

基于CiteSpace II的“共引分析”领域知识图谱分析*

□ 朱亮 赵瑞雪 寇远涛 鲜国建 / 中国农业科学院农业信息研究所 北京 100081

摘要: 文章以ISI Web of Knowledge数据库中收录的以“共引分析”(Co-citation Analysis)为主题的文献为基础数据,以信息可视化软件CiteSpace II为工具,绘制出“共引分析”领域的知识图谱,并对其进行分析,得出“共引分析”领域的经典文献、主要研究热点及研究前沿等。

关键词: 共引分析, 可视化分析, CiteSpace II

DOI: 10.3772/j.issn.1673—2286.2012.12.011

1 引言

1973年,美国情报学家Small在其发表的论文中首次提出“共引(Co-citation)”概念,之后,这方面的研究与应用便广泛地开展起来,共引分析也成为最具影响力的一种引文分析方法。共引分析不仅可以用来揭示科学结构的发展现状乃至变化情况,也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]。

共引又称同被引,若两篇文献同时被后来的 n 篇文献所引用,则称这两篇文献具有共引关系,其共引强度为 n 。一般认为同被引的文献在主题上具有或多或少的相似性,这种相似性可由共引强度来体现, n 越大则这两篇文献关系越

密切。以此为原理,由共引关系入手研究一批文献及其之间的关系便成为一条切实可行之路。共引分析就是以具有一定学科代表性的一批文章(或著者或期刊)为分析对象,利用聚类分析、多维标度等多元统计分析方法,借助电子计算机,把众多的分析对象之间错综复杂的共引网状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,使分析对象之间相互关系的格局清晰可辨,在此基础上分析研究对象所代表的学科及文献的结构和特点^[2]。当前,共引分析研究主要有两个系列:以Small为代表的文献共引分析和以White为代表的作者共引分析。除此之外,共引分析也被扩展到其他与文献相关的分析对象上,如期刊共引分析。

2 工具和方法

通常情况下,研究人员进入一个新的领域,首先会研读这个领域的经典文献,了解其发展脉络。经典文献往往代表了一个领域的产生、转折和重大突破,所载的观点、理论和技术方法会对该领域今后的发展产生重要影响;其次是学习、理解这个领域的研究热点和前沿,把握其发展现状。研究前沿(Research Front)是科学研究中最先进、最新、最有发展潜力的研究主题或方向,通常代表了科学发展的难点、热点以及发展趋势。研究前沿往往来源于新的科学发现或进展,这些新的科学发现或进展常伴随着相应文献计量学特征出现,跟踪和探测这些文献计量学特征就有可能揭示研究

* 本文系国家“十二五”科技支撑计划项目课题“基于STKOS的知识服务应用示范”(编号:2011BAH10B06)的研究成果之一。

前沿的存在^[3]。目前,学科领域经典文献、热点、前沿的识别和探测方法有许多,但基于文献引文和主题词的文献计量分析方法仍是其中的主流。建立在这些方法之上的系统和工具也层出不穷,其中表现最为出色的当属陈超美博士开发的CiteSpace系列软件。陈超美将研究前沿定义为“一组突现的动态概念和潜在的研究问题,其知识基础是研究前沿概念所在文献的引用文献簇,研究前沿与知识基础相互作用并动态发展^[4]”。为提高研究前沿揭示的时效性,CiteSpaceII采用爆发词算法来辨认研究前沿专业术语概念。在CiteSpaceII中,研究前沿是基于从文献题目、摘要、系索词中提取出的突变专业术语而确定的。运行软件可生成共引与共

词混合网络图谱,通过它研究人员能够直观地辨识出科学前沿的演化路径及学科领域的经典文献。基于此,本文利用CiteSpaceII绘制知识图谱对共引分析的核心文献、研究热点、研究前沿进行分析讨论。

