基于文献计量的科技监测方法与应用系统比较研究*

朱亮，孟宪学，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：近年来，科技监测理论及应用发展迅速，作为科技监测方法体系最重要的组成部分，文献计量学方法在科技监测领域相关的应用成果已较多。本文首先对几种常见的基于文献计量的科技监测方法进行了总结，其次，对现有部分常用的科技监测应用系统进行了分析与比较，以期为今后的相关研究提供参考。

关键词：科技监测；文献计量；引文分析；共词分析
中图分类号：G257
DOI：10.3772/j.issn.1673—2286.2015.01.010

引言

计算机及互联网技术的飞速发展，推动了各类科技信息资源的迅猛增长，导致信息过载，海量信息资源的复杂性远远超出了人们的理解能力，为科研人员准确把握研究领域的结构和演变带来了困难，阻碍了科学研究的发展。如何科学地梳理一个学科领域的发展历程、认识其发展趋势、追踪其研究热点与前沿，从而帮助科研人员找到研究的创新突破口、发掘潜在研究空间，这便为现代图书情报研究提出了新的要求。为此，科技监测（Science and Technology Monitoring）应运而生。科技监测是指以科学技术信息、数据分析为基础，以数据挖掘（Data Mining）、信息萃取（Information Extraction）、知识发现（Knowledge Discovery）、可视化技术（Visualization）等信息科学前沿技术为手段，对科学技术活动进行动态监测、分析及评估的方法[1]。从方法体系来看，科技监测采用的方法各式各样，各种方法的针对性与侧重点不一，有些方法针对研究热点与前沿的分析和识别，有些方法则侧重于领域新兴趋势的发现。在这些方法中，文献计量学方法占据了主流地位，相关理论及应用成果已较多，本文将重点对基于文献计量的科技监测方法及应用系统进行分析和比较。

1 基于文献计量的科技监测方法分析

1.1 基于文献外部特征统计的科技监测方法

文献外部特征统计是文献计量方法体系的重要组成部分，也是科技监测最常用的方法之一，通过对题名、关键词、作者等文献外部特征值进行统计分析，可

* 本研究得到国家“十二五”科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与示范应用”课题“基于STKOS的知识服务应用示范”（编号：2011BAH10B06）资助。

从时间、空间角度对学科领域的发展及演化情况进行宏观地解释。这其中最典型的当属词频分析，词频分析是利用能够揭示或表达科技文献核心内容的关键词或主题词在某一研究领域科技文献中出现的频次高低来确定该领域研究热点和发展动向[2]。其依据是一篇科技文献的关键词或主题词是文章核心内容的浓缩和提炼，因此，如果某一关键词或主题词在其所在学科领域的文献中反复出现（超过给定阈值，即高频词），则可反映出该关键词或主题词所表征的研究主题是该学科领域的研究热点。

这种方法虽然简单易行，但其不足也很明显，主要表现在词频阈值确定缺乏科学统一标准，主观性较强，且将有一定集合意义、可能代表研究热点、新研究方向的低频词汇排除在分析对象之外，将对科技监测结果造成影响。

1.2 基于引文分析的科技监测方法

引文分析是利用图论、模糊集合、数理统计方法以及比较、归纳、抽象、概括等逻辑方法，对科学期刊、论文、著者等各种分析对象的引用或被引用现象（即引证关系）进行分析，以揭示其数量特征和内在规律，评价、预测科学发展趋势[3]。因此，可以说是科技文献间的相互引证关系催生了引文分析。引文分析包括直引分析、共引分析和文献耦合分析。当前，引文分析常被国内外学者用于描绘科学结构的发展历程、评价科研成果及科研人才、追踪学科发展动向与趋势等[4]。

