

引文分析的新阶段:从引文著录分析到引用内容分析

New Stage of Citation Analysis: from Citation Description Analysis to Citation Context Analysis

刘盛博 丁 堃 张春博

(大连理工大学科学与科学技术管理研究所暨 WISELab, 大连, 116023)

[摘要] 引用内容分析是基于引文分析基础理论,借助文本挖掘和自然语言处理等技术,从施引文献的全文入手,聚焦于引用的片段,对引用频次、引用位置和引用文本的内容主题进行的挖掘和研究。文章从引用内容概念出发,探讨了引用内容分析与引文分析的一般关系。引用内容分析是引文著录分析的精致化,是引文分析理论发展的新阶段。接下来,从引文描述性统计和引文网络视角,阐述了引用内容对传统引文著录分析的比较优势。引用内容文本分析则是引用内容分析的独有优势。最后,简单总结了引用内容分析的基本框架,并指出其未来存在的四个研究取向。

[关键词] 引用内容分析 引文分析 引文著录分析 引用频次分析 引用位置分析 引用内容文本分析

[中图分类号] G350 **[文献标识码]** A **[文章编号]** 1003-2797(2015)03-0025-10 **DOI:** 10.13366/j.dik.2015.03.025

[Abstract] Based on the basic theory of citation analysis, citation context analysis takes the mining and research on the citation frequency, citation location and the text topic of citation content from citing articles in full-text to referenced fragments by the method of text mining and natural language processing. Starting from the concept of citation context, this paper discusses the general relationship between citation context analysis and citation analysis. Citation context analysis is more detailed than citation description analysis, and then has been the new stage of the theory of citation analysis. Then, the paper expounds its comparative advantage prior to citation description analysis from the perspectives of descriptive statistics of citation and citation network. Topic analysis based on the citation text is the unique advantage of citation context analysis. In the end, we sum up the basic framework of citation context analysis, and point out the four research orientations in the future.

[Key words] Citation content analysis Citation analysis Citation description analysis Citation frequency analysis Citation location analysis Citation content text analysis

1 引言

自引文分析理论创生的 50 多年来,其研究主要围绕两方面来展开。一方面是以引文著录信息为分析载体的传统引文分析,另一方面则是深入施引文献正文内容进而探查引用功能和引用动机的引用内容分析。与引文著录分析相比,引用内容分析研究明显

较少,且多集中在施引文献的主题内容研究上,较少深入到文献全文内容层面。然而参考文献在文章中的引用内容能够为我们提供更多的引用相关信息,对了解参考文献对于施引文献的作用和价值,挖掘论文作者引用该文献的意图与观点倾向性具有更直接的作用。

[基金项目] 本文系为 ISTIC-THOMSON 科学计量学联合实验室开放基金“基于全文信息的科技论文评价研究”和高等学校学科点专项科研基金“基于 SIPOD 的专利知识测度体系及其应用研究”(20110041110034)的研究成果之一。

[作者简介] 刘盛博,男,博士后,研究方向:知识计量,Email: liushengbo1121@gmail.com; 丁堃,女,教授,博士生导师,大连理工大学公共管理与法学学院副院长,研究方向:学科知识测度、创新管理; 张春博,男,博士研究生,研究方向:科学计量、创新管理。

引文分析的新阶段:从引文著录分析到引用内容分析

New Stage of Citation Analysis: from Citation Description Analysis to Citation Context Analysis

刘盛博 丁 堃 张春博

随着电子数据库建设的逐渐完善和信息处理技术的不断发展,更多可解析的全文数据库得到开发,例如 PubMed/BioMed Central、Citeseer 和 arXiv 等数据库都可以提供可进行数据格式解析和文本内容挖掘的全文信息。这些可解析的全文数据库为深入论文内部,进行全文层面的引文分析研究提供了良好的数据基础。而全文信息中包含了论文著录无法提供的引用内容相关信息,如引文发生的位置、共被引发生的距离、共被引发生的位置、引文内容涉及的主题等信息,可以进一步开展引用动机、引用类型、引用功能和引用内容主题等方面的理论研究以及包括检索、评价和知识的演化、发现和预测在内的应用研究。可以说,基于施引文献全文的引用内容分析,既可以深入拓展引文分析理论创生时的一些基本理论命题,又可以开辟引文分析新的研究和应用域,是引文分析理论发展的新阶段。本文将从基本理论与研究方法角度,探讨引用内容分析与传统基于著录的引文分析的联系与区别,并深入揭示出引用内容分析对于引文著录分析的比较优势和补充作用。

2 引用内容的概念

2.1 引用内容概念的提出和发展

引用内容概念是伴随着引文分析的产生而提出的。19 世纪 60 年代到 70 年代,引文分析在《科学引文索引》数据库建立之后逐渐兴起,Chubin^[1]、Oppenheim^[2]、Spiegel-Rösing^[3] 等人在做引文年份、引文类型、引文频次等分析的同时,也将引用内容作为分析对象。他们在对引用内容进行研究时,将其描述为“content of reference”或“content of citation”,并没有明确界定引用内容的范围,在具体研究过程中,主要采用主观判别的方式来获取引用内容。

对引用内容最具影响力的定义是由 Small^[4] 在 1982 年提出的,他将引用内容表述为“citation context”,将其定义为“The text surrounding the references”,即参考文献及其标识周围的文本内容。例如,句子“This comparison is made using BLASTX [18]……”就可以视为参考文献[18]的“citation context”。O'Connor^[5]、McCain^[6] 也使用了此定义,分别研究了引用内容在信息检索中的应用和引文分类中的作用。

