DOI:10.13530/j.cnki.jlis.170006

共词分析过程中的若干问题研究*

李　纲　巴志超

摘　要　为完善和优化共词分析方法,本文从共词分析过程中概念术语的词源选择、高频词的选定、术语相关性计算以及多元统计分析四个方面系统地总结共词分析存在的局限性。在词源选择方面,论述不同类型的文献分析单元、术语的规范化以及术语表征差异性对共词分析的影响;在高频词选定方面,分析国内外相关研究在设定高频词阈值、考虑术语语义类型特征以及低频关键词处理等问题时存在的不足,并提出相应的解决方法;在术语相关性计算方面,认为术语之间不仅存在着直接的频次共现,还存在间接的语义相关,总结现有的术语语义相关性度量方法,并对其相关特征进行分析;在多元统计分析方面,对共词分析中常采用的统计分析方法和应用策略进行探讨。本文基于严谨、客观的态度对共词分析的优缺点做出评价,有利于该方法的不断完善和发展,同时也为继续从事共词分析研究的人员提供理论借鉴和实践参考。图 1。参考文献 82。

关键词　共词分析　词源选择　术语规范化　高频词选定　语义关联　多元统计分析

分类号　G250.7

Co-word Analysis: Limitations and Solutions

LI Gang & BA Zhichao

ABSTRACT

Co-word analysis is a content analysis technique based on the assumption that the subject of a paper can be summarized in a limited number of key terms. If two terms co-occur within one paper, the two research topics they represent are related, and the higher frequency of the co-word means stronger correlation in terms pairs. However, the basic work of co-word analysis is still words and extremely sensitive to the selection of terms, and the quality of co-word analysis depends on a variety of factors, such as the quality of terms and indexes, the high-frequency terms extraction, and the adequacy of statistical methods. Therefore it is necessary to delve into the limitations of co-word analysis at different stages to improve and optimize it.

The co-word analysis conducted in the present study involved six sequential steps: determination of problem analysis, term source selection, high-frequency terms extraction, relevance calculation of terms, multivariate statistical analysis, and visual presentation of results. This paper focuses on those six key issues to analyze and demonstrate the main problems based on the induction and summarization of the existing relevant research. Results indicate the following conclusions. 1) In the term source selection, solely making use of keywords and index words, which is called "indexer effect" by researchers, is the biggest problem of

* 本文系国家自然科学基金项目"科研团队动态演化规律研究"(编号:71273196)的研究成果之一。(This article is an outcome of the project "Research on Dynamic Evolution Rule of Scientific Research Team" (No.71273196) supported by National Natural Science Foundation of China.)

通信作者:巴志超,Email:bazhichaoty@126.com,ORCID:0000-0001-5626-5604 (Correspondence should be addressed to BA Zhichao, Email:bazhichaoty@126.com, ORCID:0000-0001-5626-5604)

early co-word analysis. Keywords are uncontrolled words, and problems of homonyms and synonyms will be brought out. Meanwhile, terms expression differences exist among different parts of analysis units, and some errors of co-word analysis will be induced if those differences are ignored. In order to solve the above problems, the textual semantic structure and the phenomenon of different quality with different quantity of terms can be considered. 2) Researchers engaged in co-word analysis have never been out of the pattern that adopts high-frequency term to develop the multivariate statistical analysis. The extraction of high-frequency terms not only makes low-frequency terms more marginalized, but also causes isolation of high-frequency terms that have low correlation with clusters. Considering the discipline and multi-semantic types of terms to distinguish the representation capabilities of subject areas, we can have a comprehensive and in-depth understanding of the research characteristics of this field. 3) Two co-occurrence terms may correlate each other directly or indirectly, but these semantic relationships between co-occurrence terms are not considered at all, which may affect the soundness of the results of co-word analysis ultimately. Thus we summarize the existing calculation methods of semantic correlation and point out the limitations of each method. 4) Finally, in the multivariate statistical analysis, taking the co-word clustering and co-word association analysis method as example, we discuss the problems of their application in the new data environment and put forward the improvement method and suggestion.

Co-word analysis has been most commonly utilized in mapping or tracing patterns and trends in term association. Although co-word analysis has been improved in many aspects, it still has some limitations. This paper tries to provide theoretical and operational references for co-word analysis researchers and enhancing the reliability and effectiveness of co-word analysis, and it is of great significance to jump out of the traditional theory and practice strategy of co-word analysis. 1 fig. 82 refs.

KEY WORDS

Co-word analysis. Term source selection. Vocabulary standardization. High-frequency words selection. Semantic association. Multi-statistical analysis.

0 引言

共词分析(Co-word Analysis)方法是内容分析方法之一,也是目前情报学领域常用的研究方法之一。其基本原理是通过统计文献集中词汇对或名词短语的共现情况,来反映关键词之间的关联强度,进而确定这些词所代表的学科或领域的研究热点、组成与范式,横向和纵向分析学科领域的发展过程和结构演化[1-2]。共词分析方法具有操作灵活性以及分析结果的直观性等特点,已成为快速识别学科发展的重要工具,并在各学科领域中得到广泛应用。通过CNKI精确检索,主题中包含“词共现”或“共词分析”的期刊论文有1 765条(检索时间为2016年10月28日),可见此方法的重要性。当前,对共词分析方法的研究更多是侧重于以共词分析过程为主线,探究分析对象的改进、测度指标改进、可视化方法调整和融合其他方法等方面的理论研究,以及基于词、主题、时间维度和拓展应用四个层次的具体应用研究[3]。相比单纯的词频统计方法,共词分析方法不仅注重关键词的文档频率,更加注重获取关键词之间的联系,从而能够更好地揭示关键词之间的语义关系[4]。然而,共词分析方法的基础性工作仍然是对词的分析,其对关键词的选择具有较强的敏感性,词源选择、高频词选定、术语之间相关性计算以及采用的多元统计分析方法等都会对

共词分析的结果产生较大影响。

共词分析过程主要包括六个步骤：确定分析问题，概念术语的词源选择（分析单元的确定），高频词选定，术语的相关性计算并构建相关矩阵，多元统计分析，结合相关学科知识对统计结果进行科学分析。共词分析过程是共词分析理论研究的出发点，也是将其应用于实证的基准。研究共词分析过程中不同阶段存在的局限性，完善和优化在实际应用时的具体方案，形成系统而有效的方法也是共词分析研究的主要目的[5]。然而，分析过程的第一步及最后一步都是研究者主观意识的选择，属于完全有监督的科研行为，且不同学科领域需要解决的问题以及针对统计结果得出的结论侧重点都不相同。本文在系统地归纳和总结现有研究和应用的基础上，论证和分析共词分析过程中概念术语的词源选择、高频词选定、术语相关性计算以及统计分析四个关键问题，并结合自身的实践经验提出一些解决问题的思路和看法，希望能在理论和实践两方面为从事共词分析研究的人员提供一定的借鉴。另外，从共词分析方法的传统理论和实践惯势中跳出来，对当前数据环境下共词分析的理论和应用做一些系统性的探索，也具有重要的意义。

1 词源选择

概念术语的词源选择，也可称为分析单元的确定，是共词分析的前提和基础，也是共词分析过程中最为重要的环节，概念术语的提取质量会直接影响到共词分析的最终结果。从语言层面讲，概念（Concept）是某些特征独特组合而形成的知识单元，而术语是表达或限定科学概念的约定性语言符号，是指在特定学科领域用来表示概念称谓的集合。一门学科中的概念术语可被视为控制该学科实践的基础器件，也是构成该学科领域内科学共同体共享概念系统的基本元素[6]。科学工作者更多地借助概念术语来辨识学科的知识结构，而在共词分析中概念术语的提取主要是对文献中标题词、摘要以及关键词等不同分析单元的词的选取。词仍然是共词分析的基本元素，从不同分析单元获取表征能力较强的核心词是开展共词分析研究所需考虑的基本问题。然而，共词分析对词的选择非常敏感，文献中词的选择存在一定的随意性和惯用性，不同知识背景或认知状态的作者所具有的选词习惯，使用非规范词，词能否表征论文核心内容的完整性等，这些方面都会给共词分析结果带来较大差异。因此，有必要讨论词源选择问题，分析不同词源的选择依据，以及不同因素对共词分析结果的影响，寻求一种词源选择最优的方案。下面主要从不同类型的分析单元选择、概念术语规范化以及术语表意差异性三个方面进行论述。

1.1 不同类型的分析单元

在共词分析方法中，词源选择通常是从关键词[7-8]、标题和摘要提取词[9]、主题词[10-11]、人工标引代码[12]以及文献正文内容[13-14]等不同类型的分析单元中抽取概念术语，但不同类型的分析单元所承载的文献内容不同，使得最终的共词分析效果也不同。由于关键词本身具备标引性和便利性，且是文献核心内容的浓缩和提炼，高度概括文献的基本内容，因而较多研究采用关键词作为共词分析的单元。然而，基于自标引关键词存在一定的随意性和惯用性，即所谓的“标引者效应”。目前中文科技期刊文献仍然主要采用关键词自由标引，即不依赖于受控词表，完全由标引者根据文献主题内容自主拟词进行标引。由于标引者的知识背景、选词习惯以及对关键词的理解认识不同，造成当前关键词的标引工作仍存在通用词过多、词性不当、主题词漏选、标引深度不合适以及标引不一致等问题[15]。标引者在标引关键词时，往往只是局部重现文献研究的内容，难以准确、全面地反映文献研究内容的主题，因此，将共词分析结果完全依赖于标引关键词的共现情况并不十分可靠。Whittaker[16]指出标引结果会受到标引