3 数据来源

本文研究数据来源于Web of Science,引文数据库选择Science Citation Index Expanded (SCI-EXPANDED),1995年至今;Conference Proceedings Citation Index - Science (CPCI-S),1995年至今。时间选择为所有年,文献类型选择为ARTICLE、PROCEEDINGSPAPER、REVIEW,检索式为主题=“co-

citation analysis” or “cocitation analysis”,共检索到225条记录,数据下载日期为2012年6月14日。

检索到的225条文献记录,分别来自106种不同的出版物,共有375个作者,分布于美国(81篇)、中国(17篇)、加拿大(16篇)、中国台湾(15篇)、英国(14篇)等30多个国家和地区,基本全部为英文文献。其总被引频次为3127,平均每篇被引13.9次。文献年代分布如表1所示。

统计结果表明,在2005年之前,以“共引分析”作为主题的文献量每年基本都在个位数,只有2001及2003年达到两位数,表明在这一时期,“共引分析”的研究还处在非稳定阶段,且热度不高;从2005年开始,“共引分析”的研究热度得到明显提升,以“共引分析”作为主题的文献年平均发文量达到20以上,且文献流较为稳定、连续。

表1 文献年代分布表

出版年	文献量	出版年	文献量	出版年	文献量
1995	1	2001	12	2007	22
1996	6	2002	9	2008	25
1997	3	2003	13	2009	18
1998	7	2004	7	2010	26
1999	9	2005	19	2011	18
2000	4	2006	14	2012	12

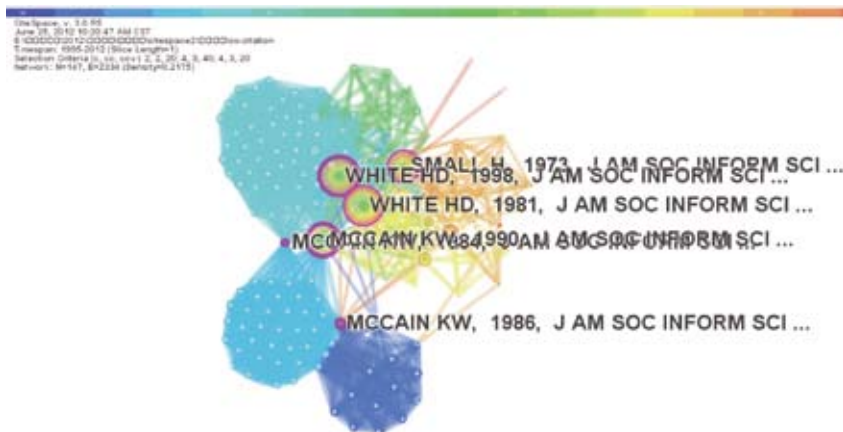


图1 共引分析文献共被引知识图谱

4 结果分析

4.1 关键节点文献分析

将检索到的225条数据记录输入到CiteSpaceII中进行共被引分析,节点类型(Node Types)选择为被引文献(Cited Reference),时区选择为1995—2012年,时间跨度选择1年,分别在前、中、后三个时间分区中设定引文数量、共引频次和共引系数三个层次的阈值,阈值选择为(2, 2, 20)、(4, 3, 40)、(4, 3, 20),其余各个时间分区的阈值由线性内插值来决定。运行软件,得到共引分析研究文献共被引网络结构组配(表2)和知识图谱(图1),网络中包含147个节点和2334条边。

表2共被引网络结构组配

时间分区	C\CC\CCV	文章数量	节点数量	连线数量
1995-1995	2\2\0.2	31	0	0
1996-1996	2\2\0.22	210	22	149
1997-1997	2\2\0.24	257	0	0
1998-1998	2\2\0.27	296	42	808
1999-1999	2\2\0.29	962	58	1290
2000-2000	3\2\0.31	213	0	0
2001-2001	3\2\0.33	394	3	3
2002-2002	3\2\0.36	434	2	1
2003-2003	3\2\0.38	733	18	87
2004-2004	4\3\0.4	240	0	0
2005-2005	4\3\0.38	593	5	7
2006-2006	4\3\0.35	448	7	17
2007-2007	4\3\0.32	815	5	3
2008-2008	4\3\0.3	1019	19	103
2009-2009	4\3\0.28	692	11	29
2010-2010	4\3\0.25	1089	17	83
2011-2011	4\3\0.22	806	11	42
2012-2012	4\3\0.2	418	5	10
合计		9650	225 (147)	2632 (2334)