随着全文数据库的发展,人们从不同角度研究引用内容时,在 Small 的引用内容定义基础上,从语句的数量角度限定了引用内容的文本范围,并给出定义。Nanba 和 Okumura^[7,8] 将引用内容定义为“reference areas”,指的是参考文献区域内,与引文相关的一个或多个句子。Mei^[9] 等人在利用引用内容生成文本概要的研究中,将引用内容定义为引用标签周围的五句话,其中有一句是包含引用标签的句子,另外四句分别是含有引用标签句子前面的两个句子和后面的两个句子。Teufel 等^[10] 将引用内容视为“text windows”,即“文本窗口”通过设置文本窗口的大小,来控制引用内容所包含的句子数量,并从引用内容中抽取索引词来提高文献的检索效率。Nakov^[11] 在 2008 年提出“cintance”的概念,指的是引用句子集合,用于表示引用内容。Kaplan 等人^[12] 用“citation-site”或“c-site”来表示引用内容,每一个“citation-site”可以包括多个句子,而每个句子称为“c-site sentence”,同时他们采用“anchor”来表示引用标签,每一个包含引用标签的句子称为“anchor sentence”。

2.2 引用内容概念的总结

综合前人学者定义,从突出施引文献和参考文献的内容关联角度出发,本文认为引用内容就是指能够表征施引文献引用参考文献的文本内容,这些内容通常用一个或几个句子来表达。一般情况下,文本内容既包含了量的信息,也有质的信息。量的信息包含引用的多与少,具体反映在文本中就是指引用内容中包含的句子数量、引用句子中主题词的多少以及参考文献在施引文献中被引用的次数。量的信息可以表征参考文献对施引文献的影响力。质的信息包括以下三类:一是引用内容发生的位置,引用发生在施引文献中的不同位置时,所体现的作用并不相同;二是引用动机,引用内容中的文本语义可以反映出施引作者在引用参考文献时的引用意图;三是引用主题,此主题揭示的是施引文献与参考文献的直接联系。

总体来说,一条引用内容应包含以下五个要素:

- ① 文本内容 T: 表征施引文献引用某条参考文献时的上下文内容;
- ② 参考文献 R: 引用内容所对应的参考文献,可以

采用参考文献编号来识别;

③施引文献 D: 此条引用内容所在的文档, 可以采用文档编号来识别;

④引用句子编号 P: 包含引用标签的句子在施引文献中的句子编号, 用于表示引用内容所发生的位置;

⑤句子长度 L: 引用内容包含的句子数量。

通过对这些要素本身以及要素间关联的分析, 进而形成引用内容分析的基本研究内容。

3 引文分析与引用内容分析的一般关系

3.1 引文分析的概念和类型

引文分析就是利用各种数学及统计学的方法和比较、归纳、抽象、概括等逻辑方法, 对科技期刊、论文、著作等各种分析对象的引用或被引用现象进行分析研究, 以便揭示其数量特征和内在规律, 达到评价、预测科学发展趋势的目的^[13]。随着科学引证行为的不断规范和《科学引文索引》工具的广泛使用, 更多文献的获取、参考和引证已成为一种基本的科研现象, 进而形成大规模科学引文网络。这种引证文化的形成和技术条件的便利性, 也直接推动了引文分析理论的不成熟, 在一定意义上已成为现代科学计量学和文献计量学的研究根基^[14]。

引文分析可以打破传统的学科分类界线, 从多维角度来反映学科间的相互交叉、相互渗透关系; 并借助引文关联和网络化的图谱形式, 形象地描绘出科学发展过程中的学术流派、演进路径和热点前沿。此外, 基于引文数据统计的科学管理和评价, 也是引文分析重要的理论研究和应用实践领域。

如前文所述, 从引文分析的对象和内容来看, 引文分析理论基本形成以下两种研究类型:

(1) 基于著录信息的引文分析。主要基于参考文献的著录信息展开的研究, 包括以引用频次为代表的引文描述性统计分析和引文网络分析, 一方面应用在论文、期刊和研究主体的评价, 另一方面也常用于研究文献情报的规律, 揭示科学发展特征。

(2) 基于全文信息的引文分析。主要基于引用内容展开的研究, 包括了引用视角下的引用频次分析、引用位置分析和引用内容文本分析^[15], 并借助传统引文分析等理论方法, 深入揭示引用动机、引用功能和

引用行为规律。

文献著录信息包括两方面, 一方面是施引文献著录信息, 另一方面是参考文献著录信息。施引文献著录信息包括了文献的作者、机构、标题、关键词、摘要、期刊、年份等信息; 参考文献的著录信息包括了参考文献的作者、标题、期刊、文献发表年份、期、卷等信息。早期由于施引文献全文的解析和引用内容的获取较为困难, 即使有也均是对较小规模施引文献的人工研读和手动记录, 因而人们对引文分析的研究多倾向于对参考文献中的著录信息进行分析。而随着研究数据和信息处理技术等条件的成熟, 引文分析理论和实践也进入引用内容分析研究的新阶段。

3.2 引用内容分析的涵义

顾名思义, 引用内容分析与传统引文分析相比, 其核心的新发展是关注对内容的分析, 具体来说是从施引文献的全文入手, 并聚焦于引用的片段 (即引用句及其附近的句子)。