者概念化科学领域方式的影响,基于这些标引词的共词分析类似于标引者自身的构想,而非科学家实际的研究工作。King[17]也提出基于不同研究领域的专业标引者所选择的关键词存在较大的不一致性。除自由词标引外,另外一个词源选择对象是借助于主题词。主题词是由相关专家严格规范化,能与概念一一对应的受控词汇。由于主题词的规范性和组配性,主题词也成为共词分析理想的分析单元。钟伟金[11]对比基于关键词和主题词进行共词分析应用时,发现两类词源在高频词、类团成员以及聚类质量上存在较大差别,且选用不同分析对象时不能简单地套用相同的数据处理过程。主题词重视语义之间的逻辑组配,在聚类效果上要比关键词表现得更加强势。但主题词同样来源于受控词表,也存在标引用词的选择问题,一定程度上会降低分析结果的客观性。

为解决关键词的这种"标引者效应",相关研究[18-21]提出利用文献中的标题和摘要提取词进行共词分析。文献的标题能够以简练、确切的词汇组合表达文献最核心的主旨,而摘要是以提供文献内容梗概为目的。利用标题和摘要提取词代替标题词,能避免标引的主观性,解决关键词的稀少、缺失问题,能更加接近文献作者的学术思维。唐晓波等[19]针对文献关键词缺失的问题,提出通过分词技术从标题中增补关键词,以获取能够全面描述文献内容的关键词。李秀霞等[20]也考虑将文献内容信息(标题、摘要和关键词)引入到作者共被引分析中,计算作者之间的语义关联。Leydesdorff[21]认为使用标题提取词可以更加接近作者的研究观点,提出使用标题词代替关键词作为共词分析的基础数据。然而,基于标题和摘要提取词进行共词分析也存在以下两个问题:一是作者为引起读者的兴趣会故意选取具有时代感、学术感的标题词,存在一定的"观众效应"[5],同时,对标题和摘要内容进行细粒度分词时,很可能会破坏词汇原本的含义;二是汉语词汇中普遍存在的同、近义词现象,同一概念在不同的标题、摘要中可能使用不同的词语进行表示。Whittaker 通过实验得出,基于关键词和标题提取词进行共词分析的结果是相似的,两者并无较大差异,最主要的区别是基于关键词共现能够更加详细地描述学科领域的结构特征[16]。

和标题摘要提取词类似的另一种做法是从文献正文中自动标引取词。文献正文内容中也包含着大量有价值的信息,Glenisson 等[22]通过研究发现:如果只将共词分析限定在标题、摘要以及关键词字段,共词分析的结果将会存在21%的错误率。Kostoff 等[23]也认为通过标引获得关键词的方法忽略了低频词汇,而基于全文自动标引的方法可以保留虽低频但较重要的词汇。为此,巴志超等[24]提出对标题、摘要和文献全文同时进行 LDA 语义建模的方法,以挖掘表征能力较强的核心关键词作为研究对象进行共词分析。索红光等[25]提出通过《知网》(HowNet)作为知识库构建词汇链实现文本关键词自动标引,结合词频与区域特征选择关键词。Anjewierden 等[26]引入本体实现对 PDF 文档的自标引,首先利用软件工具对 PDF 文档进行组块划分,再对组块进行标引。全文自动标引取词能够提高共词分析的信息利用质量,避免标引者效应及受控词表的过时问题,无需花费过多的时间来构建词表,能维持和保持词表的新颖性[5],能够较全面地反映文献的内容,充分实现对自标引关键词的增补。然而,该方法需要借助于机器学习算法融合语言学信息抽取关键词,研究难度和代价较大,抽取高质量的关键词更是一个挑战,因此在实际研究中较少采用。

1.2 术语的规范化

根据上述讨论,共词分析涉及的术语主要是从关键词、标题摘要、主题词以及文献正文内容等分析对象中抽取的词。术语是科学发展和交流的载体,反映科学研究的成果。通过审定和统一术语并实现其规范化,是保持和促进不同学科独立性的重要基础,也是使之具备与其他成熟学科渗透融合的基础条件[27]。然而,由

于汉语文化博大精深，可供表意的词汇丰富，数量庞大，涵义细致，使得不同文献中存在多个名称表达同一科学概念，或同一个名称在不同文献中表达不同科学概念的情况，使用未经规范化的术语直接作为共词分析的“构建单元”，必然会干扰数据映射的结果。如多个名称表达同一科学概念时，这些术语可能会被替换而分散使用，使得频次较低的术语会被舍弃或被过滤掉。而当同一个名称在不同文献中表达不同的科学概念时，得到的共词聚类结果也因这些宽泛、多义的词汇干扰造成类团界限模糊，无法被很好地解释和描述。因此，有必要在高频词选定之前对术语进行规范化。

术语的规范化问题包括由于术语的不规范造成共词分析时所选数据源与研究主题之间存在偏差而导致研究结果偏离实际情况的现象[5]。术语的规范化处理应该满足单义性、科学性、系统性和国际性等原则。当前术语规范主要存在以下问题。①对术语的泛用和单义性的曲解。术语的泛用表现在对同义词、近义词、缩写词及中英文混用等现象中，如“互联网”“网络”“Internet”以及“Web”等，以及在复杂网络领域中对“节点”“结点”“网络信息单元”“网络结构单元”和“node”等词的运用。这些词含义相同，但表述相异，需要对其进行整合和关联。术语单义性的曲解是指术语只依附于某一特定专业或学科范围，脱离其专业学科来笼统地使用术语，必然会造成对术语的曲解，如“文件”在档案学中和信息技术学科中是两个不同的概念，此文件非彼文件。有些术语一旦脱离专业学科背景，只采用音译，就会产生误解。②术语表达的粒度与历时变迁问题。不同粒度的术语在表征文献研究内容的主题时具有不同的描述层次，如果术语表达的过于宽泛会导致其专指性不够，无法直接用于共词分析研究。对于文献中指示性较弱的词汇，如“理论”“信息”“意义”和“问题”等，以及从背景学科继承而来的基础知识点，如“知识服务”“信息服务”等词，由于外延过宽，语义表达的过于空泛，无法解释文献具体的研究内容，甚至还会增加共词分析中共词矩阵的维数，给后续数据处理带来干扰[28]。如果术语表达粒度不够细，表达不够抽象化与具体化，也无法准确地表达作者的研究意图，如关键词“用户研究”，单从词义字面理解无法判断作者是研究用户的信息需求，还是研究用户的兴趣或行为特征，从而会造成理解歧义，因此需要对关键词“用户研究"进行细化[29]。另外，术语并非一成不变的，而是具有历时性特征，表现在时间上的动态发展和历史演变，其概念内涵与外延会随着时间的推移发生动态性变迁，名称也随之发生变化，出现一个术语有多种含义或存在多种理解与解释的情况。术语的变迁反映了术语在不同时期的规范程度，也折射出学科在不同阶段的发展水平。因此，对术语的使用，也要从历时性角度考证术语的词源，追溯术语的历史演变与动态性变迁，在不同的时期把握术语的不同概念。③新术语构造的理据和规范问题。随着社会与科学的不断发展，与之同步会产生大量对新事物、新事件以及新现象定义的新术语。一般是在继承基础上的创新，反映的是由若干简单概念抽象而来的复杂性概念，或无法用现有受控语言表达而采用自然语言进行自由标引，如采用“领域度”“领域均匀性”和“领域分布度”等新术语反映关键词对研究领域的表征能力，以及根据词频将关键词划分为“高频词”“中频词”和“低频词”等。新术语的构建不能随意标引，必须规范化，依据一定的构造理据，尽可能选自其他受控词表或比较权威的参考书目，使之词性规范、概念明确，并符合科学性、通用性的需求[30]。

当前，对术语规范化的处理方法主要分为两种：一种是基于受控词典或分类词表进行术语规范，另一种是基于人工干预方式的规范处理。前者可借助于词表对收集到的术语进行规范，或直接借助已规范的术语如主题词、叙词表进行共现分析。如张晗等[31]以 MetaMap 为工具将文本中的关键词映射为 UMLS 本体库中的概念词，然后借助 SemRep 实现源文本主题概念及其语义关系的规范化；Ding 等[32]采用关键词、

标题词以及摘要提取词作为共词分析的对象，并借助于 TITT、LISA 和 LCSH 三个叙词表对关键词的单复数、统一性(同/近义)进行整合和关联，以降低术语泛用对共词分析结果的影响；Hu 等[7]也采用叙词表作为基准对术语进行标准化，合并、修正术语并过滤过于宽泛的通用术语。也有学者如钟伟金[11]、崔雷[33]、许海云[34]等为排除关键词标引不规范性的影响，直接借助受控术语进行共词分析。人工干预的处理方式更多的是借助自我的经验制定相应的规则、方法来实现人工干预处理。如李武等[35]针对中英文混合、英文缩写以及英文关键词等不同形态的表达进行统一替代处理。陆宇杰[36]等针对术语表达粒度过于宽泛或抽象、具体的情况，通过人工审核，采用融入、整合、分解三种不同思路，对零散的、无单独存在意义的词进行规范化，以保证每个术语的高内聚、低耦合。Zong 等[37]通过合并同近义词和去除过于宽泛的词语来对博士论文中的关键词进行标准化。Dehdarirad 等[38]为精炼数据集(如不同表达方式的关键词、单复数、大小写等)，采用人工方式进行归并、整理，将指向同一概念的术语采用词频最高的词作为基准术语，并将高度相关但概念不同的术语进行分离。当前由于受到中文信息处理、语义理解等技术条件的限制，术语的自动规范化处理还未完全实现，更多的还是借助于计算机辅助术语规范工作，由于人工标引过于费时、费力，术语规范的自动化处理是亟需解决的关键问题。

