基于共词分析法的学科主题演化研究进展与分析

唐果媛[1,2] 张薇[1]
[1] 中国科学院文献情报中心 北京 100190 [2] 中国科学院大学 北京 100049

摘要: [目的/意义] 分析学科主题演化趋势,对科研人员研究学科知识、决策层规划学科布局都有重要意义。相比于词频分析法和共引分析法,共词分析法的优势是能深入文献内部,从微观角度揭示学科主题演化规律。分析中国国内基于共词分析法的学科主题演化研究现状,以期为相关研究人员提供参考和借鉴。 [方法/过程] 采用人工判读法提炼出基于共词分析法的学科主题演化研究分析流程的 5 个步骤,并对每个步骤中研究人员使用的策略、分析手段和工具进行归纳总结。 [结果/结论] 数据集的来源数据库主要有综合类、专门类和引文类等 3 种,检索策略有基于词、基于期刊和复合检索策略等 3 种; 共词分析对象来源主要为作者关键词,关键词选取主要基于关键词词频、关键词共现词频和前两者相结合 3 个角度; 构建共词矩阵时使用得最多的归一化系数为 ochiai 系数; 最常用的主题演化分析手段为聚类分析和社会网络分析图谱; 使用得最频繁的工具为 SPSS 软件。

关键词: 学科主题演化 共词分析法 主题演化阶段 检索策略
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2015.05.020

1 引言

科学技术作为第一生产力,对推动社会的发展有着非常重要的作用。党的十八大报告明确提出“科技创新是提高社会生产力和综合国力的战略支撑,必须摆在国家发展全局的核心位置”。2014 年 6 月,习近平总书记在中国科学院第十七次院士大会、中国工程院第十二次院士大会上的讲话中提出: 科技是国家强盛之基,创新是民族进步之魂; 实施创新驱动发展战略,最根本的是要增强自主创新能力,最紧迫的是要破除体制机制障碍,最大限度解放和激发科技作为第一生产力所蕴藏的巨大潜能。科技创新,情报先行,学科情报研究是为了把握科技发展动向,支持科研决策与管理,以促进国家科技事业的可持续发展,充分发挥“尖兵、耳目和参谋”的作用。随着科学技术的飞速发展,学科发展态势呈现出学科演变更替加速和学科交叉融合加剧两大特征。学科主题演化研究作为学科情报研究的一个方向,分析学科主题演化趋势,可以有效地揭示学科知识发展变化及其相互作用的特征和规律,帮助科研人员追溯学科发展轨迹,准确辨识缺失环节、薄弱环节或可能的新知识增长点,对决策层规划学科布局、调整学科方向和促进学科发展具有重要的参考价值。

学科主题演化是指以词语为表征的学科主题在时间维度上的发展变化过程,与空间变化相比,学科主题的时间演化体现的是学科主题的新陈代谢过程,体现了某一学科的发展态势和未来走向,是研究学科发展规律的重要内容[1]。

目前对学科主题演化分析的科学计量方法主要有词频分析法、共引分析法和共词分析法。相比于以单纯的关键词统计排序为主的词频分析法,共词分析法不仅能分析高频词,而且更关注这些词之间的联系,从而反映出概念之间的关系。相比于以文献作为分析对象、需要庞大的引文索引作为基础的共引分析法,共词分析法的优势在于能深入文献内部,以文献内部的关键词作为分析对象,从更微观的角度去揭示学科主题演化规律。因此,本文拟分析基于共词分析法的学科主题演化研究的进展情况。

笔者在先前的研究[2]中,利用自己构建的分类标准将共词分析法的研究分为了 5 类理论研究(分析对象的改进研究、指标改进研究、可视化方法调整研究、

作者简介: 唐果媛(0000-0001-8992-0230) ,硕士研究生,E-mail: tanggy@mail.las.ac.cn; 张薇,副主任,研究员。
收稿日期: 2015-01-07 修回日期: 2015-02-10 本文起止页码: 128-136 本文责任编辑: 杜杏叶

融合其他方法的研究及综述类研究）和4个层次（基于词、基于主题、时间维度和拓展）的应用研究，并在此基础上，分析了共词分析法在国际上和中国国内的研究现状。其中，基于共词分析法的学科主题演化研究属于共词分析法的第3个层次的应用研究，即时间维度上的应用研究。本文将在先前研究的基础上，通过人工判读中国国内的相关文献，进一步分析基于共词分析法的学科主题演化研究在中国国内的进展情况，以期为相关研究人员提供参考和借鉴。

