国内外共词分析研究综述

李 颖（西藏大学图书馆）
贾二鹏（西安理工大学图书馆）
马 力（滁州学院图书馆）

【摘 要】 共词分析法是科学计量学中最常用的方法之一。论文从共词分析的起源与原理入手，介绍共词分析的一般研究流程，总结目前共词分析的国内外研究现状及其优势与不足，并阐述了共词分析法与学科演进态势研究之间的紧密联系。

【关键词】 共词分析；关键词；科学计量学；学科演进态势

【Abstracts】 Co-word analysis is one of the most widely applied means in scientific metrology.This paper introduces the common study process of co-word analysis by dealing with its origin and principles,summarizes the current study condition,advantages and disadvantages of co-word analysis at home and abroad, and illustrates the close connection between co-word analysis and the evolution situation study of subjects.

【Keywords】 Co-word analysis; key word; Scientific metrology; Subject evolution situation

1 共词分析概述

1.1 共词分析的起源与原理

共词分析法最早是由法国文献计量学家在20世纪70年代中后期提出并对其进行详细描述的，发展至今已将近40年，之后经过Callon、Whittaker、Courtial、Turner等学者的反复研究、修正与补充，共词分析理论日趋完善。到了20世纪90年代中后期，作为内容分析方法之一的共词分析法已基本走向成熟，并被广泛应用于分析各个学科领域的研究结构，取得了较为丰硕的研究成果。

共词分析（Co-word Analysis）是通过对反映文献主题内容的关键词进行统计分析，研究文献内在联系和科学结构。它之所以能够被用来研究学科领域的研究热点与结构，是基于以下两点：一是众多科研人员关注的研究热点并不是单一孤立的，而是由一系列在内容上存在密切关系的关键词或主题词构成的；二是不论科学研究者在社会和知识背景上存在何等差异，当他们面对同一研究课题和概念时，所使用的文献主题词汇是基本一致的。

共词分析的思想来源于文献计量学的引文耦合与共被引分析。与之同理，当两个能够表达某一学科领域研究主题或研究方向的专业术语（多表现为文献的主题词或关键词）同时出现在一篇文献中，则表明这两词之间存在一定的相关关系，如果两词共现的次数越多，就表明它们的关系越密切。利用现代统计技术如因子分析、聚类分析和多维尺度分析等多种多元分析方法，就可以依据这种关键词之间的“距离”对某一学科领域内的重要主题进行分类，从而归纳出该学科的研究热点、研究结构。不仅如此，利用现代信息技术特别是可视化技术，便可将分析的结果以更加直观形象的方式表示出来，进而达到明确清晰的分析效果，便于理解。通常鉴于共词分析能够用来研究当前该学科领域文献所集中关注的研究主题，因而比较适合应用它来探讨新兴学科的研究热点与演进趋势。

Monarch[1]曾对共词分析的历史进行过深入研究，他认为共词分析技术就是通过分析某一学科领域相关文献的代表性术语之间连接强度，从而得出该学科领域的研究发展方式和趋势。因此，共词分析的一个主要途径就是通过对这些代表性术语之间概念图谱或知识网络结构的确定，来详细描述某一学科领域的研究主题。

1.2 共词分析的一般研究流程

共词分析是一种内容分析方法，在对一组词中两两词语出现在同一篇论文中的次数进行统计的基础上，利用聚类分析方法对这些词的亲疏关系进行分析，并通过可视化方式反映出这些词所代表的学科或主题领域的研究结构。相比同被引聚类分析，共词聚类分析方法被分析和聚类的对象不是文