1.3 术语表征差异性

共词分析方法中，传统的理论是假设术语的独立性，而不考虑术语之间表达的差异性。如在基于关键词共现分析过程中，统计共现频次时同等对待每个关键词而不考虑关键词的权值，即对于在文献中出现频次相同的关键词，其重要性可视为一致，这种假设显然是不合理的。从信息理论学和语言学的角度来讲，不同标引源所承载的文献内容不同，不同词性、位置等要素的关键词对文献表达的贡献程度也是不同的[39]，有些关键词能够表达某学科领域特有的核心概念，而有些关键词却只能代表一些辅助性、边缘的概念或主题。因此，忽略关键词之间的权重差异必然导致最终的共词分析效果存在一定程度的失真。相关研究成果[40-43]也证明，有效地区分术语之间的差异性，考虑文献的篇章语义结构以及术语“同量不同质”的现象，在一定程度上能够很好地改善共词分析的结果。侯汉清等[40]计算中文图书的主题与图书的书名、内容提要、目次、参考文献等不同标引源之间的关系，分析不同标引源的主题表达能力，在此基础上对标引源进行加权设计用于图书自动标引。李纲等[41]探讨了基于关键词加权的合理性和必要性，提出一种基于关键词加权的共词分析方法，通过赋予标题和摘要中更加契合论文主题的关键词以更高的权重，从而获得更好的共词分析结果。吴清强等[42]认为共词分析忽略了对共现词所在文献属性的考虑，认为发表在高影响因子的期刊或被引次数较高的文献中的词更具有代表性，并根据文献的来源期刊、被引次数等属性赋予关键词不同的权值。钟伟金[43]认为在文献的标引中存在主要主题词与次要主题词的差别，通过加权计算凸显主要主题词在聚类过程中的主导地位，减弱次要主题词的干扰，能够达到降低共词聚类敏感性的目的，通过实践数据验证了这种加权共现分析方法的有效性。

综上可以看出，从标引源差异、文献属性差异以及词重要性差异等不同角度进行加权共词分析方法是必要的、合理的。加权共词分析方法是一种在传统共词分析方法的基础上，通过映射函数考虑不同影响因子对共词之间概念关系影响的分析方法[42]。通过构建不同的映射模型，可以对概念之间的不同关系进行强化或衰减，从而展现出不同的研究目的。例如，对文献全文中的前言、结论和讨论等部位进行区分，或对文献段落、句子或词的位置进行加权，可以挖掘更细粒度的研究内容；对文献发表的时间进

行加权，可以获取所在领域的最新研究热点问题。在不同类型文献单元的选择和术语规范化处理过程中，主要问题是如何借助自然语言处理实现关键词的自动抽取和规范标引，而在术语表征差异性问题上更多的还是需要考虑如何对术语的差异性进行度量。通过考虑不同的加权策略，将差异性度量方法引入到传统的共词分析方法中，也是一个值得研究的课题。

2 高频词的选定

由于受到工具、人力的限制，以及为了结果分析、呈现的需要，研究者通常会选择一部分词作为共词分析的对象。然而，关键词词频的幂律分布特征给选词工作带来一个难以调和的矛盾，如何设定关键词的筛选标准以界定待分析对象也是一个难以解决的问题。词频是关键词筛选最直接的方式，较多的研究为简化统计分析的过程，降低低频词对统计过程的干扰，以频次阈值简化的方法选择高频词作为共词分析的对象。选用高频词进行共词分析能够有效地揭示领域的研究热点和主题，但现有研究并未对高频词数量的选择达成共识。如果选择过少，不能有效涵盖研究领域知识点的整体分布，无法获取知识网络在多尺度上的结构和特征信息；如果降低词频阈值扩大高频词选择的数量，即会增加数据处理的成本，同时也无法兼顾“有效揭示领域的研究热点和主题”与“涵盖研究领域知识点的整体分布”两者之间的平衡。因此，有必要对高频词选择涉及的关键问题做出系统的分析。下面从高频词阈值的设定、术语语义类型特征以及低频关键词处理三个方面进行阐述。

2.1 高频词阈值的设定

对高频词阈值的设定主要有两种方法：一种是基于经验判定法，研究者凭经验在选词个数和词频之间取得平衡；另一种是结合齐普夫第二定律——低频词分布定律和 Donohue[44] 高低频词分界公式判定高频词的阈值。如 Wang 等[45]、Sedighi[46] 对所有术语词频进行统计并从高到低排序，根据个人经验选取前 N 个高频词作为分析的样本数据。Zhang 等[47] 利用 NoteExpress 2.0 工具获取 4 575 篇文献中的所有关键词，并选择词频不低于 26 的 178 个关键词进行分析。为了在获取研究领域的热点的同时能揭示热点之间的关联关系，相关学者[16,48] 提出同时考察词频和共词，或共引频率选择高频词。但这类方法都是凭借研究者个人经验选词，主观性较强，缺乏一定的理论指导。为此，Yang 等[49] 借助 Donohue 提出的高低词频分界公式获取医学信息学领域 35 个频次超过 36 的高频 MeSH 词语，Yan 等[50] 设定 $T(T=(-1+\sqrt{1+8I})/2)=120$ 时，只得到 7 个频次超过 120 的高频词。齐普夫定律是以低频词（词频为 1）作为高低频词分界的依据，当研究领域范围过大时容易产生太过抽象、具体的词以及领域外不相关的词语。另外，鉴于汉语的特殊性，同近义词过多导致中文文献中的词频分布和英文文献存在较大差别，特别当词频分布中存在较多极高或极低频次的术语时，齐普夫定律确定阈值的效果并不理想。Zhang 等[51] 和杨爱青等[52] 又提出通过 g 指数来确定高频词的阈值，通过真实数据验证 g 指数要比齐普夫定律效果好，克服了齐普夫定律对低频词规范化的限定，并降低了术语选择的主观程度。

除上述方法外，相关学者还提出通过新的指标（如网络节点度数、特征向量中心性、中介中心性等）或方法（K-core 分解、惩罚性矩阵分解、词汇链、核心/边缘结构模型等），将术语集合转化成网络模型实现术语的抽取。但由于在术语构建的网络中，上述指标与词频仍然线性相关，因而抽取的术语与高频词并无太大差异。上述研究方法也缺乏对共现词和对语义关系的揭示，只是直观地假设共现就必然存在相关，且认定在共现词数相同的情况下术语之间的相关强度完全相同。在确定研究的主题和分析的问题后，不应完全根据绝对词频获取高频词而忽

视术语之间的关联。尽管有些术语会在文献中经常出现,但与文献中其他术语的关联性不大,或与文献的主要研究内容或主题并无关系,基于这些术语进行共词分析,在揭示领域研究热点时必然与实际存在偏差。如果考虑术语在文献中的语义内涵,挖掘术语在其背后隐藏的语义信息,通过语义建模获取能够对文献表征能力较强的核心术语进行共词分析,可进一步提高共词分析的科学性和有效性。另外,研究的主题和分析的问题必然有其依托的背景学科,以开阔的视野从局部、全局等视角考察术语在领域内、外的统计特征,分析术语与学科领域的关系。通过区分术语对学科领域的揭示能力,获取更能够表征领域研究特色的术语作为共词分析的对象,相比传统的高频词分析方法,更能全面、深入地了解领域的研究特点。

2.2 术语语义类型特征

研究者在进行高频词选择时都有其目的性,高频词通常用于交代学科领域的产生与发展史,分析研究主题结构,标明研究领域的归属,限定研究成立的范围,描述研究依据的理论与方法,以及概括研究涉及的知识内容[53]。然而,根据术语不同的行使功能,也可按照上述目的将其划分为不同的语义类型。当研究者在明确其研究问题后,可选择具有相同语义类型的术语进行共现分析。如研究者在分析领域主题的结构时,可优先、偏向性地选择表达“研究主题”“理论方法”的术语进行分析,这样得出的结果会更加明显、清晰。胡昌平[53]将论文关键词划分为五种语义类型:研究主题、所属领域、限定范围、理论方法、子知识点,针对情报学科人工标注了部分随机数据集,并统计不同类型之间的共现比例,通过分析得出以下结论:情报学学科中跨领域、跨主题以及多理论方法融合的研究相对较少,关键词的语义类型之间存在共现的集中性和同类的互斥性。上述结论有一定的合理性,但也存在一定的缺陷。不同类型的术语有其特定的流通领域,且只在特定领域流通度较高,术语在不同领域类别之间不均匀分布,会导致跨领域、跨主题的关键词相对较少。但单纯地只统计自标引关键词的语义类型无法全面、有效地解释学科领域的研究特点。作者在标引关键词时往往会选择研究主题、研究子知识点两种类型的词,而较少选择文献涉及的多种研究方法、限定研究范围等类型的词,标引的随意性和惯用性也会导致统计结果存在偏差。本文提出的考虑关键词语义类型的特征对高频词选择、共词分析的影响,不失为一个好的研究思路。

统计不同研究领域中术语语义类型之间的共现情况,并按照类型之间的共现强度构建不同的术语链,如知识点—主题—知识点、领域—范围—主题等,链内相连的语义类型之间共现概率高,且语义可靠性更强。基于这种思路可为术语差异性处理和术语缺失时有目的性的增补提供一定的借鉴。例如当确定研究问题后,可获取相应语义类型的术语进行分析;当该语义类型的术语相对稀少、缺失时,可根据该领域内术语语义类型的共现强度有针对性地选择同类型的术语或链内的术语进行增补;当计算不同类型术语之间的共现时,可对链内术语之间的共现关系进行加权处理。这种思路也使得对传统共词分析过程的扩充成为可能,新的共词分析过程可分为:确定研究领域与分析问题—术语的词源选择—统计术语的语义类型和共现关系—相应类型术语的选定—共词频率计算—多元统计分析—结合相关的学科知识对统计结果进行科学分析。