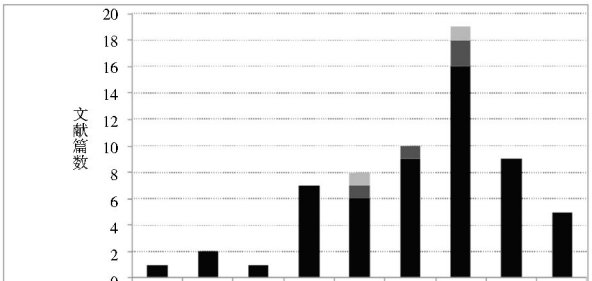
2 数据来源与基本情况

2.1 数据来源

本文的数据有两个来源：第一个来源为文献《国内外共词分析法的发展与分析》中中国国内共词分析法的第3个层次应用研究的文献，其获得的过程包括3个步骤：第一步是在中国知网（CNKI）数据库中，以高级检索“主题＝共词分析”或者“主题＝关键词 共现分析”，数据库选择中国学术期刊网络出版总库进行检索，共检索得到574篇中文文献，剔除重复文献和不相关文献，最后得到542篇文献，文献的时间跨度为1996－2014年，检索日期为2014年4月28日；第二步是利用人工判读法，根据研究人员应用共词分析法的不同角度，将研究共词分析法的文献划分为4个层次的应用研究（即基于词、基于主题、时间维度和拓展应用研究），本文选取第3个层次——时间维度上的应用研究的64篇文献；第三步剔除与学科主题演化不相关的文献，最后得到58篇基于共词分析法的学科主题演化研究的文献。由于CNKI数据库中不包括图书情报领域的核心期刊《情报学报》2002年以后的文献，因此本文数据的第二个来源为维普数据库中《情报学报》期刊中基于共词分析法的学科演化研究的文献，其获得过程包括3个步骤：第一步在维普数据库中，以“题名或关键词＝共词分析 或 关键词共现，文献来源＝情报学报”进行检索，共获得27篇文献；第二步通过人工判读，剔除不相关文献，最后获得4篇文献。因此，本文的分析数据集共有62篇文献。

2.2 基本情况

按照文献《国内外共词分析法的发展与分析》对共词分析法研究文献的分类标准，在这62篇基于共词分析法的学科主题演化研究的文献中，有2篇纯理论研究文献，4篇理论与应用研究相结合的文献，56篇纯应用研究文献。图1展示的是这3种类型研究文献数量的年度分布，发现，在中国国内将共词分析法应用于学科主题演化研究的文献始于2006年，且为纯应用研究文献，于2012年文献数量达到峰值。



	2006	2007	2008	2009	2010	2011	2012	2013	2014
纯理论研究					1		1		
理论＋应用研究					1	1	2		
纯应用研究	1	2	1	7	6	9	16	9	5

图1 各类型研究文献数量的年度分布

在6篇理论研究文献中，理论研究的内容包括：①指数的改进研究（3篇文献）。如赵凡[3]针对现有的3种基于共词分析的学科主题动态跟踪相似算法（非相似指数、影响和出处指数、相似指数）都存在不能深入分析主题演化细微关系的缺陷，以Coulter的相似指数为基础对现有相似算法进行了改进，同时阐释了改进相似算法中涉及的相关问题。叶春蕾等人[4]针对基于词频或共现词频的共词分析法难以反映主题词对间更深层次的语义关系这一情况，提出一种基于LDA模型和信息量的改进的共词分析法，该方法以信息量取代传统共词分析法中以词频或共现词频作为主题聚类的主要指标，其中的信息量是根据LDA模型中体现主题词、主题、文档之间的三层语义关系的三层概率来计算的。同时，文中还依据信息量提出了一种用来计算主题聚类簇之间关联强度的概率指数，以更精确、客观地揭示学科领域的演化规律。马晨峰等人[5]构建了战略坐标中主题聚类簇的新颖度和关注度两个指标。②可视化方法调整研究（2篇文献）。如秦长江[6]首次在国内用社会网络分析软件Pajek绘制出了类团关系图。李秀霞等人[7]将共词聚类、战略坐标和社会网络等3种可视化方法融合为一个综合知识图谱，通过对比分析两个阶段的综合知识图谱，全面、直观地揭示个性化信息服务（PIS）的研究热点、发展脉络及演化趋势。③分析对象的改进研究（1篇文献）。如杨颖等人[8]针对高频词阈值的确定尚未达成统一以及在聚类过程中没有形成中心词等问题，将共词分析方法作如下改进，依据Donohue提出的高频低频词界分公式来确定高频词，引入钟伟金提出的粘合力来确定每个聚类的中心词。

表1展示了基于共词分析法的学科主题演化研究的具体应用领域，发现其具体应用领域分布在七大学

科中，在农业学科中还没有得到应用。其中，文献数量最多的学科领域为信息科技领域，最少的学科领域为基础科学和工程科技领域。

表1 基于共词分析法的学科主题演化研究的具体应用领域

学科	文献篇数	具体应用领域
信息科技	31	图书情报(20)、信息服务(3)、计算机类(3)、知识管理(2)、知识联盟、档案学、信息资源管理
社会科学	8	卫生筹资研究、民办高等教育、独立学院研究、体育科学、心理学、远程教育、体育志愿服务研究、体育人文社会学
医药卫生科技	8	消化性溃疡、食品安全、医学信息学、丙肝研究、卫生资源分配、泌尿生殖器肿瘤领域、卒中意识与救治研究、先心病介入诊疗发展
经济与管理科学	6	中国经济问题、科学学、教育经济学、贸易与环境问题研究、人力资源管理研究、产学研研究
哲学与人文科学	3	人文学科、非物质文化遗产研究、心理学研究
基础科学	2	基因组学(2)
工程科技	2	材料科学、太阳能技术

注：括号中的数字表示该领域中的文献数量，无括号的表示文献数量为1

3 研究进展

3.1 基于共词分析法的学科主题演化研究分析流程

基于共词分析法的学科主题演化研究主要以关键词频次和共现频次量化计算为基础，根据关键词对间关联强度的大小聚集成主题簇，并计算不同时期主题的相似度，或者绘制不同时期的学科主题网络图谱，以分析学科主题的演化轨迹[4]。