表3 共被引网络关键节点文献信息

作者	发表年	出处	中心度	被引频次
MCCAIN KW	1990	J AM SOC INFORM SCI	0.39	66
MCCAIN KW	1984	J AM SOC INFORM SCI	0.32	11
WHITE HD	1998	J AM SOC INFORM SCI	0.25	73
MCCAIN KW	1986	J AM SOC INFORM SCI	0.22	20
WHITE HD	1981	J AM SOC INFORM SCI	0.14	72
SMALL H	1973	J AM SOC INFORM SCI	0.13	61

图1中每个节点代表一篇被引文献，带有紫色圆圈标记的节点表示的是关键节点。根据陈超美博士的定义，关键节点是共被引网络图谱中连接两个以上不同聚类，且相

对中心度和被引频次较高的节点^[5]。关键节点在聚类网络中起到了连接和过渡的桥梁作用，代表了一个聚类到另一个聚类的跃迁。图中共有6篇关键节点文献，其中心

度、被引频次等信息如表3。

六篇关键节点文献中的最早者是美国费城科学信息研究所Small于1973年在*Journal of the American Society for Information Science*上发表的*Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents*，此文正是“共引分析”的开山之作，文中首次提出“共引”概念，并指出共被引文献网络可以对某个具体的科学学科领域进行分析，被引文献集为科学构成提供了一种新的研究思路^[6]。其次是1981年于美国Drexel大学White和Griffith联合发表的*Author cocitation: A literature measure of intellectual structure*一文，该文开创了作者共引分析（Author Cocitation Analysis, ACA）的先河，通过1972~1979年39位情报学家的共引情况描绘了他们在学科中的位置和情报学的学科结构^[7]。

中心度最高的是来自于McCain的*Mapping Authors in Intellectual Space: A Technical Overview*，发表于1990年，这是一篇关于作者共引的综述性文章，文中McCain将ACA的程序规整为选择作者、检索共引频次、生成共引矩阵、转化为Pearson相关系数矩阵、多元分析和解释结果等几个步骤，人们称之为传统的ACA模式^[8]。

从被引次数来看，最高者是White和McCain在1998年联合发表的*Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995*，White和McCain采用ACA，通过对24年间（1972-1995）一些代表性的文献特征（作者共引数目）归

纳总结情报学领域的结构特征和24年来的发展情况,并将分析结论进行可视化^[9]。

4.2 研究热点分析

主题词是一篇文章核心内容的高度概括和凝练,对文章的主题词进行分析,利用其频次高低来确定研究领域的热点内容是目前研究人员较常采用的方法之一。因此,本文将借助CiteSpaceII,利用文献题录中的主题词来确定共引分析研究的热点领域。

通过设置合适的阈值,选择关键路径算法(Pathfinder),运行CiteSpaceII软件得到如图2所示的高频主题词知识图谱,共有292个节点和346条边。从图中可以看出,co-citation analysis\cocitation analysis(共引分析)、intellectual structure(知识结构)、author co-citation analysis(作者共引分析)、factor analysis(因素分析)是共引分析研究的重点和热点领域。

表4列出了出现频次在10次以上的11个高频主题词,其中频次最高的是co-citation analysis\cocitation analysis,频次为58。其次是intellectual structure,频次为40,表明共引分析在描绘学科的知识结构方面应用广泛,发挥着重要的作用。排在第三位的是author co-citation analysis,这是共引分析研究的两个系列之一,一直以来都是学者们关注的重点和热点领域。除此之外,information science(信息科学)、factor analysis(因素分析)、multidimensional scaling(多维标度)、citation analysis(引文分析)、cluster analysis(聚类分析)、information