2.3 低频词的处理

高频词的选择是以牺牲低频词信息为代价来获取易于分析的研究样本,传统基于高频词的共词分析方法,通常是将低频词直接过滤掉,只选用高频词进行分析。而在共词分析中是否需要兼顾表意性较强且含有大量重要关系的低频词对共词分析的影响,一直是一个比较难以回答的问题。由于借助高频词进行共词分析方

法的利弊难以权衡，从未有学者对该问题做出合理的解释和说明。文献中术语的词频和共现关系强度满足的幂律分布特征本身给共词分析带来一个天然的、无法逾越的障碍，如何在保留高频词优势的同时，弥补其对共词分析的缺陷也是一个值得研究的重要问题。为深入地探究高频词矩阵对领域内重要共现关系的保留程度，胡昌平等[53]分析了共现关系强度分布特征对高频词矩阵构建的影响，发现共词网络实际上是建立在不可靠的单次共现关联基础之上，网内关联极度稀疏，且两个高频词之间的共现关系并非一定是强关系，而大部分拥有强关系的词由于词频较低而未被抽中，高频词矩阵对重要共现关系的保留程度并不让人满意。Serrano 等[54]、Price[55]通过实验得出，仅基于高频词的共词分析难以有效地涵盖研究领域中知识点的整体分布，而保留一定比例的低频词更有利于揭示领域研究主题的热点和发展趋势。张玉芳等[56]认为在信息抽取中，某些稀有词要比高、中频词更能反映整体的特征，而不应该被滤除。邱均平等[57]认为低频词是高频词的重要补充，从不同维度对近几年新增或增速较快的低频词进行透视和分析，推断出未来我国图书馆学研究存在八个微观发展主题。赵辉等[58]在分析国家自然科学基金项目材料领域项目演化时，将关键词按照阈值划分为三类：以高频词作为常用词，以中频词作为骨干词，以低频词作为弱相关词，并认为常用词研究意义不大，过滤常用词，重点分析了其中的骨干词。

综上所述，用高频词代表领域整体的研究方向存在着天然的缺陷，而低频词却有助于获取一些隐含主题或前瞻主题的信息。在今后运用共词分析方法时，应该充分意识到孤立地以热点为依据筛选高频词并不能很好地表征领域的研究特点，只单纯地降低词频或共现频次阈值增加高频词的数量也难以有效地改善共词分析的结果。一种思路是在获得一定数量的高频词后，借助机器学习算法挖掘表意性较强且含有大量重要共现关系的中、低频词对其进行补充，两者结合形成新的共词矩阵进行分析，这样在兼顾数据处理与分析成本的同时，也能够最大程度概括领域内的知识全貌。

3 术语相关性计算

在高频词选定以后，需要两两统计高频词之间的共现关系，据此构建成共词矩阵以便进行统计分析。传统的共词分析方法通常构建共现矩阵对术语关系进行量化计算，然而，构建的二值/多值矩阵中统计的原始词对频次是绝对值，难以反映词与词之间真正的相互依赖程度，缺乏对术语对之间语义关系和关系强度的揭示。同时，多值矩阵中存在的频次悬殊数据以及较多的零值会对最终的统计结果造成影响。为此，相关学者提出采用关键词共现指数表达的方法，通过引入关键词共现相对强度指标对词对频次进行包容化处理，以生成相似矩阵和相异矩阵，如采用 E 指数[59]、Ochiia 系数[60]等。但这几种方法只是为减少低频词对共词分析过程的干扰，以区分对待低频词以及高频词之间的共现强度，仍无法挖掘出关键词之间的语义关联信息。

文献是由若干个具有语义信息的术语按照一定的逻辑结构排列组合而成，这些术语不仅存在着物理位置上的相关，还存在句法、篇章结构上的支配从属关系以及隐含的语义关系[61]。再加上汉语术语本身具有的语义模糊性以及术语之间关系的不确定性，如果仅将共词分析停留在词频、共词频率统计的语法层面，必然导致共词分析结果的失真。然而，目前大部分的共词分析研究往往只偏重于高频词的抽取，而忽视术语之间“边”的考虑，割裂了术语之间的关系，缺乏对术语之间语义关系的识别。因此，在共词分析中不仅要考虑术语之间的直接共现相关，还要考虑术语之间的间接语义相关。另外，传统共词分析方法只是直观地假设共现就必然相关，且认定在共现次数相同的情况下术语之间的相关强度也完全相同。然而，从单篇文献

以及学科领域范围内的整个文献集合研究角度而言,直接共现相关和间接语义相关的强度是不同的,且不共现的术语之间也存在一定的相关性。下面从术语语义相关性的计算方法和术语相关特征分析两个方面进行阐述。

3.1 术语语义相关性计算方法

术语语义相关性表示两个术语在同一分类体系下通过各种语义关系产生的关联程度,这与语义相似性不同。语义相似性是一种特殊的语义相关性,只考虑术语在分类体系中术语之间的上下位关系,而语义相关性则是涵盖各种类型的语义关系。已有技术方法主要有以下三种实现思路。

一种是基于文献内容的方法,借助文献集的统计规律特征间接反映术语之间的相关关系。如 Wang 等[62]提出通过计算术语在文献中的词频、信息量、词频共现等指标统计其内外部特征,并采用 DT-SVM 算法实现对术语的有效抽取,在此基础上整合欧几里得距离、Jaccard 系数等相似度方法实现对术语关系的语义计算。Li 等[63]认为仅根据自标引关键词进行共词分析存在着较大的局限性,提出在 TF-IDF 算法的基础上采用不同的加权策略从标题、摘要中抽取关键词,利用关键词自身权值替换频次,并结合有向亲和系数(Directional Affinity,DAff)来度量关键词之间的关联度,用于构建领域特征词分类体系。Ohsawa 等[64]基于动态的窗口滑动模型统计术语在句子中出现的位置、频次等信息,基于语义计算术语之间共现的强度,通过实验也验证,考虑语义信息的共现度量方法比单纯基于词频、共现频次的方法更能有效计算术语之间的关系。这类方法主要从篇章连贯性、主题分布、词频分布等角度考察并分析术语与文献结构(如文献标题、摘要、首尾段、句群等)之间的统计关系,并通过融入“显性”和“隐性”文本结构来度量术语之间的语义相关度。但该方法需要提取出有效的文献结构,并根据不同的文献结构制定不同的相似度计算策略。另外,不同术语在不同的文献中所表达的概念不同,该方法只是借助文献的内部统计特征,忽略了术语概念之间的差异性问题,未考虑文献与术语表达的主题一致性。

另一种方法是基于本体的方法,强调对术语的语义理解是术语之间关系判定的基础。通过在外部现有知识资源基础上的全局映射实现对术语概念的解释,利用 DBpedia、Cyc、知网(HowNet)、WordNet 等通用本体库以及 MeSH、OpenGALEN 等领域词典的知识体系,实现对术语语义相关性的定量计算[65-69]。如 Aouicha 等[65]考虑概念在分类体系中的信息含量、深度、上下位层次关系等属性,基于两个概念到其最近公共祖先节点的距离计算概念之间的语义相关度,并尝试借助 Wikipedia、WordNet 的概念定义和 Wiktionary 的词义描述三个不同类型的语义内容,通过加权机制计算相关性。Tastsaronis 等[66]则综合考虑 WordNet 中的多种语义关系,对不同的语义关系赋予不同的权重,并考虑路径长度和路径上概念节点的深度对语义相关性的影响。Roberto 等[68]提出基于 BibleNet 度量术语间的语义相关性,通过在不同语种的 BibleNet 中查找术语概念之间的所有路径并构建语义图,然后合并语义图获取核心子图,在此基础上基于路径长度的方法计算术语之间的语义相关性。借助本体方法这种预定义的语义描述或结构化路径可以实现对术语之间语义相关性的定量计算,尤其是针对同、近义词关系的识别方面能够获得较好的实验效果。但基础知识资源建设所固有的自动化程度低、更新缓慢、建设成本高、知识相对滞后等局限性,导致目前已有资源无法满足实际应用的需要。另外,人类语言的复杂性使得构建一个完美的本体知识库是不太可能的,一些本体知识库可以较为准确地反映术语之间的语义关系,但对于术语之间的句法、语法等特点考虑较少,一些特殊的语义关系在知识库中可能并不存在。

第三种方法是基于大规模语料库的方法,通过对大规模数据的统计分析度量术语的相关

性。如 Jorge 等[69]对规格化的 Google 距离法(Normalized Google Distance，NGD)进行扩展，提出基于多种 Web 搜索引擎度量术语之间的语义距离，通过将待比较的术语用两个等级的语义描述(同义词描述集、上下文描述集)进行表示，然后计算两种语义描述之间的相关性，将两种相关性线性合并得到术语之间的语义相关性。该方法弥补了术语相关性度量中覆盖率低、领域相关、普适性差的缺点，但在计算过程中也忽略了术语在网页中的位置以及重复出现等情况。Zhang[70]将人类认知思维的过程引入分布式度量方法中，提出一种基于关联网络的方法。首先利用自由联想法构建关联术语表；然后为解决关联术语表的覆盖率和稀疏性问题，从 Wikipedia 知识库中提取五种类型的共现术语对，对关联术语表进行扩充并生成语义关联图；最后将待比较术语映射到语义关联图的概念节点，通过计算概念节点之间的语义关联判断术语之间的语义相关性。基于语料库的分布相关性度量方法比较客观，可以不受概念层次结构的约束，综合反映术语在句法、语义、语用等方面的相关性和差异。但这种方法依赖于训练所用的语料库，而语料库通常是自动生成，存在着数据的不平衡性和数据稀疏性等问题，且计算量大，计算方法较为复杂。考虑到单一的语义相关性度量方法各有利弊，相关学者[71-72]将两种方法进行结合，提出混合式的度量方法。如 Yih 等[71]首先利用异构数据源(Web 检索结果、文本语料库、词库)构建术语的向量模型，然后利用向量之间的余弦相似度计算术语之间的语义相关性。Wang 等[72]通过 K-means 算法对语料库中的术语进行多层次聚类以构建术语的分类层次结构，进而在该结构中的不同路径上采用层次聚类和多维尺度分析计算术语之间的语义相关性。但该方法在前期构建术语的分类层次结构时，需要进行多次的聚类工作，复杂度较高。