笔者在判读本文构建的数据集的基础上，提炼出基于共词分析法的学科主题演化研究的分析流程(见图2)：首先确定学科领域的分析数据集，其次划分学科主题演化的阶段，然后从选取的数据集中选择和提取共词分析法的具体对象，接着利用选取的分析对象构建共词矩阵并进行归一化，最后采用一定的分析手段进行学科主题演化分析。接下来，笔者将对每个步骤中研究人员使用的策略、分析手段、方法和工具进行归纳总结。

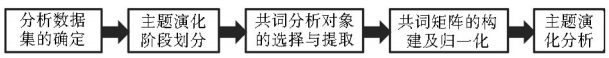


图2 基于共词分析法的学科主题演化研究分析流程

3.2 分析数据集的确定

进行学科主题演化研究的第一步是要确定所要分析学科的数据集。学科知识点的分布范围很广泛，主要集中在各类文献数据库中，同时还存在于大量灰色文献中。中国国内研究人员选择的数据源主要是各类文献数据库，包括：①综合类数据库，如中国知网数据库(30篇文献)、Web of Science数据库(10篇文献)；②专门类数据库，如生物医学领域的PubMed数据库(4篇文献)和中国生物医学文献数据库CBM(1篇文献)、农史领域的中国农史论文全文数据库(1篇文献)、经济学领域的Econlit数据库(2篇文献)；③引文类数据库，如中国社会科学引文索引数据库CSSCI(5篇文献)、全国报刊索引数据库(1篇文献)。在本文所构建的数据集中，王奉香等人[3]在利用共词分析法分析学科主题演化时选择的数据来源最多，包括4种综合类数据库、5种公共卫生类数据库、2种经济类数据库以及机构网站、灰色文献数据库和Google搜索引擎。

在构建数据集的过程中，研究人员考虑的主要因素是数据集的完整性和质量，采用的检索策略主要有3种：①基于词的检索策略，②基于期刊的检索策略，③复合检索策略。表2展示了这3种检索策略的文献数量、含义及特点。从表2可以发现，使用基于词的检索策略的文献数量最多，使用复合检索策略的文献最少。3种检索策略在数据集的完整性和质量这两个方面的侧重点有所不同，基于词的检索策略主要侧重于数据集的完整性，基于核心期刊的检索策略主要侧重于数据集的质量，而复合检索策略要么使得数据集更完整，要么兼顾数据集的完整性和质量。

3.3 学科主题演化阶段划分

对学科主题的演化分析可以是在时间序列上的逐年跟踪，也可以划分不同的时间段进行研究。前者可以及时、动态地反映学科领域发展的细节信息，后者则侧重于描述不同发展阶段的整体趋势。在本文构建的数据集中，有3篇文献采用了逐年跟踪的方法，有53篇文献采用了划分不同时间段的方法，划分的阶段数范围为2－6个，其中划分2个阶段的文献数量最多，占48.1%，其次为划分3个阶段的文献数量，占24.6%。关于学科主题演化阶段的划分没有统一标准，不同研究人员从各自的研究角度和研究目的出发对学科主题演化的发展阶段提出了不同的划分方法，笔者根据文献作者交代的学科主题演化阶段划分依据，以及通过观察划分阶段的规律，将学科主题演化阶段的划分方法分为：等距离定长法、科技文献增长规律法、学科发展历程法3种。还有13篇文献由于文献作者没有交代学科主题演化阶段的依据，其划分的阶段也无规律，故笔者没有给这13篇文献采用的主题演化阶段划分方法进行分类。

表2 3种检索策略的含义和特点

检索策略类型	文献篇数	含义	特点
基于词的检索策略	32	在选定的数据库中用关键词或主题词在标题、摘要、作者关键词、作者单位[10]等字段进行检索,并辅以其他手段进行精炼(如学科类别[11,d])	数据集的完整性较好
基于期刊的检索策略	15	选取学科领域部分期刊或核心期刊[12]作为数据源的检索策略	数据集的质量较高;数据集的完整性较弱
复合检索策略	12	指结合多种检索手段或策略,以确保数据集的完整性或者质量,包括:①主题检索和引文检索相结合[13];②特定学科数据库检索、期刊检索和主题词检索相结合[d];③主题词检索和分类号检索相结合[14-16];④主题词检索和引用次数相结合[17-18];⑤词检索和期刊检索相结合[19-20]	数据集更完整(如:①②③);兼顾数据集的完整性和质量(如:④⑤)

表3展示了学科主题演化阶段的3种划分方法,发现使用等距离定长方法的文献最多,其次为科技文献增长规律法,使用得最少的为学科发展历程法。等距离定长法的优点是操作简单,缺点是具有一定的主观性,缺乏系统严密的数据理论基础,若分段长度过大,则统计规律不明显;若分段长度过小,则在相邻阶段内,主题内容差异不大,很难得出正确结论。相对于等距离定长法,科技文献增长规律法和学科发展历程法具备了一定的理论基础,能够更科学地划分学科主题演化阶段。

表3 学科主题演化阶段的3种划分方法

学科主题演化阶段划分方法	文献篇数	含义	应用	特点
等距离定长法	31	将总时间段平均划分为若干个时间段,作为学科主题演化阶段,时间段可连续[21],也可不连续[22]	划分的时间段范围一般为2-6个;每阶段的年份数一般为2-6年和10年	操作简单;有一定的主观性;缺乏数据理论基础
科技文献增长规律法	11	根据学科领域科技文献数量的增长规律来划分学科主题演化阶段[3]	科技文献的增长规律如下:诞生期:文献数量不稳定增长;发展期:文献数量呈指数型增长;成熟期:文献数量增长缓慢,呈线性增长;饱和期:文献数量日趋减少[23]	以科技文献数量增长规律作为理论基础;主观性较弱;适合对学科的整个发展历史进行阶段划分
学科发展历程法	2	依据学科发展经历的不同时期来划分学科主题演化阶段	羌丽等[14]依据我国情报检索语言的发展经历的几个时期——建国后的发展期、文革中的瘫痪停滞期、改革开放后的繁荣期和网络时代的大变革期,采用共词分析法研究后两个阶段我国情报检索语言的研究热点	以学科发展经历的不同时期作为理论基础;划分的阶段所跨年份一般较长,不适合发展历史较短的新兴学科