内容分析法最早产生于新闻传播学领域。20 世纪初, 有人采用一些半定量的统计方法对文献的内容进行深入的分析。尤其是美国学者贝雷尔森 (Bernard Berelson) 在 20 世纪 50 年代发表的权威著作《内容分析: 传播研究的一种工具》, 确立了内容分析法在大众传播学中的地位^[16]。近年来, 随着计算机技术和网络技术的发展, 内容分析法得到大幅度改进, 并被应用于很多领域的研究中, 包括心理学、人类学、教育学、语言学、历史学、图书馆学和信息科学等学科。

传统的内容分析法把媒介上的文字、非量化的有交流价值的信息转化为定量的数据, 通过建立有意义的类目系统分解交流内容, 并以此来分析信息的某些特征, 测验文献中本质性的事实和趋势, 揭示文献所含有的隐性情报内容, 对事物发展作情报预测^[17]。而笔者认为引用内容分析指的是对科学引证过程中, 具有明确引用标识的知识传播内容 (即引用内容) 的位置分布和内容主题进行的客观、系统、定量的分析。

引用内容分析属于内容分析研究范畴, 但其分析的文本内容是引用内容, 既具有一般的文本属性, 又具有引用行为及过程所生成的独特性质。因此, 在对引用内容分析时, 在传统的内容分析法基础上, 还需

引文分析的新阶段:从引文著录分析到引用内容分析

New Stage of Citation Analysis: from Citation Description Analysis to Citation Context Analysis

刘盛博 丁 堃 张春博

要结合科学计量学尤其是引文著录分析的研究方法。

引用行为的发生和进行,事实上也伴随着知识的继承与传播,而引用片段的内容正是记录了这一过程的文本载体。传统的文献计量分析虽然可以从引用频次统计上推断引用行为特点,但无法深入引用内容层面,从内容角度揭示知识传播的特点。而内容分析法是一种以研究传播内容为主的定量与定性相结合的分析方法,可以应用于研究任何文献或信息记录的交流传播事件,因而采用内容分析法对引用内容分析具有一定的适用性。

同时值得注意的是,引用内容分析也不仅聚焦于引用片段的内容主题的挖掘和分析,如前所述,也着眼于施引文献全文视角下的观察和研究。比如某一参考文献在施引文献的实际被引频次分析。引用位置分析同样也在施引文献的全文层次上展开的,无

论是对单篇参考文献的引用位置分布研究(如某篇参考文献是出现在相应施引文献的引言处或是结论处)还是多篇参考文献间的共被引层次和距离分析(即下文提到的同一句子层次的共被引还是在同一章节内的共被引等)。

图1展示了某论文的基于全文内容的引用内容分析的示例。由于篇幅的关系,没有全面展示论文间的前向引用,即文献耦合关系。施引文献题录和参考文献题录间的直接关系,以及各参考文献题录间的直接关系,共同构成了基于题录信息的传统引文分析。而引用内容分析则深入施引文献正文层面,努力揭开文献间引用及其衍生出的文献间共被引与耦合的“黑箱”。因而引用内容分析是引文著录分析的精致化,是引文分析理论发展的新阶段。

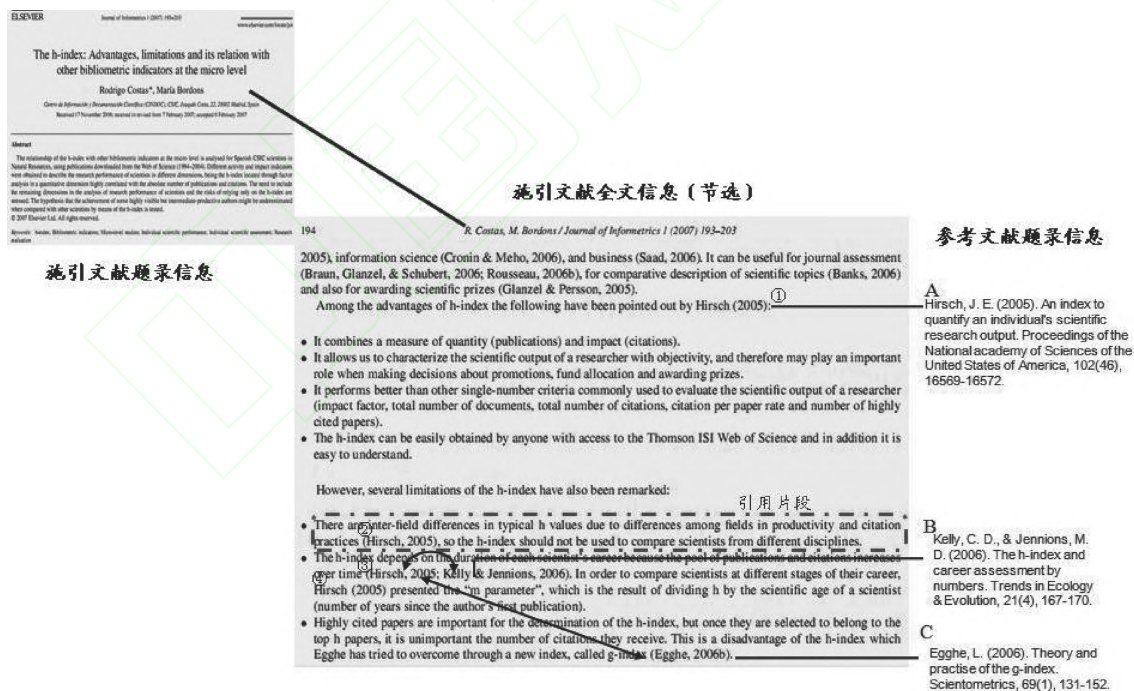


图1 引用内容分析示例

4 引用内容分析对引文著录分析的比较优势

4.1 引文描述性统计分析视角

引文描述性统计分析所要研究的内容是文献引

文的信息要素的数量特征及规律性,具体来说就是对大量的引文数据进行统计分析并总结规律。

(1)基于引文著录的引文描述性统计分析。引文描述性统计分析是相对简单的引文分析研究,也是开

展更深层次引文分析的基础。具体的统计内容包括引文年代分布、引文数量分布、引文类型分布、引文语种分布、引文国别分布等。一般来说,随着时间的由远及近,引文量呈增长趋势,即时间愈近,被引用的文献愈多。通过引文年代分析,不仅可以了解被引文献的出版、传播和利用情况,还可以研究科学发展的进程和规律。特别是在文献老化和科技史的研究中,引文年代分析更是一种广泛应用的有效方法。