3.2 术语相关特征分析

对术语进行词频统计时，术语所体现的“符号学”意义要大于它的“语义”意义，对术语所处的具体语境不予考虑。而在计算术语之间的相关性时，其所体现的“符号学”意义要小于“语义”意义。术语之间的相关性本质上是由术语所指代的领域知识概念和其所处的语境共同决定的，因此获取术语的深层次语义相关是建立关联的根本途径。术语之间的语义相关具有传递性、层次性、领域性等多种特征。其中，传递性反映的是术语之间一种特殊的间接相关。一方面，根据三元闭包理论和关联规则分析，当两个术语同时与某一个术语有着较强的相关性时，这两个术语之间很可能会存在间接的相关，而且这种间接的相关性基于传统共词分析方法是无法捕获到的。另一方面，在基于本体的预定义语义描述或结构化路径计算术语语义相关性时，构建以共同概念为重合点、多个语义关系依次连接的链式架构，也存在同种或异种关系的传递性，表现为不同语义关系类型的合成。然而，由于语义传递判断的主观性和计算机模拟过程的挑战性，通常只考虑概念体系中深度、密度、语义关系类型和路径四个因素的影响而忽略传递因素，只是将其作为语言认知研究中的附属过程，对间接相关性不加区分地统一对待[73]。因此，在计算术语的相关性时，考虑术语之间语义关系的传递性特点，以术语三元闭包为对象分析概念知识的关联特征，能够更深层次地细化语义层次，挖掘从未发现过的语义相关。

术语语义相关的层次性是考虑术语语义类型的差异性特征，根据术语所表达的语义概念对术语进行聚集、融合和细化，以构建术语相关的层次化结构。术语之间的关系主要包括等同关系、相关关系和从属关系三大类，基于等同关系在预处理环节常常被归一化，基于相关关系以及同知识层次的从属关系，可以识别术语之间的同、近义关系，而基于不同语义层次的从属关系，则可以判定具有知识延续性术语关系的远近。Xiao 等[74]针对共现网络结构分析方法中节点类型较为单一、难以揭示知识点之间的关

联差异等问题,提出以 K-core 为指标对概念知识网络进行分层,根据术语所表达的语义宽泛程度,将术语分为基础层、中间层和细节层,由内向外语义范围逐渐具体化。这样通过与上层节点的关系可以归纳出基础研究,通过与同层、下层节点的关系归纳它们与相关研究融合形成的新研究主题,通过与下层节点的关系可以归纳它们在研究中独立分化出的新主题类团。从术语语义的差异性视角出发,依据语义划分为不同层次,将同一语义层次的术语共现看作相关关系,不同语义层次的术语共现看作层次结构中的从属关系,从而可以丰富术语之间的语义关系,更深入地揭示研究主题的微观结构形态。

另外,从语言信息处理的角度来看,术语有其专用性和规定性,不同类型的术语有其自身特定的流通领域,只有在特定领域内其流通度才较高。而且术语在不同领域类别之间也存在着不均匀分布的特性,相关学者[75-76]常用领域相关度、领域一致度等指标来反映术语在相同领域的相关程度以及在领域文集中的分布情况。术语是描述领域中所对应的概念的,在其概念所代表的领域主题范围内才是分布均匀的。因此,基于术语的这种特性,术语的语义相关也存在着领域性特点。忽略术语流通的背景领域,直观地去计算术语之间的语义相关度,在分析其所代表领域的研究热点、把握领域的发展过程以及未来趋向时,必然存在一定的误差。另外,由于研究主题的背景不同,从单篇文献、领域子集以及整个领域文献集合的角度而言,术语之间的相关性强度也是不同的,在基于共词分析方法计算术语之间的相关性时,考虑术语这种语义相关的差异性,是否更有利于提高共词分析方法的科学性,也是一个值得思考的问题。

4 多元统计分析

在计算术语之间的相关程度并构建成相关矩阵后,需要借助不同的统计学分析方法揭示矩阵中的信息。由于最初确定的研究问题不同,采用的统计学描述方法也不尽相同。常用的多元统计分析方法主要有共词聚类分析法、共词关联分析法、共词词频分析法以及突发词检测法等。其中共词词频分析法是以某一领域文献中术语出现的频次高低来确定该领域的研究热点和发展动向,这种方法在文献学分析中比较常用,但局限性也较为明显,在前文中已有详细探讨。而突发词检测法是从关注术语自身发展变化出发,考察术语频次增长率的阶段性特征,来探究学科(或主题)研究发展过程中的微观因素。该方法虽然与高频词截取方法不同,但也是限定于研究特定术语的频次随时间的动态演化特性,与共词词频分析法并无本质的区别。因此,本文主要以共词聚类分析法与共词关联分析法为例,对两种分析方法进行总结,并探讨如何改进应用。

4.1 共词聚类分析法

共词聚类分析法是将术语之间的语义关系通过术语之间的共现关系进行表达,通过聚类算法将不同相关程度的术语聚集成不同的类团,以实现对领域结构的分析研究。基于该方法能够将众多术语之间错综复杂的相关网状关系简化为数目相对较少的若干类群之间的关系,并直观地表达出来。然而该方法也存在聚类过程中心缺失、成员聚集缺乏方向性、聚类结果不明确等问题[43]。术语的选择以及术语之间相关程度计算的准确性、科学性,是共词聚类分析成功的关键。钟伟金[11]分别以关键词和主题词为对象进行共词聚类分析,发现基于主题词,无论在类团表现上,还是在类团相关性、聚类质量上,都具有更好的效果。同时,提出区分对待主要主题词和次要主题词,通过对主要主题词加权进行共词聚类分析,能够有效优化共词聚类的结果。共词聚类过程主要取决于术语之间的相关依存程度,由于此过程中缺乏明确的聚类中心,导致在类团划分时容易产生“马太效

应”，使得与某个成员关系密切的其他成员也会被同时吸引进同一类团，大的类团将变得更大，小的类团会变得更小。这种吸附作用也会导致弱势类团得不到有效的生成，造成聚类结果不能全面反映学科的研究点，甚至还会出现一些与其他术语相关性较高、但表意性较差的术语“喧宾夺主”的现象，一个类团中出现较多只起到修饰、限定性作用的术语，而一些主要核心术语的作用被弱化，致使整个类团主题不明确，甚至没有意义。另外，聚类过程也较为敏感，术语对之间关系的细微变化，都可能对聚类结果产生较大影响。因此，有必要采用相应措施适当干预术语对之间的关系，弱化聚类分析对术语关系的过分敏感，以确保聚类结果具有一定的稳定性。

借鉴高频词选定中表征能力较强的核心术语的抽取策略，在类团分析过程中引入核心术语，提高核心术语的作用，排除非核心术语之间的依存关系对类团造成的干扰，不仅能够解决类团划分中的问题，还能够弥补聚类算法中存在的一些其他不足，如对类团属性的判断，研究点主题词分布，小类团的补救，以及类团之间关系的判定，都具有重要作用[77]。李佳等[78]为消除聚类计算过程中出现的上述问题，进一步明确类团内与类团之间的关系，提出通过粘合力指标反映类团内各成员的地位与作用，通过向心度反映类团之间的相关作用与地位。实验结果表明，通过对类团核心词的指定，不仅可以实现对类团概念的正确定义，还使得类团之间的关系分析变得更加具体清晰。然而，在计算术语之间的语义相关性，考虑术语之间的直接共现相关和间接语义相关时，术语之间都会建立一定的联系，不共现的术语之间也会存在相关性。因此，单纯根据上述提出的指标，通过计算共现频率平均值的方式，无法实现对类团内各成员地位的有效区分和度量，同时，以往对聚类结果采用一刀切的划分方法也容易造成聚类不全及术语只能出现在某一个类团中等现象。因此，有必要提出新的度量指标和方法，改进聚类的算法和策略，以解决聚类算法与研究点术语分布不一致、聚类结果划分缺乏弹性处理等问题。

