●刘 静[1,2]，马建霞[1]，范云满[1,2]
（1. 中国科学院 国家科学图书馆兰州分馆，兰州 730000；2. 中国科学院大学，北京 100190）

研究前沿探测方法概述

[关键词] 研究前沿；定性；定量；引文分析；词汇分析；共词分析

[摘 要] 在对研究前沿相关概念进行界定的基础上，从定性、定量等角度总结归纳了当前在探测研究前沿的方法，重点分析了基于引文的分析法和基于词汇的分析法，讨论了共词分析法的改进与完善。通过全面的归纳分析，总结比较了诸多方法的异同，提出对研究前沿探测方法的建议。

[中图分类号] G250.252
[文献标志码] A
[文章编号] 1005－8214(2014)07－0034－04

随着科学技术的快速发展，科学出版物急剧增加，人们很难再对某一专门学科领域的知识结构和发展情况给以全面的观察和考虑，也就更谈不上对科学领域研究前沿的把握和判断了。科技领域前沿是一个研究领域的最新趋势和概念现状。从微观的科学研究人员科研选题，中观的科技产业部门的技术开发、产品的市场定位到宏观的国家科技政策战略的制定，都需要了解、掌握相关领域的研究前沿。如何能够科学、准确地把握研究前沿已经成为科学研究人员及其管理者关注的焦点，更成为各国政府制定科技发展战略时面临的一大问题。因此，对于科学技术研究前沿的自动分析研究，旨在帮助科学工作者从大量的学术会议和科技文献中提取出有用的信息，具有重要的现实意义。

科学研究前沿，简称研究前沿，代表了科学发展的难点、热点以及发展趋势，从浩瀚的科技信息中探测研究前沿是科技创新的关键任务之一。针对研究前沿的专门研究是在 2005 年之后才兴起的研究热点。它涵盖内容广泛，学科交叉性强，综合了科学学、图书馆学、情报学、人工智能、机器学习、数据挖掘、内容可视化与社会网络等方面的内容。

1 研究前沿相关概念

文献计量学领域对研究前沿的定义最早由普赖斯于 1965 年提出，[1] 用它来描述研究领域的动态本质。40 多年来，不同的学者对研究前沿的概念内涵进行了不同的定义和诠释。E. Garfield [2] 把“共引文献以及引用它们的论文”作为研究前沿的定义；Small [3] 和 Griffith [4] 认为共被引文章聚类表征着当前活跃的研究领域；Persson [5] 则认为研究前沿和知识基础的区别在于：“从文献计量学来看，引文形成了研究前沿，被引文献组成了知识基础”；而 Morris [1] 将研究前沿定义为持续被一组固定的、与时间无关的基本文章引用的大量文章；Braam，Moed，Raan 等 [6] 将一个研究领域定义为“一群科学研究者关注的一系列相关问题和概念”；陈超美 [7] 将研究前沿定义为一组突发的概念及其基本研究问题。关于研究前沿的概念分类详见表 1。

总之，关于研究前沿的探测和判断多取决于分析时所采用的计量方法及所依据的数据源，研究前沿的内涵随方法和数据源的不同而有所不同。另外，研究前沿与热点主题、新兴主题、新兴趋势等概念互相交叉，边界难辨，如图 1。关于热点主题，马费成 [8] 等认为如果某一关键词或主题词在其所在领域文献中反复出现，则可反映出该关键词或主题词所表征的研究主题是该领域的研究热点。关于新兴主题，Tu Y 等 [9] 认为是指一个领域中重要的处于成长阶段但还未成为研究热点的主题。关于新兴趋势（Emerging Trend），April [10] 在 2004 年提出，是指随着时间推移引起越来越多的研究兴趣并得到愈广泛使用的一个主题领域。

[基金项目] 本文系中国科学院“西部之光”联合学者项目“基于计算情报方法的甘肃省战略性新兴产业技术创新竞争与发展研究”（项目编号：Y200201001）的研究成果之一。