论文的被引频次是引文描述性统计分析中涉及最多的研究内容,它所反映的是该论文在其领域内的影响力,可以直接用于评价论文质量,同时也可以以此为基础,对期刊和学者进行评价。被引用次数最多的论文,被认为是具有学术价值及地位的论文,而后被引用的机率也就随之增高。此种论文评价方法应用于学科领域的核心期刊遴选,最为典型的成果就是影响因子。引文频次分析也被广泛用于评估科研人员的学术水平,除了最基本的被引和他引分析,也结合科研人员的论文数量及被引用数量,计算其学术排名,最具代表性的成果就是 H 指数及类 H 指数。

然而传统的引文著录分析往往将文献与文献之间的引用关系简化为平等的线性关系^[18],随之而来是被引频次也简化为被引文献和施引文献的一对一的计数,缺乏与实际引用行为和引用性质的结合,进而造成分析结果的不准确。这种缺失和不准确主要表现在三方面:①没有考虑文本中实际存在的引文记录和引用位置的多对多关系;②没有考虑文章不同结构部分的引用,重要性有较大不同;③没有考虑实际的引用有着引用动机和引用情感上的显著差别,将正面引用、中性引用和负面引用混为一谈有失偏颇。

(2) 基于引用内容的引文描述性统计分析。在对引用内容的描述统计分析研究中,研究对象主要来源于全文信息而并不仅是参考文献的著录信息,分析的主要内容包括了引用强度(即参考文献在施引文献中的实际被引频次)、引用内容发生的位置、引用性质(或者说是引用态度的倾向)等。

引用强度的统计可以识别出一篇引文在不同文献中所体现出的不同作用。传统的基于著录信息的引文统计分析中,将所有引文对施引文献的作用视为

等同。而论文实际情况却多是不同被引文献在同一施引文献中被引频次并不相同(如图 1 所示该文对 Jorge Hirsch 2005 年提出 H 指数一文的引用)。一篇引文在一篇施引文献中出现的次数越多,说明它对这篇施引文献的作用越大。因而在对引文进行评价时,引用强度要比传统被引频次更精确可信,例如陈晓丽^[19]提出采用引文力度和引文深度角度对引文进行评价,其中的引用深度指标就是指的一篇引文在同一文献中被反复引用的次数。胡志刚^[20]比较了引用个数与引文篇数的相关性和区别,并基于引用个数预测和挖掘新的高被引论文。Hou Wen-Ru^[21]等人也通过引文在文献中具体被引次数(即引用强度),改进了传统影响因子的计算方法,并取得较好效果。

对引用内容发生位置的统计分析,可以揭示出施引者的引用行为规律,同时对不同位置上的引用内容分析,可以揭示出引文在文献不同位置出现时,所能体现出的地位与作用。Voos 和 Dagaev^[22]早在 1976 年就对引用内容的位置分布进行了研究,他们发现引用内容分布在施引文献不同位置时,引文的价值并不相同。何荣丽和魏洪善^[23]对理论型论文、实验型论文、综述型论文三种类型论文中引用内容位置分布差别进行了统计分析,结果发现,不同类型的论文中,引用内容在论文的引言、本论、结论中分布的数量并不相同。Sombatsompop^[24]等人采用引用内容分析的方式,判断引用内容发生在文章不同章节中的重要性。研究中将引用位置划分为 4 类,分别是引言、实验、结果讨论、结论。进一步提出引用位置影响因子指标,即引文在施引文献中不同位置上出现的次数与施引文献总数的比值,将这一指标应用于对论文的质量评价。Ding Ying^[25]等人将引用位置划分为摘要、引言、文献综述、方法、结果、结论六类,分别讨论了高被引文献在不同引用位置中的分布情况。其中高被引文献的统计方法分为两种:一种是根据被引文献在参考文献中出现次数来统计的,称为“CountOne”,另一种是根据被引文献在施引文献正文中出现次数来统计的,称为“CountX”。研究发现,采用“CountOne”方法统计的高被引文献最多分布在引言章节,而“CountX”方法统计的高被引文献最多分布在方法章节,说明一方面不同方法统计文献被引频次

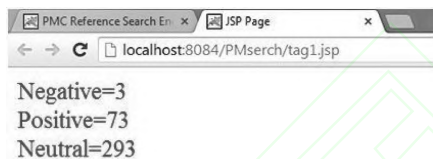
引文分析的新阶段:从引文著录分析到引用内容分析

New Stage of Citation Analysis: from Citation Description Analysis to Citation Context Analysis

