。

2 共引分析的方法及其演进

经过 30 多年的发展,共引分析研究方法可以分为宏观和微观两个层面^[8]。前者关注整体的学科结构,最终回答是什么控制科学发展规律的问题。后者的目的是回顾性地描述单个学科或者研究团体以及它们的相互依赖关系。由于一些原因,这方面的研究以著者共引研究为主。

共引分析方法随着方法原理研究的不断深入和应用层面的不断拓展而逐步完善,并且针对不同的应用场合和学科特点,总会利用不同的方法解决实际问题。共引分析方法演进中包括相当丰富的内容,但其基本的原理都是一致的。本文只结合共引分析的过程,对关键步骤中的主要方法改进做简要介绍。

共引分析方法的一般过程可以概括为:分析领域的确定;分析对象的选择及共引矩阵的形成;共引数据的处理;聚类分析和多维标度;结果分析和解释。方法的演进过程以分析过程中分析对象的选择和聚类方法中的参数修正为主。

2.1 对象选择中的方法改进

分析对象的选择就是用以进行共引分析的文献集的选择,其中最简单的方法是以高被引频次为依据的,即选择被引频次高于某数值的文献集作为分析对象。但不同学科在引文数量上有很大差异,这样必然导致学科结构分析结果失真,Small 为此提出用改进的指标分数引文量(fractional citation counting)来选择文献,就是每一篇引文都用引用它的来源文献的引文长度进行加权,以此来平衡学科差异。

选择了共引分析的文献集之后,还要对形成共引对的文献集合进行选择,进一步明确分析的对象。共

引对的选择,一般以共引强度为依据,后来为了消除不同学科,特别是规模大小不同的研究领域、学科之间共引强度的不可比性,将共引频次除以被引频次形成的标准化指标作为选择共引对的依据^[9]。

2.2 聚类分析中的方法演进

对基于选定文献之间的共引关系所形成的文献集进行逐步聚集,从而形成能够揭示科学结构的文献簇,这就是共引聚类的过程。聚类方法随着方法应用领域的拓展也在不断演进,有简单的等级聚类法(hierarchical clustering)、单一连结法(single-link clustering),常用的多维标度分析方法(multidimensional scaling analysis),以及 k-means 方法、因子分析方法(factor analysis)等。后来有人引入自组织图法(Self-Organization Map, SOM)和径路搜寻法(Pathfinder network scaling algorithm)等作为基于不同目的的共引聚类算法。

在已经实践的聚类过程中可以发现,引文率高的学科往往可以形成比较完备的聚类,而引文率低的学科如数学、工程等大学科在共引聚类中的地位明显不足。为此,Small 等人引入了可变水平聚类方法(Variable level clustering),即在聚类过程中限定类中结点数目,不断调整聚类临界值的方法,来消除不同学科间引文率不同所造成的分析结构的不平衡性。Small 还提出以类聚类(clustering of clusters)的反复聚类方法,即以一次聚类形成的簇为单元再次聚类,逐步形成大类、超类^[10]。

3 共引分析方法应用层面的拓展

共引分析作为一种信息分析方法,从引文的角度分析文献等的相互关系,揭示微观的文献所代表的主题关系、中观的学科结构和宏观的科学结构,从而将文献信息所反映的科学的、领域的、知识的关系可视化,提供了一个探索科学结构本身,及学科内部相互关系和发展脉络的全新思路,以此为基础,可以应用在科学研究的多个方面,当然应用领域也是在方法不断成熟的基础上不断拓展的结果。

3.1 共引方法用于描绘科学结构的发展历程

共引分析方法首先被应用于揭示科学结构,绘制科学结构图,即依据共引关系,用类似于地图的方式,揭示一个学科、领域、专业、单个文献和作者之间的相互关联关系,将它们的相关性用物理邻近和相对位置关系的方式表示出来。这方面的工作仍以 Small 等人的研究为代表,其发展历程可以归纳如下:

(1) 中观层次的学科、专业结构的描绘, 宏观的科学结构的揭示, 及微观的文献间关系的表示。

(2) 不同层次科学结构技术的综合, 形成逐级细化的系统的科学结构图。

(3) 连续时间内共引聚类图的历时比较, 反映科学结构的变化。

(4) 从不同学科间的共引关系中找寻某一学科到另一学科的可通路径, 从而描述知识结构。

(5) 综合层次的领域分析。

(6) 交叉学科的边界分析, 相近学科间的比较分析, 及特定学科的研究前沿总结。

(7) 结构描述的可视化、程式化。

共引分析应用的发展还可以从应用领域方面描述。应用领域经历了从单一的小学科角度, 如粒子物理、免疫学、行为科学、化学工程, 特别是信息科学领域的信息检索、信息政策、知识管理等, 到大学科的社会学、心理学、经济学、遗传学, 再到整体科学结构展示的发展历程。