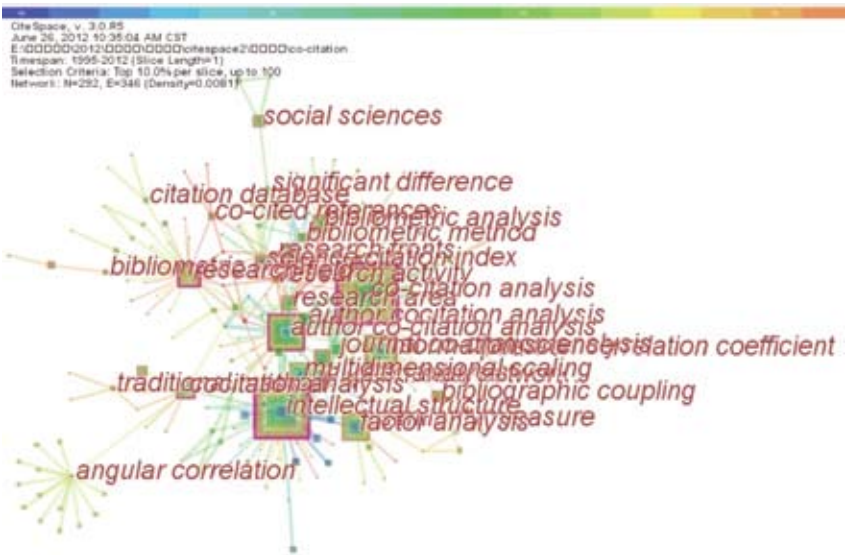


图2 共引分析研究热点知识图谱

表4 共引分析领域出现频次超过10的主题词信息

序号	主题词	频次
1	co-citation analysis\cocitation analysis	58
2	intellectual structure	40
3	author co-citation analysis	27
4	information science	24
5	factor analysis	20
6	multidimensional scaling	16
7	research field	16
8	bibliometric analysis	12
9	citation analysis	12
10	cluster analysis	12
11	information retrieval	11

retrieval (信息检索) 等都是共引分析研究的重点领域。

4.3 研究前沿分析

CiteSpace II 采用爆发词算法来辨认研究前沿专业术语概念。爆发词算法可以用于监测一个学科领域内研究兴趣的突然增长,通过考察词频的时间分布,将那些频次变化率高、频次增长速度快的“爆发词”从大量文本中抽取出来,用词频的变化趋势,而不仅仅是词频的高低来分析学科领域的前沿领域和发展趋势。

本文利用CiteSpace II 提供爆发词探测技术, 将文献记录的题目、摘要、索引词和文献记录标识中的突变专业术语提取出来, 以此来确定共引分析研究的前沿领域。针对检索到的225条文献记录, 共探测到109个突变术语(burst term)。术语类型选择“Burst Terms”, 节点类型选择“Term”和“Cited Reference”, 阈值设置为(2, 2, 20)、(4, 3, 30)、(4, 3, 20), 运行软件, 生成被引文献共引网络与突变术语共词网络构成的混合网络图谱, 如图3所示。

由混合网络图谱显示的结果来看,变化频次排在前几位的突变专业术语是:作者共引分析(author cocitation analysis)、相关系数(correlation coefficient)、信息检索(information retrieval)、引文数据(citation data),分别代表了近年来国际共引分析研究重要的前沿领域。自从McCain对作者共引分析的步骤详加阐述后,大家一直以来都按照这种约定俗成的方法来进行共引分析。但传统ACA模式需要大量

的计算与绘图操作, 计算强度较大, 严重阻碍了共引分析的应用。为此, 近几年来, 围绕着共引分析流程的优化, 国际图书情报学界开展了广泛的讨论。在2005年以前, 该讨论主要集中在相似系数的选取方面。2006年以后, 该讨论便扩展到了相似系数矩阵的生成方式方面: 是利用共引矩阵还是引文矩阵。在

应用方面,将共引分析方法与信息检索进行结合是学者们的又一重要关注点,相关实践已有不少,其中最主要的方向便是检索结果的关联展示。另外,由于用来进行共引分析的引文数据本身质量高低会直接影响最终的分析结果,因此,如何获得高质量的引文数据也已成为当前共引分析研究一个重要内容。