刘盛博 丁 堃 张春博

所得到的分析结果可能存在差异性;另一方面不同位置的内容引用也存在差异性。

引用行为不仅是对相关工作的认可,事实上也受很多非科学因素的影响,呈现出高度复杂的特征^[26]。由于施引者的引用动机不同,每次引用的态度倾向也不尽相同。有些引用对论文内容是进行褒奖和继承,而有些引用则是指出论文研究的不足。因而对论文的频次统计尤其做评价时,不能一概而论,应区别对待。引用内容分析可以借助以自然语言处理为基本处理手段的文本倾向性分析,使人们更准确地判断被引文献是“好评如潮”,还是“千夫所指”。图2是笔者基于 Pubmed 数据库所开发的引用内容检索和评价平台,对“Thomson JA, 1998, Science, V282, P1145”一文的被引评价结果。

图2 引用内容态度倾向评价分析示例^[27]

4.2 引文网络分析视角

引文网络分析中,常用的三种引文网络类型分别是直接引用网络、文献耦合网络和文献共被引网络,三种类型网络如图3所示^[28]。

(1)引用内容分析与传统引文网络分析。直接引用网络是对科学文献间有向的引用关系进行的研究,可以揭示学科领域发展脉络、预测学科发展热点,揭示科学发展过程。文献耦合是指如果两篇文献具有一篇相同的参考文献,它们的耦合强度为1,若两篇文献有n篇相同参考文献,那么它们的耦合强度为n,耦合强度越高,两篇文献在学科内容专业性上越接近,文献间联系越紧密。文献共被引,就是两篇(或n篇)文献同时被后来的一篇或m篇文献所引证,则这两篇(或n篇)被引证论文则存在共被引关系。两篇(或n篇)文献共被引次数越多,它们之间的关系越强。文献耦合和文献共被引分别是知识的扩散和汇聚的路径为起点,来研究知识的关联性,进而扩展

为学科知识单元的聚类强度和网络结构。

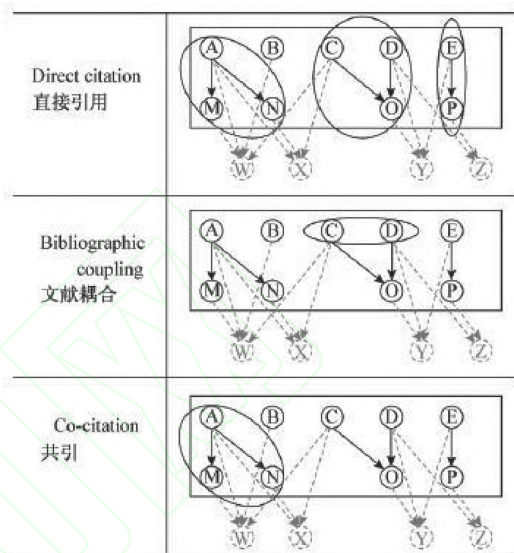


图3 三种引文网络类型

然而如前所述,传统的引文网络分析是将每篇文献看作一个节点(图3),最后只能形成和研究节点与节点之间的关系,既没有考虑节点本身的属性,节点间的关系也被均等化了。引用内容分析,则可以深入文献内容的主题和结构层面,将节点属性和节点间的关系赋予新的理解,进而改造传统引文网络分析理论。

在直接引用网络中,文献与文献间的引用关系所揭示的是文献间的继承关系,而在具体引用过程中,施引文献只是继承了参考文献中的某个知识点,并不是继承所有的知识点。引用内容分析可以直接揭示出引用过程中参考文献的哪些知识点被施引文献所引用,因此,在直接引用网络中研究引用内容,可以更深入、具体地揭示引用过程中知识的继承与演化。

在共被引网络中,共被引关系所表示的是共被引文献之间的关联性,通过分析网络的结构特征,揭示学科发展特征。计算共被引关系时,是对参考文献的著录信息进行统计,只要两篇文献同时出现在另一篇文献的参考文献著录信息中,这两篇文献间就有一次共被引关系^[29],并未考虑这两篇文献在施引文献中的位置及它们的引用内容相关性。引用内容分析则从全文结构入手分析施引文献间的位置和距离(如图1

中显示的两对共被引关系间的对比),着眼于引用片段并运用文本挖掘和自然语言处理技术分析其主题的相似性。

在文献耦合研究中,文献耦合关系表示的是施引文献间的关联性,通过对耦合网络的结构特征研究,揭示学科结构特征。具有耦合关系的两篇文献间的耦合强度是通过它们共同引用的参考文献数量来计算的^[30],而并未考虑它们共同引用的参考文献在施引文献中的位置及引用内容相似性,如果一篇参考文献在两篇施引文献中的引用位置和内容都相似,那么这两篇施引文献间的关系可能更紧密。

在引用内容分析框架下,直接引用网络主要涉及引用片段内容的主题挖掘、识别和比较,将在后文的引用文本内容主题分析来阐述。而对于文献耦合方面,目前尚未查阅到相关研究,须待进一步实验证明。因而接下来笔者将着重叙述引用内容分析在共被引关系研究上的发展。

(2)引用内容分析框架下的共被引网络研究。基于引用内容的共被引关系分析,同样有两个相辅相成的研究路径,分别是共被引片段内容主题的相关性和共被引文献的位置层次。前者同样是内容主题分析层面,暂不赘言;后者是本部分的叙述内容。

不同位置层次的共被引,共同被引用的文献间的关系程度是不同的。考虑共被引发生的位置层次及距离,可以更精确地揭示共被引关系,优化共被引网络聚类。Elkiss^[31]研究发现,共被引距离越近的两篇共被引文章,主题越相似。句子层次上的共被引关系要比章节层次上的共被引关系更紧密。