3.2 共引分析用于信息检索效率的提高

共引分析可以用于揭示科学结构, 进而揭示学科结构和知识结构, 因此从另一角度讲, 它也是一种知识聚类的方法, 可以用于相关学科、知识领域的文献、著者和主题的汇集, 进而用于信息检索方法的改进和检索效率的提高。在上世纪 80 年代就有人将共引分析结果应用于信息检索, 如 1981 年, White 提出用著者共引分析方法改进检索策略来辅助联机检索, 提高检索效率。其他还有利用共引聚类结果进行检索结果过滤的提法。后来 Rees-Potter 等人建立的动态词表系统, 便是利用引文和共引分析方法来建立、更新和维护术语词表^[11]。

3.3 用于科学交流模式的展示

引文本身是学者们信息交流模式的一种直接表现, 因而共引现象也可以从一个角度反映科学交流的模式。早在 1979 年, Price 就提到共引可以反映科学研究所形成的隐形的科学共同体——无形学院 (invisible college)。之后, 共引分析逐渐成为科学交流实证研究最重要的方法。最近 Gmur 等人还对利用共引分析方法进行科学交流研究的效果, 进行了评估和比较研究^[12]。

3.4 共引分析理论向网络结构研究的移植

在网络环境中, 站点的链接关系类似于文献的引用关系, 因而文献的共引分析方法可以移植到网络站点共引研究, 反映网络本身的结构和网络中知识的结

构。有人介绍了一个基于网络引文数据库的自动著者共引分析过程, 分析过程运用系统聚类和多维标度绘制著者聚类图, 聚类结果和结构图可以被植入到以引文为基础的检索系统 PubSearch 中^[13]。另有人提出了共链分析 (co-linking analysis), 用来揭示网络结构, 发掘其中暗含的知识, 使用的方法也是多维标度等^[14]。

4 共引分析实践的探索

4.1 共引分析的系统平台开发

Small 等人首先开发了基于共引理论的单机系统 SCI-Map 来描绘科学文献间的结构。SCI-Map 可以利用共引及其文献关系测度方法执行以聚类为基础的检索, 不仅使用户探究大的文献集合的结构成为可能, 也使得分析家们可以控制和可视化聚类过程。

1997 年的科学图示项目 (Science mapping) 加入了虚拟现实软件, 可以将大规模的文献数据创建成为大型图示, 从而畅游合成的三维空间。同年, SciViz system 运用数据可视化技术为科学图示化提供了新的可能, 可以将大量的文献数据置于一个简单的坐标空间中, 还描绘了一个实现此目标的基本框架^[15]。

4.2 用于科学研究前沿分析的应用范例——ISI 的科学前沿分析

1989 年就有人提出用共引聚类方法确定科学研究前沿的设想, 但真正实践该方法的当属 ISI。它利用共引分析方法进行科学前沿分析, 分析结果定期以热点问题、研究前沿等形式进行跟踪报道。这样可以帮助学者们确定特定领域研究的重要性及其发展程度。分析过程是通过识别 5 年内多学科中引用率最高的文献, 并用共引强度来确定研究前沿需要处理的共引文献集, 以将那些关系紧密的文献聚类。共引频率低的文献的过滤用一个标准化的临界值确定, 即用共引频次来除两文献引用频次乘积的平方根。最后, 从符合临界值的一个共引对开始, 进行单连结聚类, 以此归纳前沿问题。

共引分析理论、方法、应用与实践的发展, 使得我们相信它已经成为一种可靠实用的情报研究方法, 尤其是分析结果的客观系统性, 使得它在科学结构的分析把握方面已经发挥了很大作用。共引分析不仅可以揭示学科和科学结构及其相互关系, 还可以发现科研的前沿领域, 进行问题跟踪、趋势预测, 进而用于各种角度的评价等, 它在情报研究服务中的信息跟踪、比较研究、预测研究和科技评价中都可以发挥一定作

用。共引分析作为一种定量分析方法,结合其他分析方法和专家知识,可以对科技研究活动进行多层次的分析,挖掘数据信息,并利用可视化技术形象表达所得到的知识,从而为政府决策者、科研部门和企业的技术管理者和科研工作者提供有效的决策支持。