3.4 共词分析对象的来源和选取

3.4.1 共词分析对象的来源 吴漂生研究发现[24],在大量同专业论文的关键词集合中,隐含着该学科的研究现状、研究热点、发展规律和发展趋势等线索。因此,共词分析法以能概括文献主要内容的关键词作为分析对象。关键词的来源有4种:一是作者提供的关键词,二是Web of Science数据库提供的增补关键词(Keyword Plus,是由Thomson Reuters创建的索引词,这些索引词来自正在索引的论文的作者所引用的论文的标题[25]),三是从标题或摘要抽取的关键词,四是主题词(在标引和检索中用以表达文献主题的规范化的词或词组[26])。

笔者统计了中国国内研究人员在利用共词分析法研究学科主题演化时选取的关键词来源,发现80%的文献(48篇)选择作者关键词作为共词分析对象的来源,11.7%的文献(7篇)选择主题词作为共词分析对象的来源,1.7%的文献(1篇)选择增补关键词作为共词分析对象的来源,8.3%的文献(5篇)选择的共词分析对象来源结合了2-3种关键词来源。研究人员确定共词分析对象来源的依据是什么呢?有部分选择作者关键词的研究人员明确提出,作者关键词是作者对研究内容的高度概括,是作者经过慎重考虑所做的选择,而且作者关键词主要以词组或者短语的形式存在,这些词的逻辑组合,能较好地揭示文献的主要内容[13];选择主题词作为分析对象的研究人员则认为[27],主题词由于来源于叙词表而较为规范。但是还有部分研究人员并没有交代确定共词分析对象来源的依据。为了探明其原因,笔者通过进一步探索发现,研究人员确定共词分析对象的来源与所选择的数据库提供的字段有很大的关系,主要表现在(见表4):将主题

词作为共词分析对象来源的 5 篇文献选择的 PubMed 数据库(4 篇)和 Econlit 数据库(1 篇)都提供了主题词字段;将作者关键词作为共词分析对象来源的 46 篇文献选择的数据库都提供了作者关键词字段;将增补关键词作为共词分析对象来源的 1 篇文献选择的 Web of Science 数据库提供了增补关键词字段。因此,笔者推测研究人员会参考所选数据库提供的关键词字段来确定共词分析对象的来源。

表 4 共词分析对象的来源及所选数据库

共词分析对象的来源	文献篇数	数据库(提供的关键词字段)
作者关键词	48	Web of Science(作者关键词、增补关键词);中国知网(作者关键词);万方数据库(作者关键词);CSSCI(作者关键词);维普数据库(作者关键词);中国科学引文数据库(作者关键词);中国农史论文全文数据库(作者关键词);中国生物医学数据库 CBM(作者关键词)
主题词	7	PubMed 数据库(主题词);Econlit 数据库(主题词、作者关键词);德温特创新索引数据库
增补关键词	1	Web of Science
多种来源关键词结合(作者关键词与从标题中抽取的关键词相结合[9]、作者关键词与增补关键词和从标题摘要抽取的关键词相结合[28])	4	Web of Science;中国知网;万方数据库

3.4.2 共词分析对象的选取 研究人员在选取共词分析对象时,要么是选取所有关键词进行共词分析[29],要么是选取高频词作为共词分析的对象。选取高频词作为共词分析对象,其目的是为了简化统计过程及减少低频词对统计过程的干扰。图 3 展示的是 2007 – 2014 年各年研究人员选取高频词作为共词分析对象所采用的方法。从总体来看,研究人员选取共词分析对象的方法主要基于 3 个角度,按出现时间早晚依次为关键词词频、关键词共现词频、关键词词频和共现词频相结合。

从时间序列上来看,2007 – 2011 年,研究人员主要从关键词的词频角度出发,将关键词按词频高低进行排序,采用经验判定法,在选词个数和词频高度上进行平衡,从前往后选取一定数量的关键词或者词频高于一定阈值的关键词作为共词分析的具体对象,选取的高频关键词数量一般保持在 40 – 70 之间[30],该方法仅依据研究人员的经验,主观性较强,缺乏理论的指导。2012 年,新增了两种选取共词分析对象的方法:一种是从关键词词频角度出发的高频低频词界分公式[31],该公式为 $T=\left(-1+\sqrt{1+8\times I_1}\right)/2$,其中 I_1 是出现一次的词的个数,T 为高频低频词词频临界值;另一种是从关键词共现词频的角度,采用经验判定法选取关键词共现词频超过一定阈值的关键词作为共词分析的对象[32]。2013 年,从关键词词频的角度增加了一种方法,即把文献的被引频次和关键词词频结合起来选取共词分析对象的方法[18]。2014 年,增加了两种方法:一种是从关键词词频的角度,利用关键词 g 指数来选取高频词作为共词分析对象[33],关键词 g 指数的定义[34]是包含该关键词的论文集合中,单篇引用次数最多的 g 篇论文总共获得不少于 g^2 次引用;另一种是从关键词词频和共现词频结合的角度,采用经验判定法,从高词频的关键词中选取共现词频超过一定阈值的关键词作为共词分析的对象[16]。总的来说,随着研究的深入,共词分析对象的选取方法在不断丰富和完善。