献而是文献中的关键语词，由于词与词间的相互关系表征的是概念之间的关系，因而共词聚类分析之后所形成的类就能够比较简洁明确地描述学科领域的发展特点与研究结构。

利用共词分析研究学科领域文献的研究结构，其一般流程大致可划分为四步：

1.2.1 关键词的提取

词组作为共词分析中最重要的研究对象，有两种方式可以从期刊文献、参考论文、报告或著作等文献中进行抽取。一种是从关键词列表、标题、摘要等进行提取；另一种数据收集方式则是利用专门的语词提取软件直接从全文抽取。如果利用数据库管理软件以及SPSS统计软件对文献语词进行识别统计，那么可能会出现这样的情况，表达意思完全相同的不同语词却被看作是两个完全相异的词汇，这就导致最终的统计分析结果出现偏差。因此，最好选择受控的、统一标引的规范主题词作为分析对象。只有这样，利用文献中关键词语对的共同出现频次来反映文章中所蕴含的概念进行共词分析才能成立[3]，而直接选择文献中的关键词或主题词作为共词分析的基本单元就是比较简单又具有代表性的抽取方法。

1.2.2 高频词的选定

关键词或主题词一般都是一篇文献核心内容的浓缩和提炼，可以在很大程度上代表文献的研究主题。因此，如果某一关键词或主题词在其所在领域的文献中多次出现，则可说明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[3]。为了简化统计过程及减少低频词对统计过程带来的不必要干扰，通常共词分析是选择某一研究领域的高频主题词作为分析对象。但是共词分析对高频词数量阈值的确定通常没有达成统一见解，如果选择的关键词范围过小，则不能如实反映其所代表的学科知识点的构成；如果设置的高频主题词阈值范围过大，则又会给之后的共词分析过程带来相当大的干扰。目前来说，高频词的确定主要有两种方法：一种是根据研究者的自身经验确定选词个数与词频高度以达到平衡，但该方法具有一定的随机性与主观性；另一种方法则是利用齐普夫第二定律来辅助确定高频词的界限[4]。

1.2.3 共词矩阵的构建及其标准化

在共词分析中，一旦确定研究主题，接下来便是根据所选定的高频关键词构建基于关键词共现的共词矩阵。若两词之间的共现频率越高，则说明它们之间存在较为相近的关系。但是在实际的计量化分析过程中，由于关键词的出现频率是绝对值，难以反映彼此之间真正的相互依赖程度，故此就需要利用一些特殊的相关系数计算方法将所得到的原始共词矩阵转换成相关矩阵，便于进行之后的分析。

1.2.4 共词分析图谱的绘制

利用共词分析方法研究学科研究结构，其最终成果就是利用高频关键词之间关系所形成的知识网络结构，通过一系列的数据分析方法绘制出该学科领域的知识图谱，以期能够以更客观直接的可视化方式来反映各个研究主题之间的关系，深入揭示隐含在文献群中的知识。

绘制知识图谱的方法众多，但最常使用的有多维尺度分析和聚类分析技术。关于这方面的研究，早已有一些专门软件被研发出来：LEXIMAPPE，是20世纪80年代由法国国家科学研究中心CNRS（Centre National de la Recherché Scientifique）联合研发；CAIR（Content Analysis and Information Retrieval），是由卡内基梅隆大学软件工程实验室研发；BibTechMon（Bibliometric Technology Monitoring），是澳大利亚研究中心研发的一款用于共词分析的软件。

2 国内外共词分析法研究

2.1 共词分析研究现状

利用共词分析法研究某一学科领域或主题，可以得出其学科或主题的研究热点，并且能从横向和纵向角度揭示学科领域的发展过程、结构以及领域或学科之间的关系等等。廖胜姣等[5]在《基于文献计量的共词分析研究进展》一文中，使用文献统计的方法详细分析了目前国内外有关共词分析研究论文的特点，可以看出在国外共词分析方法早已被广泛应用到多个学科领域的研究热点与研究结构的分析中，而国内目前这方面的研究虽然增长较快，但仍有待充分发挥共词分析方法的优势，进一步加强研究。

基于共词分析的研究主要是通过对文本所包含知识点的挖掘，进而达到映射学科知识结构的目的。以前的很多研究都是应用共词分析方法将某一研究主题的趋势与定期变化概念化，或绘制术语间关系的知识图谱，或者跟踪学科或主题的研究模式与发展趋势。