表 1　科学研究前沿概念分类表

概念分类	作者	年份	研究前沿	聚类
将一组高被引文献定义为研究前沿	Price	1965	对于一篇指定引文，是由被频繁引用的近期文章（30-50 篇）所组成的动态聚类	引文的最近行为(Recentness)
	Small&Griffith	1974	共被引聚类	共引
	Garfield	1991	共引聚类与引文的总和	共引
将一组施引文献定义为研究前沿	Persson	1994	引用相同文献的文章	文献耦合
	Morris et al.	2003	经常被一组固定的、与时间无关的基本文章引用的一组文章	文献耦合
将突发或热点主题定义为研究前沿	Braam et al.	1991	集中关注的一系列相关问题和概念	共被引文章和高频词集
	Chaomei Chen	2006	在某一时段内，以突现文献 (burst article) 为知识基础的一组论文所探讨的科学问题或专题	共被引文章和引用这些文章术语的复合网络
将重点领域与优先主题、前沿技术、科学前沿问题、面向国家重大战略需求的基础研究等具有前瞻性、先导性、理论性、探索性的研究内容作为研究前沿	刘小平等	2012	从目前国内外科技前沿分析实践出发，在时间角度上将研究前沿分为未来的科技前沿问题和当前的科技前沿问题两类：未来的科技前沿是指政府的科技规划战略路线图；当前的科技前沿，是指世界科技强国的资助机构通过各类计划项目最新资助的战略投资重点领域	未来的科技前沿刚刚部署启动或者即将部署，还没有研究出成果的前沿；而当前的科技前沿各国已经部署，但还没有完成的并产生研究成果的前沿

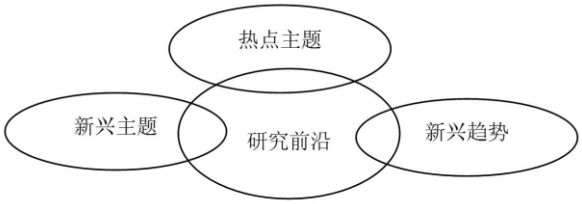

图　近似概念关系

2　研究前沿探测方法概述

在大数据时代，鉴于及时有效把握研究前沿的重要意义，人们早已对研究前沿探测展开了丰富多样的研究，并呈现出了各式各样的研究方法和研究成果(见表 2)。从基本的探测方法入手，可以将这些方法以定性、定量的视角加以总结归纳。

定性方面，文献综述法和德尔菲法是比较常用也较为成熟权威的研究前沿分析方法。它们以归纳为主，广泛收集第一手资料，从研究者的个人背景和知识积累出发，对各家不同思想、观点、方法进行综合整理、归纳分析、概括提炼，最终形成能反映该课题或专题研究水平和发展动态的阶段性的回顾总结、现状描述或技术预见、未来预测等。利用定性研究进行研究前沿探测由于其分析过程的特征优势，一般能得到相对整体全面灵活的分析结果；但也正因为定性研究中研究者即研究工具，其对研究者的素质要求过高，结果的主观性及不精确性也是显而易见的。

定量方面的探测分析一直是研究者们关注的焦点，不仅因为定性方法固有的缺陷需要克服，还因为定量方法在处理大数据方面的分析潜质与平民化特质。从文献计量的角度，可以将研究前沿探测方法分为基于引文的分析法和基于词汇的分析法。

表 2　研究前沿探测方法一览表

分类	研究方法	分析角度
引文分析法	引文分析；共引分析；耦合分析	文章共引；作者共引；期刊共引；学科共引等
词频分析法	高频词分析法；低频词分析法；共词分析法（共词聚类分析法；共词关联分析法；共词词频分析法）；突发词监测法	基于词汇
多元统计分析法	因子分析；多维尺度分析；聚类分析	科学统计
复杂网络分析法	社会网络分析	作者-主题词；作者-作者；主题词-主题词等

(1) 引文分析法。引文分析法是文献计量学领域最常用的分析方法之一，包括直接引文分析、共引分析以及耦合分析。引文分析法不仅广泛用于研究前沿探测，在主题演变、学科分析以及科研能力评价等领域的研究也扮演着举足轻重的角色。人们普遍认同运用引文分析方法探测研究前沿，但所运用的具体引文类型各不相同，Naoki Shibata 等[11]、Persson[5]、Schiebel Edgar[12] 分别基于不同的引文类型对研究前沿展开了分析和研究。但是，究竟哪种类型的引文分析法更适于研究前沿的探测，目前还未达成共识，相关研究也比较少见。其中，Klavans 和 Boyack[13]对共引分析、引文耦合、直接引用和基于引文耦合的混合引文方法开展了对研究前沿探测效果的对比。结果表明：直接引文网络可更直接、更早地揭示科学引文网络所代表的研究领域的结构特征和发展趋势；在精确度指标上，引文耦合及混合引文方法稍优于共引分析，直接引用是最不准确的方法。Shibata[14] 对同被引、引文耦合与直接引用方法的探测效果进行了对比，直接引用能较早探测大的新出现的聚簇，在探测研究前沿上效果最好，直接引用法探测效果最全面，而同被引效果最差。在实际应用中，大多数分析人员都会综合运用上述方法，以获得最佳的研究效果。