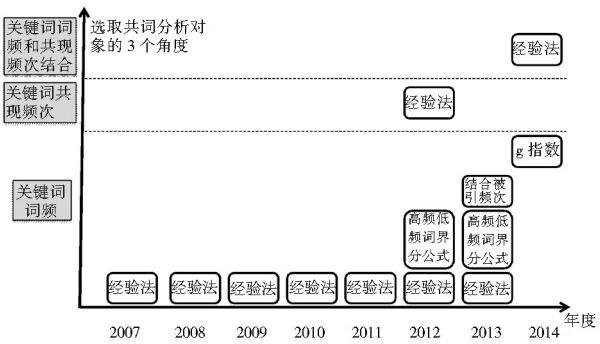


图 3 共词分析对象的选取方法

3.5 共词矩阵的构建及归一化

3.5.1 共词矩阵构建的方式 在本文构建的数据集中,研究人员在构建共词矩阵时采用了两种方式:一种是按照一定的取词门槛标准选出所有年份的高频词构建共词矩阵,得到学科领域在所有年份中的聚类主题,然后在每个时间段中选取一定数量的高频词,将每个时间段的高频词按照所有年份的聚类构成分别划分到该时段的每个聚类中;另一种是将每个时间段的关键词各自按照一定的取词门槛标准取出高频词,分别构建共词矩阵,得到学科领域在每个时间段的聚类主题。在本文构建的数据集中,只有 3 篇文献采用了第一种方式,其余文献均采用了第二种方式。这两种方式略有差异,也各有利弊。第一种方式取词统一,便于不同

时间段主题的比较;第二种方式可以更加真实地反映出学科领域研究主题的原貌。

3.5.2 共词矩阵的归一化 根据所选定的关键词构建基于关键词共现的共词矩阵,一组词共现的频率越高,则说明它们之间的关系越密切。但是在实际的计量化分析过程中,由于关键词出现的频次是绝对值,难以反映彼此之间真正的相互依赖程度,故此研究人员常利用一些特殊的相关系数(也称为归一化系数)将所得到的原始共词矩阵转换成相关矩阵,以便于之后的分析。

在本文所构建的数据集中,笔者统计了研究人员使用的归一化系数,发现有29篇文献指明使用了归一化系数,其中27篇文献明确指明了归一化系数的名称。在这27篇文献当中,共使用了6种归一化系数,使用得最多(16篇)的归一化系数公式为 $Oij=\frac{Cij}{\sqrt{Ci}\sqrt{Cj}}$,其中,Cij表示关键词对i和j在文献集合中的共现频次,Ci表示关键词i在文献集合中出现的频次,Cj表示关键词j在文献集合中出现的频次,研究人员对这个公式有3种不同的叫法:Ochiai系数、Salton系数、余弦指数;其次为基于现代统计软件的Pearson系数(4篇);排在第3位(4篇)的为国外研究人员早期提出的等价系数 $Eij=\frac{Cij}{Ci}\times\frac{Cij}{Cj}$;另外,还有3种归一化系数各出现了一次,分别为Jaccard系数[22]($Jij=\frac{Cij}{Ci+Cj-Cij}$)、TF/IDF[13]、泊松相关系数[35]。

通过以上统计分析,发现共词矩阵的归一化系数种类较为丰富,但研究人员相对来说比较偏向于使用Ochiai系数。

3.6 主题演化分析

在构建共词矩阵的基础上,研究人员采用了不同的分析手段来研究主题演化情况。笔者通过判读本文构建的数据集,统计了研究人员在研究主题演化情况时采用的分析手段,主要包括聚类分析、因子分析、战略坐标、类团分析、多维尺度分析、社会网络分析图谱和社会网络属性分析7种:①聚类分析,是依据关键词与关键词之间的共现强度,把共现强度较大的关键词聚集在一起形成聚类簇(即主题),用到的聚类方法有:系统聚类法(使用得最频繁)、Callon方法[6]、Coulter方法[28]等。②因子分析,其原始目标是用尽可能少的因子去描述众多的指标,使得较少的几个公共因子可以反映原始资料的大部分信息;在主题演化分析中,主要是依据因子个数碎石图,辅助聚类分析确定最佳分类数[36]。③战略坐标,是在聚类分析的基础上开展的,可以清晰地反映出主题的中心度和密度,便于确定研究热点。④类团分析,亦是在聚类分析结果的基础上开展的,用可视化的方法来展示类团(即主题)在一定时间内的组成、演化、消失及增长[33]。⑤多维尺度分析[19],与聚类分析和因子分析配合起来,利用平面距离展示词间亲疏关系,能够判断出某主题在学科领域中的位置。⑥社会网络分析图谱,指利用社会网络分析工具通过节点—链接图直观、形象地反映词间联系的强弱,快速定位核心词和边缘词。⑦社会网络属性分析,是指对共词网络的个体属性[37](如:点度中心度)或整体属性[29](如:平均距离、聚集系数)进行分析,以展示共词网络的演化过程。