图3 共引分析研究被引文献共引与突变术语共词混合网络知识图谱

5 结语

经过近40年的发展,共引分析的理论研究、实践应用已日渐成熟,已成为图书情报领域一种重要且高效的分析方法。本文对Web of Knowledge收录的1995-2012年间以“共引分析”为主题的文献进行统计和可视化分析,得出如下结论:

(1) 在共引分析的发展进程中, Small、White、McCain、Griffith等几位学者作出了突出贡献, 其代表作品有着重要的影响力。共引分析最早产生于美国, 这一领域的重要论文基本都发表在美国本土期刊上, 其中的典型代表是 *Journal of the American Society for Information Science*, 此刊也是近年来共引分析流程优化大讨

论的主要阵地。从发表时间上看, Small于1973年发表的*Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents*一文首次提出“共引”概念,被认为是共引分析研究的起点;1981年White和Griffith联合发表的*Author cocitation: A literature measure of intellectual structure*一文,开创了ACA的先河,在此基础上, McCain于1990年撰文*Mapping Authors in Intellectual Space: A Technical Overview*提出了传统ACA模式。

(2) 通过对CiteSpaceⅡ生成的知识图谱进行分析, 确定了共引分析研究的热点研究领域主要是: co-citation analysis\cocitation

analysis (共引分析)、intellectual structure (知识结构)、author co-citation analysis (作者共引分析)、information science (信息科学)、factor analysis (因素分析)、multidimensional scaling (多维标度)、citation analysis (引文

分析)、cluster analysis (聚类分析)、information retrieval (信息检索) 等。共引分析研究的主要前沿领域是: 作者共引分析 (author cocitation analysis)、相关系数 (correlation coefficient)、信息检索 (information retrieval)、引文

数据 (citation data) 等。

以上对共引分析领域经典文献、主要研究热点和前沿的可视化分析, 一方面, 是对共引分析发展历程的梳理; 另一方面, 也希望对今后我国相关领域的研究有一些参考作用。

参考文献

- [1] 王建芳, 冷伏海. 共引分析理论与实践进展[J]. 中国图书馆学报, 2006(1): 85-88.
- [2] 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[J]. 情报杂志, 1993, 12(2): 36-42.
- [3] 陈仕吉. 科学研究前沿探测方法综述[J]. 现代图书情报技术, 2009(9): 28-33.
- [4] CHEN C. 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.
- [5] CHEN C. The Centrality of Pivotal Points in the Evolution of Scientific Networks [C]// Proceedings of the International Conference on Intelligent User Interfaces (IUI 2005). San Diego, CA, 2005.
- [6] SMALL H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents [J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.
- [7] WHITE H D, GRIFFITH B C. Author Cocitation: A Literature Measure of Intellectual Structure [J]. Journal of the American Society for Information Science, 1981, 32(3): 163-171.
- [8] MCCAIN K W. Mapping Authors in Intellectual Space: A Technical Overview [J]. Journal of the American Society for Information Science, 1990, 41(6): 433-443.
- [9] WHITE H D, MCCAIN K W. Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995 [J]. Journal of the American Society for Information Science, 1998, 49(4): 327-355.

作者简介

朱亮 (1981-), 男, 汉族, 助理研究员, 在读博士。研究方向: 信息资源管理。E-mail: zhuliang@caas.net.cn

赵瑞雪 (1968-), 博士, 研究员, 博士生导师, 研究方向: 信息管理与信息系统、信息资源管理、知识组织及数字图书馆。E-mail: zhaorx@mail.caas.net.cn

Visualization Analysis of Co-citation Analysis Research Based on CiteSpace II

Zhu Liang, Zhao Ruixue, Kou Yuantao, Xian Guojian / Agricultural Information Institute of CAAS, Beijing, 100081

Abstract: In this paper, we retrieved "co-citation analysis" data records from ISI Web of Knowledge, and then used the visualization software CiteSpace II to draw the knowledge map of co-citation analysis. Based on this, we discussed the critical literature, main research focus and research fronts in co-citation analysis domain.

Keywords: Co-citation analysis, Visualization analysis, CiteSpace II

(收稿日期: 2012-08-01)