共引分析（Co-citation Analysis）是一种重要的引文分析方法，共引又称同被引，若两篇文献同时被其他n篇文献所引用，则称这两篇文献具有共引关系，其同被引次数（即共引强度）为n。按分析单元的不同，共引分析主要有两个系列：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（Author Co-citation Analysis，ACA）。Small认为高被引文献代表了特定的发现、方法，或是施引文献所共同认可的概念。基于这种观点，Small等人开发了单机系统SCI-Map来描绘科学文献间的结构及变化、分析科学研究前沿等[5]。同样，由于科技文献是作者研究成果最重要的表达和体现形式，通过对特定研究领域内作者的引用和被引关系进行分析，可实现对该领域智力结构的概括和描绘。1981年，White在其发表的《作者共引：科学结构的文献测量方法》中首次提出ACA[6]，1990年，McCain将ACA程序归纳为选择作者、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[7]，即传统ACA模式，针对该模式计算强度大等不足，White于2003年采用网络寻址定位（Pathfinder Network Scaling，PFNETs）技术对其进行了改进，进一步提高了其分析结果的可信度。

共引分析只能从外部特征间接描述科技文献内容的变化，分析深度有限。在用文献共引分析进行科学前沿和热点分析时，分析结果可能会漏掉一些由于新出现而未得到高被引的研究前沿领域，此外，传统的ACA都是针对第一作者进行的，无疑会使分析结果在一定程度上失真。正因如此，在实际使用中，共引分析常与其他方法联合使用，以提高分析结果的科学性。

1.3 基于共词分析的科技监测方法

共词分析属于内容分析方法的一种，其原理是对一组词两两统计它们在同一篇文献中出现的次数，以此为基础对这些词进行聚类分析，从而反映出这些词之间的亲疏关系，进而分析这些词所代表的学科和主题的结构变化[8]，实现对学科内部结构关系及学科间联系的描述，以及不同时期学科发展和相互间交叉、渗透趋势的揭示。共词分析的对象通常是文献的关键词或主题词，其分析的第一步就是从相关文献集中抽取关键词或主题词，一般是出现频次超过一定阈值，并且能够代表该学科研究主题或研究方向的高频词。其次，两两统计这些高频词在同一篇文献中同时出现的次数，形成共词矩阵。最后，围绕着这个共词矩阵进行分析[9]。

相较于共引分析，共词分析最显著的优势是能够深入文献实际内容开展研究。但与共引分析一样，共词分析也不可避免的存在着一些问题，如共词分析是基于词的分析，而词（尤其关键词）的选择带有很强的作者主观性和随意性，且共词分析是以单个概念作为分析对象，脱离了词汇的上下文关系，也就无法准确表示词汇的语义关系[10]。因此，越来越多的学者将共引分析与共词分析结合使用，发挥两种方法的互补优势。

1.4 基于爆发词的科技监测方法

魏晓俊将词频分析、基于词网络关系的共词分析、基于词频变化率的突发监测等均归纳为基于词语的科技监测方法[11]。其中突发监测算法是Kleinberg在2002

年提出的[12]，其主要思想是关注并寻找那些在一段时间内突然增长的词，即爆发词。爆发词作为热点问题的直观表现，在文献情报研究领域，有效识别并处理爆发词对科技热点监测、研究机会发现和科研趋势预测都具有重要意义。

在爆发词的探测方面，洪娜等总结了当前几种代表性方法，包括基于文档流上的词频统计、基于有限状态自动机的突发监测、基于卡方统计的热点词计算、基于文档聚类和词聚类方法的热点发现等[13]。各种方法特点不一，也有相应的局限性，如基于文档流上的词频统计方法虽操作简单，但其仅对一段时间中词的状态进行简单统计，对词的变化反应不敏感；Kleinberg提出的突发监测自动机模型在捕捉低频突发词时具有更好的效果，直接用于爆发词探测时误报率较高。因此，在实际应用中，爆发词的探测不仅要考虑词频，还要结合词所处的语义环境等，综合从词爆发的多方面特征来判别。

2 现有科技监测应用系统分析与比较

现有科技监测系统大体可分为两类：一是基于引文数据分析的监测系统，主要利用WOS等带有引文信息的文献数据进行聚类分析，识别研究前沿，探寻发展脉络，如CiteSpace、HistCite等；二是基于数据挖掘技术的监测系统，通过文献数据的主题表示、识别和聚类，发掘主题间的隐含关系，如PROTEJ、In-SPIRE等。文献计量方法在这些系统中均有不同程度的应用，最常用的当属前文所述的词频统计、共引分析、共词分析等。