Callon，Vourtial & Laville[6]以高分子化学领域为例，利用共词分析揭示了学术研究与技术研究之间相互影响的方式；Law & Whittaker[7]以环境的酸性研究为例，利用共词分析方法绘制了该领域的科学图谱；Courtial[8]等人将专利文献的题目作为分析对象，对其进行共词聚类分析，得到食品类专利文献的研究热点，同时还利用战略坐标将这些研究热点显示出来；PETERS & VAN RAAN[9]使用聚类多维尺度分析的方法，将聚类分析同多维尺度分析相结合分析了化学工程领域的研究进展；Courtial[10]根据科学交流的网络属性，使用共词分析方法分析了科学计量学领域的研究动态，并对该领域将来的发展趋势作出预测。

Noyons & van Raan[11]和Van Raan & Tijssen[12]利用共词分析方法分析了神经网络研究领域的演进态势；Coulter，Monarch & Konda[13]与KOSTOFF[14]分别将共词分析方法应

用到软件工程领域用以识别领域研究现状与研究结构；DING&AL[15]以SCI和SSCI中有关信息检索领域的相关文献为基础，分析其主题领域以及特殊时期的变化。并通过研究其相关文献的关键词的共现频率，利用多维尺度分析生成了一幅详细的领域地图及每一聚类分析类团的详细领域地图；Clemens[16]将共词分析技术应用于运输业，该项研究不仅利用共词分析研究学科的主题结构和发展，并创造性地将研究机构、作者纳入到研究范围之内，得到该领域文献作者与研究机构之间相关关系的认识；Irene[17]打破了之前仅将共词分析方法应用到自然科学领域的现状，将其引入到社会科学领域的研究中，对现代福利国家中福利的实践工作进行研究；Jacobs[18]则利用共词分析研究了特定词汇描述居民工作职能和信息资源引用的用途；SCHNEIDER&BORLUND[19]则将共词分析应用到叙词表的构建与维护中。

Stegmann[20]等利用主题词聚类分析的方法对Swanson等人所做的发现非相关文献之间隐含关系的数据进行研究，并绘制了其研究热点的战略坐标图，结果显示战略坐标能很好地再现Swanson等人的研究结果，并且发现研究结果与语词在战略坐标中的位置之间存在着一定的联系；AIZAWA & KAGEURA[21]在统计学术文献关键词共现频率的基础上利用共词分析计算了科学术语之间的联系；Rikken, Kiers & Vos[22]和Looze & Lemarie[23]将共词分析应用到医学领域的主题研究中；ONYANCHA & OCHOLLA[24]利用共词分析衡量了HIV/AIDS感染几率的相关性；Mizuki[25]从医学角度将每个病患的数据转换成能够表征所有相关因素的共现矩阵，然后利用主成分分析方法抽取其最主要的特征，最终达到找出该疾病的主要形成因素的目标。

与国外共词分析的研究成果相比，国内对共词分析在领域演进态势方面的研究探讨还不够深入，也不够系统，应用研究也相对简单和单一。

蒋颖[26]使用文献计量法分析了国内外共词分析的研究进展，通过文献统计给出了国内词共现领域研究的核心作者有史金生、崔雷、郑华川、张晗以及耿焕同。纵观这些学者发表的与共词分析相关的文章，大多都是从研究学科领域结构的角度出发，应用数学指标体系和关系范式与共词分析相结合，对具体的学科领域进行实证分析。崔雷[27]利用共词分析方法，分析了医学文献的学科结构并追踪了学科的研究热点。柴省三[28]探讨了将共词分析和引文共引聚类分析相结合的方法来研究学科领域主题和科学结构。朱东华[29]运用共词分析原理，将计算机领域的前沿技术应用到科学技术管理领域中，并对科学技术发展进行了监测分析。谢彩霞等[30]对1994—2001年间我国纳米科技领域的学术论文作了其关键词的共现分析，展示了我国纳米科技研究领域的发展动态和趋势。张晗等[31]使用共词聚类分析对生物信息学相关文献的高频主题词进行分析，很好地显示了该主题领域的研究热点，同时还绘制了战略坐标图，用以定量地分析各个研究热点的发展阶段。