图4展示的是基于共词分析法的学科主题演化的分析手段及年度分布。从总体来看:聚类分析和社会网络分析图谱是最常用的分析手段;其次为战略坐标;使用得最少的分析手段为类团分析和社会网络属性分析。从时间序列上来看:2007年,研究人员使用的分析手段包括因子分析、聚类分析、战略坐标和多维尺度分析;2009年,研究人员开始使用社会网络分析图谱;2010年,研究人员开始使用类团分析;2011年,研究人员开始使用社会网络属性分析;2012-2014年,没有增加新的分析手段。

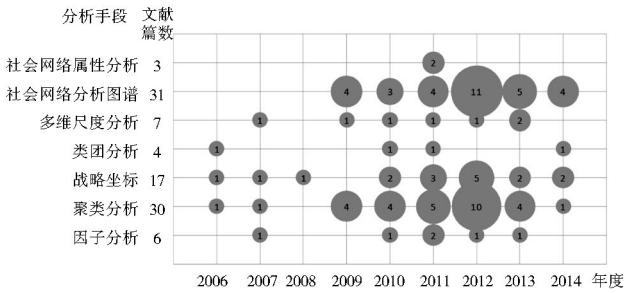


图4 基于共词分析法的学科主题演化的分析手段及年度分布

3.7 分析工具

研究人员在利用共词分析法研究学科主题演化时,常常需要借助一定的分析工具。笔者统计了国内研究人员在利用共词分析法研究学科主题演化时所使用的12种分析工具,其能用来实现切分词、统计词频、构建共词矩阵、因子分析、聚类分析、多维尺度分析、共词网络可视化和共词网络属性计算等8种功能。

图 5 展示了基于共词分析法的学科主题演化研究的分析工具及使用的功能，从总体上来看，使用最频繁的工具为 SPSS，其次为 Ucinet。

从横向来看，用来切分词的工具有中国科学院计算技术研究所研发的汉语词法分析系统（ICTCLAS 系统）；用来统计词频的工具有 6 种，分别为 ICTCLAS 系统、Excel、武汉大学 ROST 虚拟学习团队开发的 ROST 内容挖掘系统、瑞典科学家佩尔松开发的 Bibexcel、中国医科大学医学信息学系崔雷教授开发的书目共现分析系统（BICOMB）和汤姆森路透公司开发的汤姆森数据分析软件（TDA）；用来构建共词矩阵的工具有 4 种，分别为 Bibexcel、BICOMB、TDA 和汤姆森路透公司开发的文献管理软件 Endnote X4；用来进行因子分析的工具有 SPSS；用来进行聚类分析的工具有 SPSS 和美国德雷塞尔大学陈超美教授开发的文献分析工具 CiteSpace；用来进行多维尺度分析的工具有 SPSS 和加州大学欧文分校的林顿·费里曼教授编写的 Ucinet；用来把共词网络可视化的工具有 4 种，分别为 Ucinet（集成的可视化工具 Netdraw）、CiteSpace、V. Batagelj 和 A. Mrvar 共同编写的 Pajek 软件、印第安纳大学伯明顿分校的图书情报专家 K. Börner 及其团队研发的知识图谱工具 Sci2；用来计算共词网络属性的工具有 CiteSpace。

从纵向上来看，运用每种分析工具来实现的功能不多，最多为 3 项，如使用 SPSS 进行因子分析、聚类分析和多维尺度分析，使用 CiteSpace 进行聚类分析、共词网络可视化和共词网络属性计算。在利用共词分析法研究学科主题演化时，需要同时借助多种分析工具才能完成全部的分析流程。

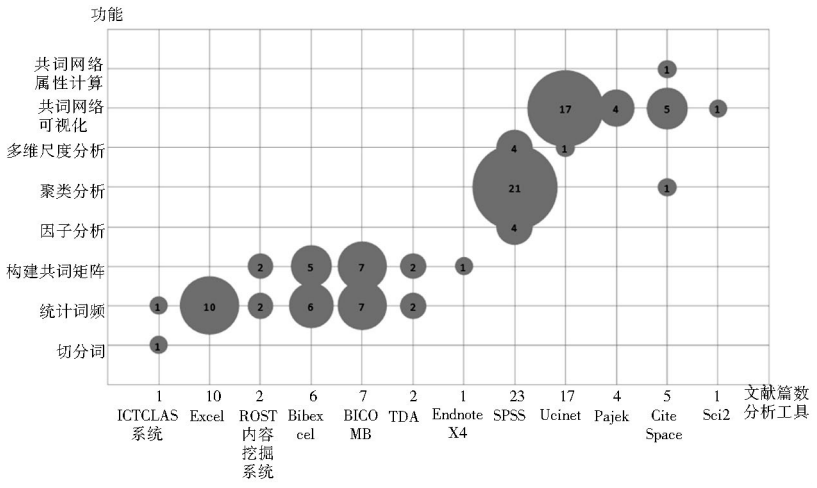


图 5 基于共词分析法的学科主题演化研究的分析工具及使用的功能

4 结语

分析学科主题演化趋势，对科研人员研究学科知识、决策层规划学科布局都有重要意义。本文主要基于笔者提炼的学科主题演化研究分析流程的 5 个步骤和使用的分析工具，分析了基于共词分析法的学科主题演化研究在中国国内的进展情况，得到以下结果：①分析数据集的确定。数据集的来源数据库有三大类（综合类数据库、专门类数据库和引文类数据库），检索策略有 3 种（基于词的检索策略、基于期刊的检索策略和复合检索策略）。②学科主题演化阶段划分。划分方法没有统一标准，使用较多的方法有 3 种：等距离定长法、科技文献增长规律法和学科发展历程法，并总结了 3 种方法的应用及特点。③共词分析对象的来源和选取。来源有 4 种：作者关键词、增补关键词、从标题或摘要抽取的关键词以及主题词；分析对象的选取主要基于 3 个角度：关键词词频、关键词共现词频、关键词词频和共现词频相结合。④共词矩阵的构建及归一化。研究人员在构建共词矩阵时采用了两种方式，主要采用的是对每个时间段的关键词分别构建共词矩阵；研究人员共使用了 6 种归一化系数，使用得最多的为 Ochiai 系数（亦被称为 Salton 系数、余弦指数）。⑤主题演化分析。采用的分析手段包括 7 种，最常用的分析手段为聚类分析和社会网络分析图谱。⑥使用的分析工具共有 12 种，其中使用最频繁的工具为 SPSS。