参考文献

- 1 赵党志. 共引分析——研究学科及其文献结构和特点的一种有效方法. 情报杂志, 1993(5)
- 2 Small, H. Macro-level changes in the structure of co-citation clusters: 1983 - 1989. *Scientometrics*, 26(1): 5 - 20, 1993
- 3 Small, H. Visualizing Science by Citation Mapping, *Journal of the American Society for Information Science*. 50(9): 700 - 813, 1999
- 4 Small, H. A passage through science: crossing disciplinary boundaries. *Library Trends*, 48(1): 72 - 108, 1999
- 5 Tsay, Ming-yueh; Wu, Chia-wen; Xu, Hong. Journal co-citation analysis of semiconductor literature. *Scientometrics*, 57(1): 7 - 25, 2003
- 6 Chinchilla-Rodriguez, Zaida; Corera-Alvarez, Elena; Herrero-Solana, Victor. A New technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1): 129 - 145, 2004
- 7 Small, H. Paradigms, citations, and maps of science: a personal history. *Journal of the American Society for Information Science and Technology*. 54(5): 394 - 9, 2003
- 8 Markus Gmur. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1): 27 - 57, 2003
- 9 Small, H; Sweeney, E. Clustering the Science Citation Index

using co-citations: 1-A comparison of methods. *Scientometrics*, 7(3-6): 391 - 409, 1985

- 10 Small, H; Sweeney, E; Greenlee, E. Clustering the Science Citation Index using co-citations: 2-mapping science. *Scientometrics*, 8(5-6): 321 - 340, 1985
- 11 Rees-Potter; Lorna K; Rees-Potter, L K. Dynamic thesaural systems: a bibliometric study of terminological and conceptual change in sociology and economics with application to the design of dynamic thesaural systems. *Information Processing and Management*, 25(6): 677 - 691, 1989
- 12 Gmur, Markus. Co-citation analysis and the search for invisible colleges: a methodological evaluation. *Scientometrics*, 57(1): 27 - 57, 2003
- 13 Yulan He, Siu Cheung Hui. Mining a Web Citation Database for author co-citation analysis. *Information Processing and Management*, 38(4): 491 - 508, 2002
- 14 Faba-Perez, Guerrero-Bote, De Moya-Anegon. Data mining in a closed Web environment. *Scientometrics*, 58(3): 623 - 640, 2003
- 15 Small, H. A general framework for creating large scale maps of science in two or three dimensions: the SciViz system. *Scientometrics*, 41(1-2): 125 - 133, 1998

王建芳 中国科学院文献情报中心, 中国科学院研究生院情报学专业2004级博士研究生。通信地址: 北京市海淀区中关村北四环西路33号。邮编100080。

冷伏海 教授, 中国科学院文献情报中心博士生导师, 情报研究部副主任。通信地址同上。

(来稿时间: 2005-04-13)

(上接第55页) 由应用开发商把需要的安全逻辑(如加密、签名、时间戳等)写到业务逻辑中, 要么由安全开发商对原来的应用做二次开发, 加入安全逻辑。这一过程费时费力, 给证书的应用造成了不便。采用安全应用支撑平台代理完成所有的加密、签名等安全事务, 可以将安全逻辑从业务逻辑中分离出来, 做到证书的应用基本与应用开发无关。这样, 安全应用支撑平台就为数字证书的便捷应用创造了条件。

参考文献

- 1 国家信息安全工程技术研究中心. 电子政务总体设计与技术实现. 北京: 电子工业出版社, 2003
- 2 马朝斌. 大力加强电子政务内网的安全保密建设和管理. 信息安全与通信保密, 2003(12)
- 3 谭兴烈. 电子政务安全解决方案要解决的主要问题. 信息

安全与通信保密, 2004(5)

- 4 郭贺铨. 电子政务安全体系. 信息安全与通信保密, 2003(4)
- 5 刘远航, 崔维利. 电子政务信息安全和 PKI/PMI 体系. 网络安全技术与应用, 2002(7)
- 6 姜楠, 王健. 电子政务中基于 PKI 的角色授权管理策略. 通信技术, 2003(9)
- 7 许长枫, 刘爱江, 何大可. 基于属性证书的 PMI 及其在电子政务安全建设中的应用. 计算机应用研究, 2004(1)

李纲 武汉大学信息管理学院教授。通信地址: 武汉。邮编430072。

王毅彦 武汉大学信息管理学院博士研究生。通信地址同上。

(来稿时间: 2005-04-25)