还有一些学者是根据共词分析的特性，将之应用到系统设计中用来进行关系数据的挖掘。如宋爽[32]研究了将共现分析应用到基于空间分布、时间分布和内外关联映射的文本知识挖掘中的适用范围及一般操作流程。肖雷[33]以急性白血病为例，利用聚类分析方法通过自然语言和主题词两种途径研究疾病与基因的共现关系，并对二者之间的关系进行挖掘。陈颖[34]在基于摘要信息的中文信息检索可视化系统设计中引入基于词共现的概念空间方法与信息检索可视化技术相结合实时生成概念空间图，最终实现检索过程以及结果的可视化。郝丽云[35]借助词频统计分析、语义过滤、共词聚类分析等方法对文献的标题词以及主题词等文献单元进行分析，探索非相关文献的知识发现过程。张浩[36]在硕士论文中应用主题词共现聚类分析方法，研究来自于MEDLINE数据库中有关生物体类相关文献中的高频主题词之间的潜在语义关系和关联规则。

纵观以上国内外共词分析的发展脉络以及研究现状，可以从应用领域方面将基于共词分析的研究成果归纳为以下四个主题：（1）揭示特定领域内的研究主题及其层次之间的关系，以及其对应的研究方向间的关系，划分科学子研究领域并确定其研究结构；（2）从横向和纵向角度揭示特定领域内研究主题之间以及同其他研究主题之间的关系；（3）考察特定研究领域内研究主题发展的历史脉络及其子领域的演进态势；（4）通过词间关系的数据挖掘达到学科主题知识发现的目的。

总体来说，目前我国关于共词分析的应用研究还较为单一，在技术方法的创新上也存在不足。因此，有必要从理论和应用相结合的角度上，对其进行系统而全面的研究，以弥补我国现有研究的不足，丰富共词分析方法的内涵及应用。

2.2 共词分析法的优势与不足

共词分析因其适用于研究学科领域的文献主题与内容以及其方法本身所具有的优势，在当前众多的文献计量方法中，常常被用于以定性和定量相结合的方式研究学科主题的演进状态和对学科领域动态发展过程的监测与跟踪。现有的两种最主要的基于共词分析的学科主题动态跟踪方法：一种是基于知识图谱的定性分析方法，另一种是基于相似性测度的定量分析方法。这两种方法可以相互补充，也成为学科主题动态跟踪的重要手段之一[37]。

比起利用简单的关键词分析文献主题，共词分析是通过对高频关键词之间的亲疏关系与相关性进行分析，并用可视

化的方式揭示这些高频主题词所在学科领域与研究主题的结构发展变化，因此能够弥补仅仅使用主题词来分析揭示文献主题之间内在关系的不足。但是要有足够长的时间来使学科主题的文献与文献主题词达到一定的积累量，才能够进行深入的分析研究。由此可知，利用共词分析时会存在一定的时滞性，不能够完全地反映某一学科主题的发展趋势和最新生长点[38]。

与共被引分析法相比，共词分析法是对当前已发表文献主题的直接统计，所寻找的是当前研究论文所集中关注的主题，其所反映的主题是在趋势形成之后的研究焦点，因此利用它更适合寻找新兴学科的研究范式；而共被引分析法则是通过分析以往发表论文的引用情况来表现目前的研究焦点，而论文通常是在发表后的三年内才能达到被引高峰，因此更适合于寻找成熟学科的研究范式，这也是因为新兴学科的研究往往人数众多却不集中，研究主题比较分散，被引用情况不稳定，而关键词却能很好地体现表征这一新兴学科领域的发展方向的研究结构和研究热点，从而更加有助于探索该学科领域的演进态势[39]。