综上所述，利用共词分析法来分析学科主题演化研究在中国国内已发展得比较成熟，形成了一个相对完善的分析流程。在每个分析流程上，研究人员选择的策略、分析手段以及使用的工具等呈现多样性。特别是在主题演化分析步骤中，可视化手段非常丰富。因此，笔者推测，随着社会网络技术的发展，可视化软件逐步增多，未来，基于共词分析法的学科主题演化的研究会在社会网络和可视化方面有进一步的发展。

但是，在利用共词分析法来探索学科主题演化规律时，也存在一些不足之处，如：①剔除低频关键词或者共现强度弱的共词对，将不利于探测潜在的主题或处于上升期的主题，而且，在剔除低频词或低共现共词对时，没有统一的标准，存在一定的主观性。②共词分析法把每个关键词都视为同等重要，但实

际上属性不同的关键词(主要关键词和次要关键词)的重要程度是不同的,揭示主题的能力也有差异。可见,共词分析法在探索学科主题演化规律的应用中还需进一步完善。

参考文献:

[1] 王春秀,冉美丽. 学科主题演化定量分析的理论基础探析[J]. 现代情报,2008(6):48-50.

[2] 唐果媛,张薇. 国内外共词分析法的发展与分析[J]. 图书情报工作,2014,58(22):138-145.

[3] 赵凡. 基于共词分析的学科主题动态跟踪相似算法改进研究[J]. 情报杂志,2010,29(1):173-176.

[4] 叶春蕾,冷伏海. 基于共词分析的学科主题演化方法改进研究[J]. 情报理论与实践,2012,35(3):79-82.

[5] 马晨峰,谷祖莎,沈君. 我国贸易与环境问题研究的文献计量分析——基于聚类和战略坐标方法的对比分析[J]. 科技管理研究,2013(17):227-232.

[6] 秦长江. 基于共词知识图谱的人文学科研究热点可视化的实证研究[J]. 图书馆理论与实践,2010(12):29-33.

[7] 李秀霞,邵作运,郑春厚. 我国图书情报界PIS研究的共词可视化分析[J]. 情报杂志,2012,31(8):109-113.

[8] 杨颖,崔雷. 应用改进的共词聚类法探索医学信息学热点主题演变[J]. 现代图书情报技术,2011(1):83-87.

[9] 王奉香,刘彩,王敏,等. 国外卫生筹资研究文献主题分布及变化趋势[J]. 中国卫生经济,2009,28(5):24-27.

[10] 杨立英. 基因组学领域演进的科学计量研究[J]. 科学观察,2007,2(1):11-19.

[11] 周卫华,林彦汝. 近十年国际知识管理领域的研究态势——基于SSCI的文献计量分析[J]. 新世纪图书馆,2012(9):33-37.

[12] 王红. 近十年我国图书情报学科研究热点的共词分析[J]. 情报学报,2011,30(7):765-775.

[13] 任红娟,张志强. 基于文献计量的科学知识图谱发展研究[J]. 情报杂志,2009,28(12):86-90.

[14] 羌丽,屈卫群. 我国情报检索语言研究透视——以共词分析为方法[J]. 图书馆杂志,2010,29(10):10-15.

[15] 王琪,徐成立. 知识图谱视野下我国体育科学研究的发展路径——基于1991-2009年体育学博士论文关键词共词网络的可视化分析[J]. 体育学刊,2010,17(12):118-125.

[16] 姜霖,王子朴,王晓虹. 基于CSSCI的体育人文社会学论文关键词分析[J]. 西南民族大学学报(人文社会科学版),2014(1):229-238.

[17] 赵丽梅,张庆普. 我国知识管理研究前沿演进趋势知识图谱[J]. 科学学与科学技术管理,2012,23(1):90-98.

[18] 马海群,姜鑫. 我国档案学研究热点与前沿演进的知识图谱分析[J]. 档案学研究,2013(4):16-22.

[19] 樊霞,吴进,任畅翔. 基于共词分析的我国产学研研究的发展态势[J]. 科研管理,2013,34(9):11-18.

[20] 蔡治东,虞荣娟,汤际澜. 知识图谱视野下我国体育志愿服务研究热点综述[J]. 体育科技,2014,35(1):5-7,10.

[21] 吴潇泽,王小华,谌志群. 基于共词分析的科技文献趋势挖掘[J]. 计算机系统应用,2011,20(4):59-63.

[22] 张晗,王晓瑜,崔雷. 共词分析法与文献被引次数结合研究专题领域的发展态势[J]. 情报理论与实践,2007,30(3):378-380,426.

[23] 庞景安. 科学计量研究方法论[M]. 北京:科学技术文献出版社,1999:299-300.