(2) 词汇分析法。基于词汇进行研究前沿探测的分析方法主要包括词频分析法和共词分析法。词频分析中的高频词能有效探测研究热点，低频词有助于预测新兴主题和新兴趋势。当前的词频分析主要集中在

以关键词或主题词为对象的词频分析，如喻培珍[15]和郭凌辉[16]分别利用基于主题词和关键词的词频分析探测相关领域的热点及前沿。虽然词频分析法相对简单，分析结果直接且易于理解，但由于词频具有波动性及词频阈值的人工干预，通常采用的固定阈值在分析时易出现误差，加之需要专家依据知识背景将词分成特定研究主题，使得词频分析法的分析结果主观性太强。

而共词分析法能够在最大程度上发挥词频分析的优势，对文献资料的挖掘更深入准确，越来越多的研究者将目光转向共词分析，如 Luan CJ 等[17]、Xin Ying An 和 Qing Qiang Wu[18]、刘丽[19]等人都曾运用共词分析探测研究前沿。同时，共词分析法得到了持续改进：分析词从索引词、关键词发展到自由词，从单个词语、双词短语再到多词短语,词语共现范围被限定在同一句子之内、数十个词之内、同一段落之内或者同一篇论文之内等等；[20] QingQiang Wu[21]等基于 LDA 概率主题模型，集成共现理论和聚类指标构建了主题分割模型 ATNLDA，深入挖掘文献主题及其之间的关系以探究主题演化规律；叶春蕾、冷伏海[22,23]提出基于概率模型的主题识别方法，将 LDA 主题模型与共词分析相结合改进主题识别方法，体现了主题词、主题和文档间的层次语义关系。值得注意的是，在利用共词分析处理词汇语义关系的问题上，他们都引进了 LDA 模型，并取得了可观的分析结果。LDA（Latent Dirichlet Allocation——潜在狄利克雷分布模型）是由 Blei、Ng、Jordan 2002 年提出的完全概率语言模型，应用到文本建模范畴，就是对文本进行“隐性语义分析”（LSA）。LDA 模型不仅具有强大的理论支撑，还具有较易控制的参数设置以及良好的泛化能力，能够以词组的形式充分反映主题词—主题—文档间的语义关系，改善了共词分析不能有效表达词汇间语义关系的状况，使得分析结果更加准确、成熟、可靠。与单纯的主题词统计、排序，进而分析研究热点的文献计量方法相比，共词分析不仅专注于高频词，更关注词间联系，更好地反映了概念及语义之间的关系。但是，因其分析对象是已发表的文献，故具有时滞性，无法及时有效反映还未形成热点的前沿趋势等潜在前沿主题；而且其词频阈值的选择也不可避免地会影响到聚类效果，进而影响到主题探测效果。虽然如此，在反映当前论文关注主题的同时，共词分析仍较基于引文的分析方法更灵活、简单、直观。

3 结论及展望

（1）针对研究前沿主题的判定方法研究。虽然当前的方法可以识别出研究前沿，但定性方法过分依赖研究者主观经验和知识，而定量方法仅能做到主题聚类，具体前沿主题仍依赖研究者主观经验或者专家知识，且关于研究前沿的定义及判定标准随所采用方法的不同而不同。故针对研究前沿主题判定方法的研究已迫在眉睫。虽然已有研究者尝试设计一套指标来辅助判定研究前沿，但公认的客观可信赖的指标体系还有待进一步研究。

（2）针对研究前沿探测的混合方法研究。虽然各种研究前沿探测方法都有自身特点，但是受数据源和分析原理影响，都有不可避免的缺陷。比如基于引文方法的优势是其辨别力，但其缺点在于低估文档间的关系以及分析的滞后性；而词频分析虽简单易行，但只是从宏观角度考察学科发展动向，对研究前沿更深入的分析还多依赖专家判读。将共词法与其他方法相结合的混合方法在研究前沿探测方面已成趋势，其中引文法与共词法结合的突出效果已广为接受；但当前的混合方法常以简单直接的方式混合，并未考虑边界语义，如此简单组合各方法可能会导致预想不到的问题。

（3）针对不同类型的前沿探测方法的比较研究。目前已有多种针对科学研究前沿探测的方法，但探测效果参差不齐，研究者在面对不同的选题和目的时暂无可依据的方法遴选金指标。为能更加科学准确及时迅速地探测科学研究前沿，针对各种探测方法及其比较的研究是十分必要和紧迫的。