基于共被引位置的层次距离和共被引强度间正相关的思想,学者们也通过赋值,进一步量化共被引关系,深化了共被引理论,并被用于提高文献检索效率的研究。Gipp 和 Beel^[32]根据共被引发生位置,将共被引关系划分相同句子层次共被引、相同段落层次共被引、相同章节层次共被引、相同期刊层次共被引和相同期刊不同版本层次共被引 5 个层次。他们对发生在这 5 个层次上的共被引关系进行赋值,并采用相关文献检索的方法来验证位置层次赋值对共被引分析的有效性,结果发现,加入共被引权重后的检索

效率要比传统未加入权重的检索效率提高 2 倍。Eto^[33]将共被引文章划分为 4 类,分别是同引用标签共被引、同句子共被引、同段落共被引和不同段落共被引。当共被引发生在不同位置时,给予不同的相似性权重,而加入共被引权重后,信息检索效果有明显提高。Boyack 等人^[34]在对共被引位置进行权重赋值时,通过计算共被引文献位置间的字符数来对他们的共被引关系进行赋值,字符数越少,它们的共被引关系权重越大,实验结果发现,加入共被引位置后的共被引聚类效果比传统共被引聚类效果提高 30%。

鉴于相关研究均是通过主观判断对共被引权重赋值,缺少合理依据、方法不够准确,笔者提出了一种基于引用内容相似度的共被引权重赋值方法,既将引用内容和引用位置结合起来,又提高了检索的效率^[35]。表 1 是以 3 本 BMC 期刊为研究对象,做的共被引位置层次及其相似度分析结果。

表 1 共被引位置层次及其相似度示例

	句子层次	段落层次	章节层次	文章层次
共被引次数	2131	1146	1150	2359
相关共被引次数	2131	884	733	1321
平均相似度	1	0.77	0.64	0.56

4.3 引用内容文本分析

如果说引用频次分析和引用位置分析是传统引文分析理论在描述性统计分析和引用网络分析的延伸,引用内容的文本主题挖掘和分析则是引用内容分析所独有的研究内容,也是其内容分析导向的集中体现。当然,在很多实际研究中,引用内容的文本主题分析也常与前述的其他的引用内容分析方法相结合。

引用内容的文本主题挖掘就是利用文本挖掘和自然语言处理技术,对施引文献的引用片段进行数据格式解析、主题及特征词抽取、索引链接建立以及语义分析和科学计量分析。引用内容的文本分析主要通过主题词和特征词的抽取和语义分析,来对参考文献对于施引文献的引用功能和引用对象类型以及施引文献对参考文献的引用情感倾向和引用的主题知识进行探索、识别和归类,并可以在学术评价、信息检索和科学知识演化及发现等领域得到应用。

前文已述,通过对引用内容的知识主题、对象类

引文分析的新阶段:从引文著录分析到引用内容分析

New Stage of Citation Analysis: from Citation Description Analysis to Citation Context Analysis

刘盛博 丁 堃 张春博

型尤其是情感倾向的挖掘分析,更能真实全面地审视受评对象的学术影响力。Anderson^[36]利用引用内容,分析了组织学习领域中 Walsh 和 Ungson 发表的一篇经典文章,分析结果包括这篇文章具体的哪些知识被人们所引用,哪些知识对以后研究具有重大影响以及哪些知识被人们所批判,这些分析结果都是对这篇经典文献的综合评价,而这些评价结果是无法单纯通过引用频次来揭示的。Chang^[37]也采用类似的方法分析了 Taylor 在 1968 年发表的一篇论文《Question-Negotiation and Information-Seeking in Libraries》的学术影响,包括了这篇论文的哪些概念影响力最大,影响的主题包括哪些以及这篇文章在被引文献中的作用。

文本内容的相似性,可以提高论文检索的精度和效率。主要研究手段就是将引用内容中的索引词抽取出来,与引文本身的索引词相结合,进而提高引文的检索效率。O'Connor^[38,39]假设引用内容可以提供一些关于被引文献的信息,并基于仿真实验研究了引用内容中的词在信息检索中的应用。他将引用内容中的词作为被引文献的索引词,结果显示这种方法可以提高检索效率。Bradshaw^[40]利用引用内容信息改进搜索引擎效果,提出了一种参考文献直接索引(Reference Directed Indexing)方法,将引用内容中的检索词与引文的被引频次相结合,进而提高引文的检索效率。Mercer 和 Di Marco^[41]也利用引用内容中的线索词对生物医学文献进行分类,然后将这些分类信息加入到检索索引中,进而提高检索效率。

引用内容是施引文献与被引文献的直接关联信息,一方面可以揭示出被引文献的哪些知识被他人利用,另一方面则可以揭示出施引文献的研究基础。在具体的应用研究中,可以通过“参考文献—引用片段内容—施引文献—引用片段内容—参考文献”,来揭示直接引用和共被引等关系结构中所蕴含的科学技术知识演化和发现。Small^[42]早在 1986 年就利用引用内容中的主题词来揭示共被引网络的聚类主题。首先,提取了高被引文献的所有引用内容;其次,抽取引用内容中的重要主题词;最后在共被引网络可视化时,采用引用内容主题词来代替被引文献节点,进而实现采用引用内容主题词来揭示共被引聚类主题