2.1 HistCite

1955年，加菲尔德提出了利用相互引用关系分析科学文献的思想[14]，以此为理论基础，2001年，用来直观反映学科领域在某一阶段的重要文献及它们之间的引用关系的引文编年可视化系统HistCite[15]正式问世。HistCite处理的数据主要是来自于Web of Knowledge，数据中的每条记录都详细标明了当前文献引用的文献和被其他文献引用的次数。HistCite可以输出重要文献、作者和期刊等多种列表，从而帮助研究人员快速了解学科领域发展的起源、经典文献、知名专家和学者、重要研究机构等。

作为一个引文分析可视化系统，HistCite的最主要功能是生成引文编年图。HistCite可以选取文献集合中GCS（文献在整个Web of Knowledge数据库中的被引频次）或LCS（文献在被统计文献集中的被引频次）超过用户确定阈值的文献，并根据时间先后顺序生成编年图。通过引文编年图，研究人员可以观察到学科领域发展的沿革和继承关系，以及在某一阶段的发展程度。

2.2 CiteSpace

CiteSpace[16]是由美国德雷塞尔大学陈超美博士开发的一款信息可视化软件，主要用于对特定领域文献（集合）进行计量，以探寻学科领域演化的关键路径及转折点，探测学科领域研究前沿等。CiteSpace分析的数据主要来源于Web of Knowledge和PubMed，目前也已支持对CSSCI、CNKI等中文引文数据的分析。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[17]。为提高研究前沿揭示的时效性，CiteSpace采用爆发词算法来辨认研究前沿专业术语概念。在CiteSpace中，研究前沿是基于从文献题目、摘要、索引词和文献记录标识符中提取出的突变专业术语而确定的。运行软件可生成由文献共引网络和施引文献共词网络共同构成的混合知识图谱，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。

2.3 In-SPIRE

In-SPIRE[18]是由美国太平洋西北国家实验室开发的一个可视化工具包，使用地形图发现文献之间的关系和相似文献集。In-SPIRE能够揭示普通主题和一个大型文献集的相关隐藏关系，并基于词的分布、频率和与其他关键词的相似度来可视化文本文献。IN-SPIRE可以根据用户指定的列来判断文档之间的相似性。在相似度计算完成之后，IN-SPIRE运行聚类算法生成若干主题（文献的集合），每个主题的名字是最频繁出现在文献中的关键词。IN-SPIRE提供Galaxy和ThemeView两种不同的视图，将主题看成沉积层，它们一起构建起自然地貌，其山峰高度表示该领域的主题强度。

2.4 PROTEJ

生物医学主题监测和追踪系统PROTEJ[19]由美国Berkeley大学研发，主要实现对生物医学领域文献信息的主题监测和追踪，从而帮助用户把握领域研究现状和趋势。PROTEJ每30分钟更新一次生物医学文献数据，然后经过XML解析、特征抽取、维度缩减、聚类分析、主题识别等一系列工作流程，最终将属于某主题的新文献识别出来并通过Email发送给目标用户。

2.5 科技监测系统比较

现有科技监测应用系统各有特点，有的侧重于文献关联分析，有的则专注于文献内容分析。从所采用的方法体系来看，每个监测系统所含文献计量特征的程度也不尽相同。此外，为提高监测与分析结果的可读性与直观性，绝大多数监测系统均实现了对结果信息的多维度可视化展示。本文主要从数据源、分析字段、文献计量学方法、可视化图形等方面对HistCite等6个系统进行比较分析，详见表1。

从上表可以看出，文献计量学方法在各个监测系统中得到了不同程度的应用，足见其在科技监测方法体系中的重要地位。总体来看，现有绝大多数科技监测系统具备了结构化操作、动态交互等特点，但也还存在一些需要改进的方面，如数据源要求偏高，来源单一，引文数据多选择Web of Science数据库；系统辅助功能有待加强，如提供使用细则、参考实例等以帮助用户快速掌握系统使用方法。