共词分析法自20世纪末发展至今，由于其自身具有灵活、简单、结果直观等优点，已经被广泛应用于各个学科领域的研究。但是，不可否认共词分析自身仍存在一些不足之处。比如在对高频关键词进行聚类分析时，很难同时兼顾聚类与结果有效性之间的平衡；由于文献主题词积累所引发的分析时滞性也导致其在揭示学科领域研究结构及其动态变化上不可避免存在滞后。总的来说，虽然共词分析方法本身存在一些缺陷，但因为共词分析是以语词之间在文献中的共现频次为基础对学科研究主题进行提炼分析，并且以词与词之间相关度构成的网络关系所表征的主题作为监测对象，因此能够比较客观地揭示科学研究领域中研究主题的演进态势。

3 结语

某一学科或研究领域都会在自己的历史演化进程中，逐渐形成属于自己特有的组织结构，包括诸如研究者、研究机构、研究内容、研究成果等多维结构。作为科学研究成果的有形载体之一，科技论文中包含着大量与研究主题相关的线索与信息，因此针对科技论文所包含的数据进行分析，就可以客观全面地理清科技论文所代表的研究领域之历史脉络与演进态势[40]。而关键词或主题词作为文献研究内容的一个重要表征方式，以及学科发展过程的重要标志之一，对它进行研究可以揭示学科或研究领域在演进过程中的诸多细节信息。共词分析方法通过对相关研究文献的主题词之间相关度的分析，确定出这些主题词之间所形成的概念图谱或知识网络结构，并通过多元统计方法绘制出一系列类似的图谱，这样就可以细致地描绘出某一学科领域的研究主题，从而得出某一学科领域的研究结构和演进态势[42]。

参考文献

1 Monarch I.Information science and information systems: converging or diverging?Proceedings of the 28th Annual Conference of the Canadian Association for Information. [EB/OL].http://www.slis.ualberta.ca/cais2000.monarch.htm. 2000,2009-10-07.

2 冯璐,冷伏海.共词分析方法理论进展[J].中国图书馆学报,2006(2):88~92

3 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(2):163~171

4 魏瑞斌.基于关键词的情报学研究主题分析[J].情报科学,2006(9):1400~1404,1434

5 廖胜姣,肖仙桃.基于文献计量的共词分析研究进展[J].情报科学,2008(6):855~859

6 Callon M,Courtial J P,Laville F.Co-word analysis as a tool for describing the network of interactions between basic and technological research:The case of polymer chemistry[J]. Scientometrics, 1991(1):153~203

7 Law J,Whittaker J.Mapping acidification research:A test of the co-word method[J].Scientometrics,1992(3):417~461

8 Courtial J P, Callon M, Sigogneau A. The use of patent titles for identifying the topics of invention and forecasting trends [J]. Scientonetrics, 1993(2): 231~242

9 Peters H P F,van Raan A F J.Co-word based science maps of chemical engineering,Part I:Representations by direct multidimensional scaling[J]. Research Policy,1993 (22):23~45

10 Courtial J P.A co-word analysis of scientometrics[J].Scientometrics, 1994(3):251~260

11 Noyons E C M,van Raan A F J.Monitoring scientific developments from a dynamic perspective:Self-organized structuring to map neural network research[J]. Journal of the American Society for Information Science,1998(1): 68~81

12 Van Raan A F J,Tijssen R J W.The neural net of neural network research[J]. Scientometrics,1993(1):169~192

13 Coulter N, Monarch I,Konda S. Software engineering as seen through its research literature:A study in co-word analysis[J]. Journal of the American Society for Information Science, 1998(13):1206~1223

14 Kostoff R N.Science and Technology Metrics[EB/OL].

http://www.dtic.mil/dtic/kostoff/Metweb5_IV.htm,2009-11-07.