的研究。Nanba 和 Okumura^[43,44]将一篇引文的所有引用内容信息收集起来,通过分析引用内容中的主题词,总结出这些引用内容的概要,用这些概要信息来描述这篇引文的主要研究内容。Mei^[45]和 Mohamad^[46]发现,通过引用内容总结出来的概要信息与引文本身的摘要信息不同,说明引文在被继承过程中,可以体现出原文中未被重点指出的重要价值。Elkiss 等人^[47]研究了引用文摘的生成,利用一篇文献的引用句子集合生成引用文摘,用于描述被引文献主题。Chang^[48]比较了《Little Science, Big Science》一文在自然科学学科中和人文社会科学学科中被引用时的引用主题差异。研究结果发现,此文在自然科学学科中和人文科学学科中的引用主题基本相同,但引用动机存在一定差异。

5 结论与展望

从引用内容的概念出发,本文将引用内容定义为能够表征施引文献引用参考文献的文本语句及其内容主题,并从质量和数量两个角度,深入解读了引用内容的涵义。而引用内容分析就是从施引文献的全文入手,聚焦于引用的片段,对引用频次、引用位置和引用文本的内容主题进行的挖掘和研究。作为下一代的引文分析理论^[49],引用内容分析对传统的基于引文著录的信息分析,具有功能价值上的拓展和突破性的创新。表 2 列出了笔者总结的引用内容分析的基本框架。

表 2 引用内容分析基本框架

分析起点	理论内容模块	应用研究领域
施引文献全文	引用频次分析	学术评价
	引用位置分析	信息检索
引用片段	引用内容主题分析	科学知识演化和发现

引用内容分析是在引文分析基础理论的指导下,深入施引文献的全文及文本内容层面,挖掘文献间的关系,进而所作的相关研究。因而引用内容事实上形成了两个分析起点:一方面从施引文献的全文层面入手,分析参考文献的实际被引用频次,挖掘单篇参考文献的位置分布以及多篇参考文献间的相对位置、层次和距离;另一方面聚焦于引用片段,运用文本挖掘和语

句处理技术,进行引用内容主题层面的研究。总体而言,前者是对传统引文分析理论的研究深化和功能拓展;而后者正是因其内容分析的技术和特征,具备了引文著录分析所没有的揭开论文引用“黑箱”的优势。

通过前文可见,引用内容分析主要应用在学术评价、信息检索以及科学知识演化和发现三个领域。学术评价和信息检索又共同形成了学术推荐上的应用;科学知识演化和发展可以应用在文献摘要总结、科学知识的梳理及可视化、学科发展规划等方面的研究。当然,这些应用研究并非只是在一类理论研究下实现和深化的,而是综合研究的成果。如文中已提到的,完善的学术评价既要分析其实际的被引用频次,也要通过其引用位置分布,分析其起到的引用功能;同时结合内容倾向分析等,更全面地看待其被引频次。

就引用内容分析的理论发展,笔者认为接下来可以有四个层面的研究取向。第一,传统引文分析理论对引用内容分析的指导和渗透。引用内容分析是引文分析理论的新发展,必然具有内容相通性,可以迁移过来去耕耘这片新开辟的研究域。第二,各内容模块受各自研究方法和技术的发展,研究会不断深化和精致化。比如当前基于引用位置(层次和距离)的共被引分析存在赋值主观的问题,如何更客观地基于文本本身的赋值值得关注。文本挖掘技术尤其是主题模型的不断更新,也推动了内容抽取的准确性。第三,各内容模块间的交叉综合是未来引用内容分析的趋势。因为一个实际研究问题本身由多个要素组成,只有多维度研究,才能更深刻地认识和解决研究问题。如前述的学术评价研究问题。第四,其他学科尤其是社会科学的研究思想和方法的吸收和运用,会深化学术和实践问题,甚至开辟新的研究路径。例如在引用动机分析上,将基于引用内容分析的客观结果统计与以问卷调查为代表的主观心理统计结合起来,会更系统化地揭示论文的引证行为。

参考文献

- 1 Chubin DE, Moitra SD. Content analysis of references: Adjunct or alternative to citation counting? [J]. *Social Studies of Science*, 1975, 5(4): 423-441
- 2 Oppenheim C, Renn SP. Highly cited old papers and the reasons why they continue to be cited[J]. *Journal of the American Society for Information Science*, 1978, 29(5): 225-231
- 3 Spiegel-Rosing I. Science studies: Bibliometric and content analysis[J]. *Social Studies of Science*, 1977: 97-113
- 4 Small H. Citation context analysis[J]. *Progress in communication sciences*, 1982, (3): 287-310
- 5, 39 O'Connor J. Biomedical citing statements: computer recognition and use to aid full-text retrieval[J]. *Information processing & management*, 1983, 19(6): 361-368
- 6 McCain KW, Turner K. Citation context analysis and aging patterns of journal articles in molecular genetics[J]. *Scientometrics*, 1989, 17(1): 127-163
- 7, 43 Nanba H, Okumura M. Towards multi-paper summarization using reference information[C]. *International Joint Conference on Artificial Intelligence*. Stockholm: LAWRENCE ERLBAUM ASSOCIATES LTD, 1999: 926-931
- 8, 44 Nanba H, Okumura M. Automatic detection of survey articles [C]. *Research and Advanced Technology for Digital Libraries*. Vienna: Springer, 2005: 391-401
- 9, 45 Mei Q, Zhai C. Generating impact-based summaries for scientific literature[J]. *Proceedings of ACL-08*. Columbus: HLT, 2008: 816-824
- 10 Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function[C]. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney: Association for Computational Linguistics, 2006: 103-110
- 11 Nakov PI, Schwartz AS, Hearst M. Citances: Citation sentences for semantic analysis of bioscience text[C]. *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*. Sheffield: ACM, 2004: 81-88
- 12 Kaplan D, Iida R, Tokunaga T. Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach[C]. *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Singapore: Association for Computational Linguistics, 2009: 88-95
- 13 庞景安. 科学计量研究方法论[M]. 北京: 科学技术文献出版社, 2002
- 14 刘则渊, 陈悦, 侯海燕, 等. 科学知识图谱: 方法与应用[M]. 北京: 人民出版社, 2008
- 15 赵蓉英, 曾宪琴, 陈必坤. 全文本引文分析——引文分析的新发展[J]. *图书情报工作*, 2014, 58(9): 129-135
- 16 邱均平, 余以胜, 邹菲. 内容分析法的应用研究[J]. *情报杂志*, 2006, 24(8): 11-13
- 17 陈淑平. 内容分析法成为图书馆学情报学研究方法的相关问