表1 科技监测应用系统综合比较表

名称	主要数据源	分析字段	文献计量学方法			可视化图形
			特征统计	引文分析	共词分析	
HistCite	Web of Knowledge	引文	√	√		引文编年图
CiteSpace	Web of Knowledge、PubMed	题名、摘要、索引词等	√	√	√	轴节点图
In-SPIRE	XML、文本型数据	文本全文	√		√	星系山脉图
PROTEJ	XML、文本型数据	题名、摘要	√		√	文献列表
VxInsight	Web of Knowledge	摘要	√		√	三维山脉图
ThemeRiver	XML、文本型数据	文本全文	√		√	河流图

3 结束语

本文重点对基于文献计量的科技监测方法和应用系统进行了分析与比较，但科技监测的方法体系还包括了许多其他内容，如复杂网络理论、主题模型分析等。随着科技监测理论及应用研究的深入，特别是越来越多的学者将网络信息资源纳入科技监测的范围，将文献计量学方法与其他技术方法有机融合，发挥各自所长，将是今后科技监测领域需要重点研究的一个内容。

参考文献

[1] 朱东华,袁军鹏.基于数据挖掘的科技监测方法研究[J].管理工程学报,2004, 18(4):135-139.

[2] 马费成,张勤.国内外知识管理研究热点——基于词频的统计分析[J].情报学报,2006, 25(2):163-171.

[3] 耿海英,肖仙桃.国外共引分析研究进展及发展趋势[J].情报杂志,2006, 25(12): 68-69,72.

[4] 杨微微,吕娜.基于文献计量学的科技监测理论研究[J].情报杂志,2011,30(10): 21-24,48.

[5] Small H. A SCI-Map case study: Building a map of AIDS research[J]. Scientometrics,1994, 30(1):229-241.

[6] 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.

[7] 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.

[8] 蒋颖.1995-2004年文献计量学研究的共词分析[J].情报学报,2006, 25(4) :504-512.

[9] 肖伟,魏庆琦.学术论文共词分析系统的设计与实现[J].情报理论与实践,2009,32(3): 102-105.

[10] 安新颖,钟华.科技监测的理论综述与应用系统对比分析[J].情报理论与实践,2010,33(5): 124-128.

[11] 魏晓俊.基于科技文献中词语的科技发展监测方法研究[J].情报杂志,2007, 26(3) :34-36,39.

[12] KLEINBERG J. Bursty and hierarchical structure in streams[C]. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002:1-25.

[13] 洪娜,张智雄,等.基于决策树的潜在爆发词探测方法[J].情报学报,2012,31(3) :228-241.

[14] Garfield E. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas[J]. Science,1955,122(3159):108-111.

[15] HistCite [EB/OL][2014-12-21]. http://interest.science.thomsonreuters.com/forms/HistCite/.htm.

[16] Chen C. Searching for intellectual turning points: Progressive Knowledge Domain Visualization [J]. Proc. Nat. Acad. Sci.,2004,101 (Suppl.):5303-5310.

[17] 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.

[18] In-SPIRE [EB/OL] [2014-12-21]. http://in-spire.pnnl.gov/.htm.

[19] PROTEJ[EB/OL] [2014-12-21]. http://courses.ischool.berkeley.edu/i256/f06/projects/ye_chen_nguyen_presentation.pdf.

作者简介

朱亮，男，1981年生，助理研究员，研究方向：文献计量、情报分析研究，Email：zhuliang@caas.cn。

Review of Science and Technology Monitoring Method and Application Systems Based on Bibliometrics

ZHU Liang, MENG XianXue, ZHAO RuiXue, KOU YuanTao, XIAN GuoJian
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: In recent years, the science and technology monitoring theory and application developed rapidly. As the most important part of method system of science and technology monitoring, many related application achievements base on bibliometrics method has been reached. This paper first summarizes several common science and technology monitoring method based on bibliometrics, secondly, analyzes and compares the existing part of science and technology monitoring application systems, hopes to providing reference for related research in the future.

Keywords: Science and technology monitoring; Bibliometrics; Citation analysis; Co-word analysis

（收稿日期：2015-01-04）
编辑：刘伟