（4）针对研究前沿探测的数据源多样化研究。当前的研究前沿探测方法多以期刊论文等为数据处理对象，较少涉及其他形式的数据源。虽然文献是科研产出的主要形式之一，但并不足以代表所有科研成果所涵盖的信息量。诸如各国各部门的科技规划、战略蓝图、路线图、各类机构资助的重点领域的项目申请书内容和研究报告、专利相关文件以及重要组织、学会、科研团体撰写的有关研究前沿的研究报告和战略文件等，从某种程度上讲，这些资料更能及时有效反映科学研究前沿。今后的研究前沿探测方法可以考虑从多样化的数据源入手，也许能得到意外的收获。

[参考文献]

[1] Morris S A，et al. Time line visualization of research fronts [J]. Journal of the American Society for Information Science and Technology，2003，54（5）：413—422.

[2] Garfield E. The new 1956–1965 social–science citation

index. 1. Analysis of 1988 Research fronts and the citation-classics that made them possible [J]. Current Contents, 1989, 41: 2—8.

[3] 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.

[4] Griffith B C, et al. The structure of scientific literatures II: toward a macro-and microstructure for science [J]. Social Studies of Science, 1974, 4 (4): 339—365.

[5] Persson O. The intellectual base and research fronts of JASIS 1986-1990 [J]. Journal of the American Society for Information Science, 1994, 45 (1): 31—38.

[6] Braam R R, et al. Mapping of science by combined co-citation and word analysis, I. Structural aspects [J]. Journal of the American Society for Information Science and Techology, 1991, 42 (4): 233—251.

[7] 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.

[8] 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.

[9] Tu Y N, Seng J L. Indices of novelty for emerging topic detection [J]. Information Processing & Management, 2012, 48 (2): 303—325.

[10] Kontostathis A, et al. A survey of emerging trend detection in textual data mining [M]. New York: 2004: 185—224.

[11] Shibata N, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications [J]. Technovation, 2008, 28 (11): 758—775.

[12] Schiebel, E. Research fronts and areal density of bibliographically coupled publications [C]// Proceedings of 13th international conference of the international society for scientometrics and informetrics (ISSI 2011). Proceedings of the International Conference on Scientometrics and Informetrics, 2011, 756—762.

[13] Boyack K W, Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front mostaccurately? [J]. Journal of the American Society for Information Science and Technology, 2010, 61 (12): 2389—2404.

[14] Shibata N, et al. Comparative study on methods of detecting research fronts using different types of citation [J]. Journal of the American Society for Information Science and Technology, 2009, 60 (3): 571—580.

[15] 喻培珍, 秦惠基. 从主题词频率变化分析我国放射诊断新技术发展趋势 [J]. 医学图书馆通讯, 1995, 4 (3): 26—27.

[16] 郭凌辉. 知识发现 (KD) 研究热点与前沿的信息可视化分析 [J]. 图书馆理论与实践, 2011 (8): 27—30.

[17] Luanc, et al. Quantitative studies on frontiers of international patent bibliometrics [J]. Studies in Science of Science, 2008 (2): 20.

[18] An X Y, Wu Q Q. Co-word analysis of the trends in stem cells field based on subject heading weighting [J]. Scientometrics, 2011, 88 (1): 133—144.

[19] 刘丽. 公共图书馆研究热点领域知识图谱: 共词分析视角 [J]. 图书馆理论与实践, 2012 (7): 62—65.

[20] 王立学, 冷伏海. 简论研究前沿及其文献计量识别方法 [J]. 情报理论与实践, 2010, 3 (10): 54—58.

[21] Wu Q Q, et al. Topic segmentation model based on ATNLDA and co-occurrence theory and its application in stem cell field [J]. Journal of Information Science, 2013, 39 (3): 319—332.

[22] 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.

[23] 叶春蕾, 冷伏海. 基于概率模型的主题识别方法实证研究 [J]. 情报科学, 2013 (2): 135—142.

[作者简介] 刘静 (1990—), 女, 硕士研究生, 研究方向: 计算机信息处理与检索; 马建霞 (1972—), 女, 研究馆员, 硕士生导师, 研究方向: 文本挖掘与情报计算研究; 范云满 (1980—), 男, 硕士研究生, 研究方向: 计算机信息处理与检索。

[收稿日期] 2013—12—16 [责任编辑] 王钧梅