引文分析的新阶段:从引文著录分析到引用内容分析

New Stage of Citation Analysis: from Citation Description Analysis to Citation Context Analysis

刘盛博 丁 堃 张春博

- 题探讨[J]. 情报资料工作, 2009, (1): 19-22
- 18 陆伟, 孟睿, 刘兴帮. 面向引用关系的引文内容标注框架研究[J]. 中国图书馆学报, 2014, 40(6): 93-109
 - 19 陈晓丽. 引文评价中的引文方式与力度因素[J]. 图书馆, 2000, 6: 43-45
 - 20 胡志刚, 陈超美, 刘则渊, 等. 从基于引文到基于引用——一种统计引文总被引次数的新方法[J]. 图书情报工作, 2013, 57(21): 5-10
 - 21 Hou WR, Li M, Niu DK. Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution[J]. Bioessays, 2011, 33(10): 724-727
 - 22 Voos H, Dagaev KS. Are All Citations Equal? Or, Did We Op. Cit. Your Idem? [J]. Journal of Academic Librarianship, 1976, 1(6): 19-21
 - 23 何荣利, 魏洪善. 引文在论文中的分布和被引用内容的调查与分析[J]. 图书情报工作, 2000, 44(2): 26-29
 - 24 Sombatsompop N, Kositchaiyong A, Markpin T, et al. Scientific evaluations of citation quality of international research articles in the SCI database: Thailand case study [J]. Scientometrics, 2006, 66(3): 521-535
 - 25 Ding Y, Liu X, Guo C, et al. The distribution of references across texts: Some implications for citation analysis[J]. Journal of Informetrics, 2013, 7(3): 583-592
 - 26 祝清松, 冷伏海. 引文类型识别研究进展[J]. 图书情报知识, 2013, (6): 70-76
 - 27 刘盛博, 丁堃, 张春博. 基于引用内容性质的引文评价研究[J]. 情报理论与实践, 2015(已录用, 未出版)
 - 28 Boyack KW, Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? [J]. Journal of the American society for information science and technology, 2010, 61(12): 2389-2404
 - 29 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 and technology, 1973, 24(4): 265-269
 - 30 Kessler MM. Bibliographic coupling between scientific papers [J]. American Documentation, 1963, 14(1): 10-25
 - 31, 47 Elkiss A, Shen S, Fader A, et al. Blind men and elephants: What do citation summaries tell us about a research article? [J]. Journal of the American society for information science and technology, 2008, 59(1): 51-62
 - 32 Gipp B, Beel, J. Identifying related documents for research paper recommender by CPA and COA[C]. Proceedings of international conference on education and information technology. Berkeley: International Association of Engineers, 2009: 636-639
 - 33 Eto M. Evaluations of context-based co-citation searching[J]. Scientometrics, 2013, 94(2): 651-673
 - 34 Boyack KW, Small H, Klavans R. Improving the Accuracy of Co-citation Clustering Using Full Text[J]. Journal of the American society for information science and technology, 2013, 64(9): 1759-1767
 - 35 刘盛博, 张春博, 丁堃, 等. 基于引用内容与位置的共被引分析改进研究[J]. 情报学报, 2013, 32(12): 1248-1256
 - 36 Anderson MH, Sun PY. What have scholars retrieved from Walsh and Ungson (1991)? A citation context study [J]. Management Learning, 2010, 41(2): 131-145
 - 37 Chang YW. The influence of Taylor's paper, Question-Negotiation and Information-Seeking in Libraries [J]. Information Processing & Management, 2013, 49(5): 983-994
 - 38 O'Connor J. Citing statements: Computer recognition and use to improve retrieval [J]. Information processing & management, 1982, 18(3): 125-131
 - 40 Bradshaw S. Reference directed indexing: Redeeming relevance for subject search in citation indexes. Proceedings of the 7th European conference on digital libraries [C]. Trondheim: Springer, 2003: 499-510
 - 41 Mercer RE, Marco CD. A design methodology for a biomedical literature indexing tool using the rhetoric of science [C]. BioLink workshop in conjunction with NAAACL/HLT. Boston: Association for Computational Linguistics, 2004
 - 42 Small H. The synthesis of specialty narratives from co-citation clusters [J]. Journal of the American Society for Information Science, 1986, 37(3): 97-110
 - 46 Mohammad S, Dorr B, Egan M, et al. Using citations to generate surveys of scientific paradigms [C]. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder: Association for Computational Linguistics, 2009: 584-592
 - 48 Chang YW. A comparison of citation contexts between natural sciences and social sciences and humanities [J]. Scientometrics, 2013, 96(2): 535-553
 - 49 Ding Y, Zhang G, Chambers T, et al. Content-based citation analysis: The next generation of citation analysis [J]. Journal of the Association for Information Science and Technology, 2014, 65(9): 1820-1833

(收稿日期: 2015-01-15)