4.2 共词关联分析法

共词关联分析法是根据文献中术语之间的逻辑关联关系，在满足一定支持度和可信度的条件下，通过关联规则寻找出频繁共现的两个或多个术语组合。如李牧南[79]提出构建基于摘要文本的突现词和主题关键词的关联规则挖掘模型，通过“关键词”和“突现词”的直接关联获取研究的前沿热点。相比于传统共词分析，该方法在研究前沿的发现与甄别时揭示的信息更加丰富，起到一定的辅助性作用。崔雷等[80]尝试根据书目文献数据库中主题词/副主题词之间的语义关联规则抽取知识，通过选取经过验证的关联规则获取具体的药物与疾病之间的知识，并通过实例验证该方法具有一定的可靠性。高继平[81]等针对当前共词分析主要关注两个术语之间的共现关系，而忽略多词共现的知识结构的现象，提出采用关联规则挖掘算法实现多词共现分析方法。在共词关联过程中会涉及四个重要的概念：支持度、可信度、期望可信度和作用度。支持度描述的是两个术语在文献集中同时出现的概率，即 $\mathrm{Support}(A\rightarrow B)=P(A\cup B)$；可信度描述的是在术语 A 出现的前提下，术语 B 也同时出现的概率，即 $\mathrm{Confidence}(A\rightarrow B)=P(B\mid A)$，反映的是关联规则的预测强度；期望可信度表示术语 B 在所有文献集中出现的概率，即 $P(B)$；作用度表示可信度与期望可信度的比值。基于共词关系的关联规则方法，无需在特定背景知识的前提下实现有效关联，但是由于不考虑文献篇章语法结构、不同标引源所承载的信息量、术语表达的差异性，以及术语之间存在的直接共现相关关系和间接语义相关关系等问题，只是通过共词统计和数据挖掘的手段获取术语之间的关系，必然会导致结果出现较大误差。同时，当设定较小的支持度和可信度时，会产生大量的冗余规则，导致规则的精确

度也会大大降低[82]。因此,有必要结合术语本身的语义类型、历时性特征,以及术语之间语义相关的领域性、层次性和学科性等特点,构建具有时态约束的加权、多层次的扩展型关联规则,通过较少的规则表示全部的有效关联信息。

5 总结与展望

通过上述归纳和梳理可知,共词分析方法的基本元素仍然是词,对词的选择以及词对的处理都具有较强的敏感性,共词分析的不同阶段都会对共词分析结果产生较大的影响。另外,共词分析方法往往只可以用于某具体的学科结构或研究领域分析,成立的条件也通常是基于多种假设前提:文献中标题、关键词等专业术语都是作者经过深思熟虑后认真选取的,能够代表文献的核心内容,反映该领域的研究现状;同一文献中使用的不同术语之间存在一定的相关性,且当有较多作者同时选择使用某一对术语时,这对术语所表达的关系在作者所属的研究领域具有特别重要的意义;作者标引的关键词是一种规范性的科学概念。只有当上述假设前提都成立时,共词分析方法才具有科学性。然而,上述假设前提与现实中的实践是有一定距离的,文献中词的选取存在着一定的随意性和惯用性,不同知识背景或认知状态的作者所具有的取词习惯,非规范词的使用,词能否表征论文内容的完整性等,都会导致共词分析结果存在偏差。为此,本文以共词分析过程为基准,研究共词分析过程中不同阶段存在的局限性,总结出共词分析方法的研究框架(见图1),从而完善和优化在实际应用中的具体方案,并指出未来研究中需要深化的方面。

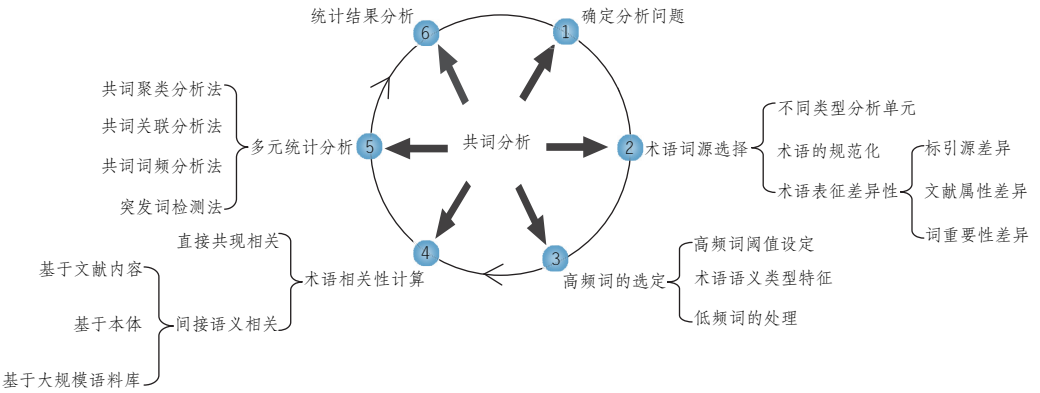

图1 共词分析过程研究框架

本文研究的结论主要包括以下方面。

在词源选择方面,论述不同类型分析单元的选择、概念术语规范化以及术语表意差异性三个方面对共词分析的影响。首先,仅根据标引关键词和主题词作为共词分析的"构建单元",将数据映射的可靠性完全依赖于标引结果的做法并不十分可取,标引的随意性、主观性、不一致性等问题往往会导致共词分析的结果产生较大的偏差。标引关键词和主题词往往只是局部重现文献研究的内容,难以准确、全面地反映文献的核心主题,而借助从标题、摘要、关键词、文献正文等不同文献分析单元抽取、挖掘表征能力较强的核心术语作为共词分析的研究对象,能够有效提高共词分析的信息利用质量,避免标引者效应以及受控词表的过时问题。其次,术语的规范化处理对共词分析结果具有重要的意义。当前对术语的使用存在以下问题:术语的泛用和单义性曲解,忽略术语表达粒度

与历时变迁，缺乏新术语构造的理据和规范性等。审定和统一术语的规范化，能大幅度提高共词分析的科学性。但现有术语规范化处理仍是以人工方式为主，还未能自动化规范，实现术语的自动化规范处理已成为共词分析领域亟待解决的重要问题。最后，考虑术语表意的差异性，从标引源差异、文献属性差异以及词重要性差异等不同角度进行加权共词分析是必要的、合理的。

高频词的选定方面，当前共词分析的研究人员仍未走出以截取高频词为基础进行多元统计的套路。提取高频词进行共词分析的方法不仅将低频关键词边缘化，忽视代表新研究方向的关键词，还会造成高频词的孤立，最终导致词与类团的相关性较低，严重影响共词分析的有效性和可靠性。本文从高频词阈值的设定、考虑术语语义类型特征以及低频关键词处理三个方面，分析高频词的选定对共词分析的影响，发现简单地根据频次阈值选择高频词作为共词分析的对象，在揭示领域研究结构和热点时与实际情况存在一定偏差。首先，考虑术语的学科性特点，以开阔的视野从全局视角考察术语在文献中的语义内涵，通过区分术语对学科领域的揭示能力，获取更能表征领域研究特色的术语作为共词分析对象，更能全面、深入地挖掘领域的研究特点。其次，结合术语的语义类型特征，根据不同语义类型之间的共现强度构建术语链，能够为术语的差异性处理和术语缺失时有目的性的增补提供借鉴，也使得对传统共词分析过程的扩充成为可能。最后，低频词有助于获取一些隐含主题或者前瞻主题的外观，在获取一定数量的高频词后，挖掘表意性较强且含大量重要共现关系的中、低频词对其进行增补，这样在兼顾数据处理与分析成本的同时，能够最大程度概括领域内的知识全貌。

术语相关性计算方面，传统共词分析方法只是将文献简单地模拟为“词汇包”，通过词对频次的二值或多值矩阵来量化术语之间的关系，缺乏对术语对之间语义关系的度量。这种思路只是将共词分析停留在共词频率统计的语法层面，忽视术语之间“边”的考虑，割裂术语之间的多种关系，必然导致共词分析结果的失真。术语之间不仅存在直接的共现关系，还存在间接的语义相关，在总结现有术语语义相关性计算方法以及各方法存在的局限性的基础上，对术语相关性特征如传递性、层次性、领域性进行分析，提出有效提高共词分析方法科学性的建议和思路。最后，在多元统计分析方面，以共词聚类分析法和共词关联分析法为例，对两种分析方法进行总结并对其应用改进做出探讨。

共词分析方法以其灵活性和分析结果的直观性，已得到广泛应用。但该方法在应用于不同的数据环境时，实施过程中还存在一些弊端，本文也试图从共词分析方法中传统的理论常态和实践惯势中跳出来，对当前数据环境下共词分析的理论和应用做一些系统性的探索。未来的研究主要包括以下内容。①文献中术语之间的关系类型包括位置关系、概念关系、数量关系、变异关系以及共现关系等，当前研究更多考虑术语的共词和变异关系，从某一侧面体现术语的一个角度特征关系，无法全面地反映术语之间的语义关系。未来研究应融合各种关系，在考虑语言结构特征的同时，结合语义概念结构获取术语之间的关系，进行共词分析，更能全面、有效地反映学科领域的结构特征。②由于关系媒介的不同，基于共引分析法在描述文献结构特征分析对象之间的关系时，往往会考虑方向性问题。而基于共词分析方法只是通过提出测度指标定量计算数值，如借助共词分析方法探究作者合作关系、研究兴趣的相似性时，只是简单地测量作者合作强度、研究兴趣的相似程度，而不考虑关联的方向性。然而作者之间的合作关系、研究兴趣的相似性也是存在方向性的。由于地位和角色的不同，某位作者的研究兴趣可能是多方向的，而有些作者的研究兴趣只呈现单一方向，甚至是该作者多个方向中的一个分支。如何从地位与角色分析的角度探究合作关系、研究兴趣相似性等关系的方向性

问题,并借助数学内积形式来表示这种方向的　差异性研究,也是一个值得探讨的课题。

参考文献

[1] Liu L Q,Mei S Y. Visualizing the GVC research:a co-occurrence network based bibliometric analysis[J]. Scientometrics,2016,109(2):953-977.

[2] 冯璐,冷伏海.共词分析方法理论进展[J].中国图书馆学报,2006,32(2):88-92.(Feng Lu,Leng Fuhai. Development of theoretical studies of co-word analysis[J]. Journal of Library Science in China,2006,32(2):88-92.)