[24] 吴漂生. 从关键词词频看我国读者工作的发展[J]. 现代情报,2005(10):28-31.

[25] Web of Science[EB/OL]. [2015-02-10]. http://images.webofknowledge.com/WOKRS5161B5_fast5k/help/zh_CN/WOS/hp_full_record.html.

[26] 百度百科——主题词[EB/OL]. [2015-02-10]. http://baike.baidu.com/link?url=_DRzKOQJrBFp5Hf4HtC3mvovAd0TQyQGGANlv6KAENIxG4lxNmEsoi-wqnHr0I2w.

[27] 蒋颖,魏众. 中国经济转型与发展的国际研究——基于ECONLIT数据库的共词分析[J]. 经济学动态,2008(8):76-83.

[28] 韩红旗,安小米. 科技论文关键词的战略图分析[J]. 情报理论与实践,2012,35(9):86-90.

[29] 林德明,刘则渊. 复杂网络研究领域演进中的复杂性[J]. 数学的实践与认识,2011,41(17):174-182.

[30] 肖明,杨楠,李国俊. 基于共词分析的我国用户信息行为研究结构探讨[J]. 情报杂志,2010,29(S2):12-15,26.

[31] 于跃,潘玮,王丽伟,等. 人类基因组测序文本数据挖掘研究[J]. 医学信息学杂志,2012,33(4):39-44.

[32] 周玉芳. 知识图谱视野下科技查新研究的发展分析[J]. 现代情报,2012,32(6):25-28,32.

[33] 冯佳,张云秋. 国内泌尿生殖器肿瘤领域研究热点分析[J]. 医学信息学杂志,2014,35(1):41-47.

[34] 吴明智,高硕,杨错. 基于关键词词频和g指数的高校图书馆学科服务研究热点分析[J]. 医学信息学杂志,2013,34(1):61-64.

[35] 黄维,陈勇. 中国教育经济学发展轨迹的知识图谱研究——基于《教育与经济》所载论文的关键词共词分析[J]. 教育与经济,2010(3):68-72.

[36] 王涓. 2000-2009年国内心理学论文研究热点的计量分析[J]. 心理科学,2011,34(5):1209-1215.

[37] 刘竟,王慧. 近十年我国搜索引擎研究的可视化分析[J]. 图书情报研究,2011,4(4):37-42.

作者贡献说明:

唐果媛:提出论文研究框架,判读相关文献并提取信息,进行数据统计与分析,撰写论文;

张薇:修正论文研究框架,指导论文修改。

Development and Analysis of Subject Theme Evolution Based on Co-word Analysis Method

Tang Guoyuan[1,2] Zhang Wei[1]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] University of Chinese Academy of Sciences, Beijing 100049

Abstract: [Purpose/significance] It is significant to analyze the trend of subject theme evolution for researchers researching subject knowledge and policy makers planning subject layout. Compared to the word frequency analysis method and co-citation analysis method, co-word analysis method has the advantage of being able to deep inside the literatures, and revealing the rules of subject theme evolution. This paper studies the development of subject theme evolution based on co-word analysis method at home, so as to provide reference for the related researchers. [Method/process] This paper uses manual interpretation method to extract five steps of research process of subject theme evolution based on co-word analysis. Furthermore, it summarizes the strategies, analysis methods or tools of the researchers using in each step. [Result/conclusion] There are three kinds of main source databases of data sets: comprehensive database, special database, and citation database. There are three kinds of retrieval strategies: based on the words, based on the journals and composite retrieval strategy. The main source of co-word analysis object is author keyword. There are three kinds of main perspectives: based on keywords frequency, based on keywords co-occurrence frequency, and combination the previous two kinds. The Ochiai coefficient is used most widely of normalized coefficients when the co-word matrix is built. Clustering analysis and social network analysis map are used most widely of subject theme evolution analysis methods. SPSS software is used the most frequently.

Keywords: subject theme evolution co-word analysis theme evolutionary stage search strategy

2014年度“复印报刊资料”转载学术论文指数排行榜及重要转载来源期刊发布

2015年3月31日下午，中国人民大学人文社会科学学术成果评价发布论坛在中国人民大学逸夫会议中心召开。会议由中国人民大学主办、中国人民大学书报资料中心和中国人民大学人文社会科学学术成果评价研究中心承办。中宣部、教育部、国家新闻出版广电总局、全国人大、中国人民大学有关部门的领导出席了会议。人文社科成果评价研究界专家学者、学术期刊与学术机构代表约400人参会。论坛发布了由中国人民大学人文社会科学学术成果评价研究中心与中国人民大学书报资料中心联合研制的《2014年度“复印报刊资料”转载学术论文指数排名》和《2014年版“复印报刊资料”重要转载来源期刊》两项成果。

“复印报刊资料”的转载量、转载率、综合指数等已逐渐被学术界和期刊界视为人文社科学术评价的参考依据之一，越来越多的科研机构以此作为评价论文、作者、期刊、机构水平和影响力的重要参数；也成为北京大学《中文核心期刊要目总览》、中国社科院《中国人文社会科学核心期刊要览》、教育部“名刊名栏”工程，武汉大学中国科学评价研究中心大学排名、期刊排名的重要指标之一。“复印报刊资料”强调“质量为本”的论文评价和精选实践，为我国人文社科成果评价提供了有价值的新视角。

2014年度转载学术指数排行榜，共有约600种人文社科期刊和300家作者机构榜上有名。在图书馆学情报学学科，《图书情报工作》被转载量名列第一，综合指数排名第二。