15 Ding Y, Chowdhury G G, Foo S. Bibliometric cartography of information retrieval research by using co-word analysi[J].Information Processing and Management, 2001(37):817~842

16 Clemens,Alexander K.Co-occurrence and Knowledge Mapping to Indentify Hot Topics and Key Players in the Field of Mobility and Transport[EB/OL]. http://www.semantic-web.adfile-upload/roottmpphpA4EeZH.pdf,2009-11-07.

17 Irene W. Bibliometric Analysis of the Welfare Topic[J]. Scientometrics, 2000(2):203~236.

18 Jacobs N. Co-term network analysis as a means of describing the informational landscapes of knowledge communities across sectors[J].Journal of Documentation,2002 (5): 548~562

19 Schneider J W, Borlund P. Introduction to bibliometrics for construction and maintenance of thesauri:Methodical considerations[J].Journal of Documentation,2004(5): 524~549.

20 Stegmann J,Grohmann G.Hypothesis generation by co-word clustering [J].Scientometrics,2003 (1): 111~135.

21 Aizawa A,Kageura K.Calculating association between technical terms based on co-occurrences in keyword lists of academic papers[J]. Systems and Computers in Japan, 2003 (3): 85~95

22 Rikken P, Kiers H A L,Vos R. Mapping the dynamics of adverse drug reactions in subsequent time periods using Indscal[J]. Scientometrics,1995(3):367~380

23 Looze M D,Lemarie J.Corpus relevance through co-word analysis:An application to plant proteins [J].Scientometrics,1997(3):267~280

24 Onyancha O B, Ocholla D N.An informetric investigation of the relatedness of opportunistic infection to HIV/AIDS [J].Information Processing and Management,2005(41):1573~1588

25 Morris S A.Manifestation of emerging specialties in journal literature:a growth model of papers, references, aexemplars, bibliographic coupling, co-citation, and clustering coefficient distribution[J].Journal of the American Society for Information Science and Technology, 2005(12): 1250~1273

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

27 崔雷.专题文献高频主题词的共词聚类分析[J].情报理论与实践,1996(4):49~51

28 柴省三.内容词—共引聚类分析及其在科学结构中的应用[J].情报学报,1997(2):68~73

29 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(4):135~139

30 谢彩霞,梁立明,王文辉.我国纳米科技论文关键词共现分析[J].情报杂志,2005(3):69~73

31 张晗,崔雷.生物信息学的共词分析研究[J].情报学报,2003(5):613~617

32 宋爽.共现分析在文本知识挖掘中的应用研究[D].南京:南京理工大学,2006.6

33 闫雷.急性白血病相关基因的文本挖掘分析[D].沈阳:中国医科大学,2006.9

34 陈颖.基于摘要信息的中文信息检索可视化系统研究与实现[D].哈尔滨:黑龙江大学,2007.4

35 郝丽云.非相关文献知识发现的医学研究与实践[D].南京:中国人民解放军军事医学科学院,2007.5

36 张浩.MEDLINE数据库中生物体类主题词相关语义关系的构建与评价[D].沈阳:中国医科大学,2008.9

37 韩真.基于共词分析的主题类型划分方法比较研究[J].图书馆,2009(2):46~47,53

38 郑华川,于晓欧,辛彦.利用共词聚类分析探讨抗原CD44研究现状[J].中华医学图书情报杂志 2002(2):1~3

39 崔雷.专题文献高频主题词的共词聚类分析[J].情报理论与实践,1996(4):49~51

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

41 刘则渊,尹丽春.国际科学学主题共词网络的可视化研究[J].情报学报,2006(5):634~640

（作者信息：李颖，西藏大学图书馆助理馆员，邮编：850000；贾二鹏，西安理工大学图书馆助理馆员，邮编：710048；马力，安徽滁州学院图书馆馆员，邮编：239000。收稿日期：2011-05-23）

编校：刘勇定