[3] 唐果媛,张薇.国内外共词分析法研究的发展与分析[J].图书情报工作,2014,58(22):138-145.(Tang Guoyuan,Zhang Wei. Development and analysis of co-word analysis method at home and abroad[J]. Library and Information Science,2014,58(22):138-145.)

[4] 杨建林.关键词选择策略及其对共词分析的影响[J].情报学报,2014,33(10):1083-1090.(Yang Jianlin. Strategies of keyword selection and their impact on co-word analysis[J]. Journal of the China Society for Scientific and Technical Information,2014,33(10):1083-1090.)

[5] 傅柱,王曰芬.共词分析中术语收集阶段的若干问题研究[J].情报学报,2016,35(7):704-713.(Fu Zhu,Wang Yuefen. A discussion on some questions of term collection in co-word analysis[J]. Journal of the China Society for Scientific and Technical Information,2016,35(7):704-713.)

[6] Albrechtsen H,Hjorland B. Understandinmgs of language and cognition:implications for classfication research[J]. Advance in Classification Research Online,1994,5(1):1-16.

[7] Hu C P,Hu J M,Deng S L,et al. A co-word analysis of Library and Information Science in China[J]. Scientometrics,2013,97(2):369-382.

[8] 唐果媛,张薇.基于共词分析法的学科主题演化研究进展与分析[J].图书情报工作,2015,59(5):128-136.(Tang Guoyuan,Zhang Wei. Development and analysis of subject theme evolution based on co-word analysis method[J]. Library and Information Science,2015,59(5):128-136.)

[9] Zhang B,Mclean D C,Lu X. Identifying biological concepts from a protein-related corpus with a probabilistic topic model[J]. Bmc Bioinformatics,2006,7(1):58-68.

[10] 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.

[11] 钟伟金.共词分析法应用的规范化研究——主题词和关键词的聚类效果对比分析[J].图书情报工作,2011,55(6):114-118.(Zhong Weijin. Empirical study on effectiveness of the co-word analysis:comparative analysison the clustering results of subject heedings and keywords[J]. Library and Information Science,2011,55(6):114-118.)

[12] 王贤文,徐申萌,彭恋,等.基于专利共类分析的技术网络结构研究:1971—2010[J].情报学报,2013,32(8):89-92.(Wang Xianwen,Xu Shenmeng,Peng Lian,et al. Structural analysis of technology network based on patent co-classification analysis:1971-2010[J]. Journal of the China Society for Scientific and Technical Information,2013,32(8):89-92.)

[13] 王忠义,谭旭,夏立新.共词分析方法的细粒度化与语义化研究[J].情报学报,2014,33(9):969-978.

(Wang Zhongyi, Tan Xu, Xia Lixin. Research on fine-granularity and semantization of co-word analysis[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(9): 969-978.)

[14] Turber W A, Chartron G, Laville F, et al. Packaging information for peer review: new co-word analysis techniques [M]. Handbook of Quantitative Studies of Science and Technology. Netherlands: Elsevier Science Publishers, 1988: 291-323.

[15] 王昌度,熊云,徐金龙,等.科技期刊论文关键词标引的问题与对策[J].编辑学报,2003,15(5):349-351. (Wang Changdu, Xiong Yun, Xu Jinlong, et al. On keyword indexing to sci-tech papers[J]. Acta Editologica, 2003, 15 (5): 349-351.)

[16] Whittaker J. Creativity and conformity in science: titles, keywords and co-word analysis[J]. Social Studies of Science, 1989, 19(3): 473-496.

[17] King J. A review of bibliometric and other science indicators and their role in research evalutation[J]. Journal of Information Science, 1997, 23(4): 301-311.

[18] 任红娟.一种内容和引用特征融合的知识结构划分方法研究[J].中国图书馆学报,2013,39(5):76-82. (Ren Hongjuan. A division method of intellectual structure of combining content and citation of documents[J]. Journal of Library Science in China, 2013, 39(5): 76-82.)

[19] 唐晓波,肖璐.融合关键词增补与领域本体的共词分析方法研究[J].现代图书情报技术,2013,29 (11): 60-67.(Tang Xiaobo, Xiao Lu. Research of co-word analysis method of combining keywords extension and domain ontology[J]. New Technology of Library and Information Service, 2013, 29 (11): 60-67.)

[20] 李秀霞,邵作运.融入内容信息的作者共被引分析——以学科服务研究主题为例[J].图书情报工作,2016, 60(1): 98-104.(Li Xiuxia, Shao Zuoyun. The author co-citation analysis based on the content information: taking the subject service as an example[J]. Library and Information Science, 2016, 60(1): 98-104.)

[21] Leydesdroff L. Words and co-words as incicators of intellectual organization[J]. Research Policy, 1989, 18(4): 209-223.

[22] Glenisson P, Glanzel W, Persson O. Combining full-text analysis and bibliometric indicators. A pilot study[J]. Scientometrics, 2005, 63(1): 163-180.

[23] Kostoff R N, Eberhart H J, Toothman D R. Database tomography for information retrieval[J]. Journal of Information Science, 1997, 23(4): 301-311.

[24] 巴志超,李纲,朱世伟.共现分析中的关键词选择与语义度量方法研究[J].情报学报,2016,35(2):197-207.(Ba Zhichao, Li Gang, Zhu Shiwei. Research on keyword selection and semantic measurement of co-word analysis[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(2): 197-207.)

[25] 索红光,刘玉树,曹淑英.一种基于词汇链的关键词抽取方法[J].中文信息学报,2006,20(6):25-30. (Suo Hongguang, Liu Yushu, Cao Shuying. A keyword selection method based on lexical chains[J]. Journal of Chinese Information Processing, 2006, 20(6): 25-30.)

[26] Anjewierden A, Kabel S. Automatic indexing of PDF documents with ontologies[C]//Proceedings of 13th Belgian/Dutch Conference on Artificial Intelligence (BNAIC'01). Amsterdam, Neteherlands, 2001: 23-30.

[27] 潘书祥.汉语科技术语的规范和统一[J].中国科技术语,1998(1):6-11.(Pan Shuxiang. The standardization and unification of scientific and technical terms[J].China Terminology, 1998(1): 6-11.)

[28] Lee B, Jeong Y I. Mapping Korea's national R & D domain of robot technology by using the co-word analysis[J].

Scientometrics,2008,77(1):3-19.

[29] 邵作运,李秀霞.共词分析中作者关键词规范化研究——以图书馆个性化信息服务研究为例[J].情报科学,2012,30(5):731-735.(Shao Zuoyun,Li Xiuxia. Study on the standardization of author keywords in co-word analysis:taking library personalized information services study as example[J]. Information Science,2012,30 (5):731-735.)

[30] 刘鲁红.浅谈主题标引规范化[J].情报理论与实践,2004,27(4):367-368.(Liu Luhong. The standardization of subject indexing[J].Information Studies:Theory & Application,2004,27(4):367-368.

[31] 张晗,赵玉虹.医学文献语义共词知识网的构建:方法与实证[J].图书情报工作,2016,60(11):135-142.(Zhang Han,Zhao Yuhong. Construction of semantic co-word knowledge network for medical literature:method and empirical study[J]. Library and Information Science,2016,60(11):135-142.)

[32] Ding Y,Chowdhury G G,Foo S. Incorporating the results of co-word analysis to increase search variety for information retrieval[J].Journal of Information Science,2000,26(6):429-451.

[33] 张晗,崔雷.生物信息学的共词分析研究[J].情报学报,2004,22(5):613-617.(Zhang Han,Cui Lei. Study of bioinformatics through co-word analysis[J]. Journal of the China Society for Scientific and Technical Information,2004,22(5):613-617.)

[34] 许海云,岳增慧,雷炳旭,等.基于专利技术功效主题词与专利引文共现的核心专利挖掘[J].图书情报工作,2014,58(4):59-64.(Xu Haiyun,Yue Zenghui,Lei Bingxu,et al. Core patents mining based on cross co-occurrence analysis to patent technology-effect subject terms and citations[J]. Library and Information Science,2014,58(4):59-64.)

[35] 李武,董伟.国内开放存取的研究热点:基于共词分析的文献计量研究[J].中国图书馆学报,2010,36(6):105-115.(Li Wu,Dong Wei. Hostpots of open access research in China:an empirical study based on co-word analysis[J]. Journal of Library Science in China,2010,36(6):105-115.)

[36] 陆宇杰,张凤仙,范并思.基于共词分析的高校图书馆核心价值研究[J].大学图书馆学报,2011,29(6):34-40.(Lu Yujie,Zhang Fengxian,Fan Bingsi. Research on the core value of foreign universities[J]. Journal of Academic Libraries,2011,29(6):34-40.)

[37] Zong Q J,Shen H Z,Yuan Q J,et al. Doctoral dissertations of library and information science in China:a co-word analysis[J]. Scientometrics,2013,94(2):781-799.

[38] Dehdarirad T,Villarroya A,Barrios M. Researchtrends in gender differences in higher education and science:a co-word analysis[J]. Scientometrics,2014,101(1):273-290.

[39] 李湘东,巴志超,黄莉.一种基于加权 LDA 模型和多粒度的文本特征选择方法[J].现代图书情报技术,2015,31(5):42-49.(Li Xiangdong,Ba Zhichao,Huang Li. A text feature selection method based on weighted latent dirichlet allocation and multi-granularity[J]. New Technology of Library and Information Servrce,2015,31(5):42-49.

[40] 顾燕萍,侯汉清,王晓红.中文图书自动标引与分类加权设计研究[J].中国图书馆学报,2006,32(6):69-72.(Gu Yanping,Hou Hanqing,Wang Xiaohong. Weighting design of automatic indexing and classification of Chinese books[J]. Journal of Library Science in China,2006,32(6):69-72.)

[41] 李纲,李秩.一种基于关键词加权的共词分析方法[J].情报科学,2011,29(3):321-324.(Li Gang,Li Yi. A new method for weighted co-word analysis based on keywords[J]. Information Science,2011,29 (3):321-324.)

[42] 吴清强，赵亚娟.基于论文属性的加权共词模型探讨[J].情报学报，2008，27(2)：89-92.(Wu Qingqiang，Zhao Yajuan. Research in the weighted co-word analysis based on the attributes of articles[J]. Journal of the China Society for Scientific and Technical Information，2008，27(2)：89-92.)

[43] 钟伟金.基于主要主题词加权的共词聚类分析法效果研究[J].情报学报，2009，28(4)：214-219.(Zhong Weijin. Research into the effects of weighted co-word cluster analysis based on major descriptor[J]. Journal of the China Society for Scientific and Technical Information，2009，28(4)：214-219.)

[44] Donohue J C. Understanding scientific literatures：a bibliometric approach[M].Cambridge：the MIT Press，1973：49-50.

[45] Wang Z S，Zhao H，Wang Y. Social networks in marketing research 2001-2014：a co-word analysis[J]. Scientometrics，2015，105(1)：65-82.

[46] Sedighi M. Application of word co-occurrence analysis method in mapping of the scientific fields (case study：the field of informetrics)[J]. Library Review，2016，65(1/2)：52-64.

[47] Zhang W，Zhang Q，Yu B，et al. Knowledge map of creativity research based on keywords network andco-word analysis 1992-2011[J]. Quality & Quantity，2014，49(3)：1-16.

[48] Courtial J P，Law J. A co-word study of artificial intelligence[J].Social Studies of Science，1989，19(2)：301-311.

[49] Yang Y，Wu M，Cui L.Integration of three visualization methods based on co-word analysis[J]. Scientometrics，2012，90(2)：659-673.

[50] Yan B N，Lee T S，Lee T P. Analysis of research papers on E-commerce(2000-2013)：based on a text mining approach[J]. Scientometrics，2015，105(1)：403-417.

[51] Zhang S，Liu C X，Chang Y. Selection research of keywords in co-word clustered based on the G-index of word frequency[J]. Modern Educational Technology，2013，23(10)：54-57.

[52] 杨爱青，马秀峰，张风燕，等. g指数在共词分析主题词选取中的应用研究[J].情报杂志，2012，31(2)：52-55.(Yang Aiqing，Ma Xiufeng，Zhang Fengyan，et al. Application research of g-index in the topic words of co-word analysis[J]. Journal of Intelligence，2012，31(2)：52-55.)

[53] 胡昌平，陈果.科技论文关键词特征及其对共词分析的影响[J].情报学报，2014，33(1)：23-32.(Hu Changping，Chen Guo. Characteristics of keywords in scientific papers and their impact on co-word analysis[J]. Journal of the China Society for Scientific and Technical Information，2014，33(1)：23-32.)

[54] Serrano M A，Boguna M，Vespignani A. Extracting the multiscale backbone of complex weighted networks[J]. Proceedings of the national academy of sciences，2009，106(16)：6483-6488.

[55] Price L，Thelwall M. The clustering power of low frequency words in academic webs[J]. Journal of the Association for Information Science and Technology，2005，56(8)：883-888.

[56] 张玉芳，万斌候，熊忠阳.文本分类中的特征降维方法研究[J].计算机应用研究，2012，29(7)：2541-2543.(Zhang Yufang，Wan Binhou，Xiong Zhongyang. Research on feature dimension reduction in text classification[J]. Application Research of Computers，2012，29(7)：2541-2543.)

[57] 邱均平，丁敬达.1999—2008年我国图书馆学研究的实证分析[J].中国图书馆学报，2009，35(6)：79-87.(Qiu Junping，Ding Jingda. An empirical analysis of Chinese Library Science research in 1999-2008 [J].Journal of Library Science in China，2009，35(6)：79-87.)

[58] 赵辉,彭洁,刘润达,等.基于关键词网络的科技项目多角度演化分析[J].情报学报,2011,30(6):658-667.(Zhao Hui,Peng Jie,Liu Runda. Evolutionary analysis of science and technology projects based on keywords network[J]. Journal of the China Society for Scientific and Technical Information,2011,30(6):658-667.)

[59] 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 chemsitry[J]. Scientmetrics,1991.22(1):155-205.

[60] 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,2002,11(2):1-3.(Zheng Huachuan,Yu Xiaoou,Xin Yan. Antigen CD44 with clustered analysis of co-words:a status Quo investigation[J]. Chinese Journal of Medical Library and Information,Science,2002,11(2):1-3.)

[61] 郭红梅,张智雄.基于图挖掘的文本主题识别方法研究综述[J].中国图书馆学报,2015,41(6):97-108.(Guo Hongmei,Zhang Zhixiong. Methods of text theme identification based on Graph mining[J]. Journal of Library Science in China,2015,41(6):97-108.)

[62] Wang D X,Liu X,Luo H Z. A novel framework for semantic entity identification and relationship integration in large scale text data[J]. Future Generation Computer Systems,2015,64(11):198-210.

[63] Li S,Sun Y,Soergel D. Erratum to:a new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis[J]. Scientometrics,2016,103(2):1023-1042.

[64] Qi J,Ohsawa Y. Matrix plane model:a novel measure of word co-occurrence and application on semantic relatedness[C]//Proceedings of 2015 IEEE 15th International Conference on Data Mining Workshops (ICDM W). Atlantic,USA,2015:1246-1253.

[65] Aouicha M B,Taieb M,Hamadou A B.Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness[J]. Applied Intelligence,2016,45(2):1-37.

[66] Tsatsaronis G,Varlamis I,Vazirgiannis M. Text relatedness based on a word thesaurus[J]. Journal of Artificial Intelligence Resarch,2010,37(1):1-40.

[67] Jiang Y,Bai W,Zhang X,et al. Wikipedia-based information content and semantic similarity computation[J]. Information Processing & Management,2017,53(1):248-265.

[68] Roberto N,Simone P P. Babel relate! A joint multilingual approach to computing semantic relatedness [C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto,Canada,2012:108-114.

[69] Jorge G, Eduardo M. Web-based measure of semantic relatedness [C]//Proceedings of the International Conference on Web Information Systems Engineering(WEBIST). Auckland,New Zealand,2008:136-150.

[70] Zhang K Y,Zhu K Q.An association network for computing semantic relatedness[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin Texas,USA,2015:593-600.

[71] Yih W T,Qazvinian V. Measuring word relatedness using heterogeneous vector space models[C]//Proceeding of the Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies(NAACL HLT). California,USA,2012:616-620.

[72] Wang H,Deng S H,Su X N.A study on construction and analysis of discipline knowledge structure of Chinese LIS based on CSSCI[J]. Scientometrics,2016,109(3):1725-1759.

[73] Qiu J N,Wang G Y,Zhang K. Research on semantic transitivity laws based on association analysis in objective knowledge system[C]//Proceedings of International Symposium on Knowledge and Systems Science (KSS). Sapporo,Japan,2014:17-22.

[74] Xiao L,Chen G,Sun J,et al. Exploring the topic hierarchy of digital library research in China using keyword networks:a K-core decomposition approach[J]. Scientometrics,2016,108(3):1085-1101.

[75] Robert N,Paola V. Learning domain ontologies from document warehouses and dedicated Web sites [J]. Computational Linguistics,2004,50(2):151-179.

[76] Chen G,Xiao L. Selecting publication keywords for domian analysis in bibliometrics:a comparison of three methods [J]. Journal of Informetrics,2016,10(1):212-223.

[77] 李佳.共词聚类分析法中的主要问题与对策[J].情报学报,2010,29(4):614-617.(Li Jia.Main problems in co-word clustered analysis and their solution[J]. Journal of the China Society for Scientific and Technical Information,2010,29(4):614-617.)

[78] 钟伟金,李佳.共词分析法研究(二)——类团分析[J].情报杂志,2008,27(6):141-143.(Zhong Weijin,Li Jia. The research of co-word analysis—cluster analysis[J]. Journal of Information,2008,27(6):141-143.)

[79] 李牧南.基于关联规则挖掘竞争情报研究前沿分析[J].情报杂志,2016,35(3):54-60.(Li Munan. Probing the research frontiers of competitive intelligence based on the associations rules mining[J]. Journal of Intelligence,2016,35(3):54-60.)

[80] 崔雷,李丹,冯博.运用主题词/副主题词关联规则在医学文献检索系统中抽取知识的尝试[J].情报学报,2005,24(6):657-662.(Cui Lei,Li Dan,Feng Bo. An exploratory research of knowledge acquistition with MeSH heading/subheading association rules in medline system[J]. Journal of the China Society for Scientific and Technical Information,2005,24(6):657-662.)

[81] 高继平,丁堃,潘云涛,等.多词共现分析方法的实现及其在研究热点识别中的应用[J].图书情报工作,2014,58(24):80-85.(Gao Jiping,Ding Kun,Pan Yuntao,et al. Implementation of multiple words co-occurrence analysis and its application in the recognition of research hotspots[J]. Library and Information service,2014,58(24):80-85.)

[82] Zhang A S,Shi W Z,Webb G. Mining aignificant association rules from uncertain data[J]. Data Mining and Knowledge Discovery,2016,30(4):928-963.

李 纲 武汉大学信息资源研究中心副主任,长江学者特聘教授,博士生导师。湖北 武汉 430072。
巴志超 武汉大学信息管理学院博士研究生。湖北 武汉 430072。

(收稿日期:2016-11-28)