引文内容分析方法研究综述

祝清松[1,2] 冷伏海[1]
([1]中国科学院国家科学图书馆 北京 100190; [2]中国科学院大学 北京 100049)

摘要 文章首先对引文内容分析方法的内涵及其三个主要步骤进行阐述，然后对相关研究进展进行综述，最后对其潜在应用进行展望。

关键词 引文内容分析 引文分析 被引用次数 内在特征

A Review of Research on Citation Content Analysis Method

Zhu Qingsong[1,2] Leng Fuhai[1]
([1]National Science Library, Chinese Academy of Sciences, Beijing, 100190;
[2]University of Chinese Academy of Sciences, Beijing, 100049)

Abstract Firstly, the paper introduces the connotation and three major steps of citation content analysis. Then the paper reviews the relevant research progress. At last, the paper gives a prospect for its potential applications.

Keywords citation content analysis, citation analysis, citation frequency, intrinsic characteristic

1 引言

科技论文是科技创新活动的主要产出形式，是科技发展过程中知识的累积，是反映学科领域基础研究和应用研究的创新成果。科技论文之间并不是孤立存在的，一篇科技论文的形成一般建立在多篇科技论文的基础上，并通过科技论文之间的相互引用来实现，符合科学本身的发展规律和研究活动规律，体现了创新过程中知识流动的继承性和变化性[1]。引文分析就是建立在科技论文的引用与被引用关系(主要包括直接引用、文献耦合和同被引等[2])基础上，用于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[3]。

引文分析经过数十年的发展，在理论研究和实践应用方面都取得了长足的进展，其广泛应用于科学知识评价、学术研究动态变化和科学发展模式的揭示，并作为一个研究重点被积极地运用于发现学科间的关联，进行学科发展趋势分析和科学前沿预测等[4]。“从普赖斯、加菲尔德到斯莫尔，已确立起日臻完备的引文分析理论与方法，构成科学计量学的基础与主流，在一定意义上也可以说在科学计量学中已形成一门成熟的分支学科——引文分析学，现代科学计量学和文献计量学都构筑在引文分析学的根基之上”[5]。引文分析具有重要的学科地位，已经成为面向科技创新的战略情报研究和服务工作中必不可少的重要方法，对科技创新和科技决策的支撑具有重要的理论意义和实践价值。

然而，引文分析方法也存在很多问题，比如引文数据不全面、引用信息不明确、引用动机不清晰等。引用行为是作者的主观行为，不同作者引用同一篇科技论文的动机各不相同，同时还存在伪引和漏引等虚假引用、正面引用和负面引用、深度引用和浅度引用等复杂的引用行为[6]。引用行为的研究起源于20世纪60、70年代，代表人物有Garfield[7]和Weinstock[8]等，之后涌现出大量关于引用行为和引用类型的理论和实证研究，但是大部分的研究仍然停留在理论层面和实证调查，没有应用到实际的引文分析中来。另外，引文分析极大地依赖于引文数据库，不论是国外ISI的JCR、Scopus的SJR，还是国内的中国科学引文数据库(CSCD)、中国科技论文与引文数据库(CSTPCD)、中文社会科学引文索引(CSSCI)等都将引文看做是同等重要，各种统计指标也均以被引用次数为基础。

引用行为的复杂性对目前以被引用次数为主的引文分析提出了挑战。引文分析将所有的引文同等看

待，施引文献和被引文献之间的关联性也通常不加以区分，而实际引用行为具有复杂的动机，仅仅通过被引用次数并不能看出作者的引用动机，另外被引文献对施引文献所起的作用并不完全一样，引用作用的不同意味着引文的价值以及对施引文献的贡献和重要性也不同。因而，单纯地利用被引用次数是目前引文分析方法的最大问题。

同时，随着文本挖掘技术的提升以及全文本获取的可行性，对文献全文的分析越来越多。Ding[9]提出基于内容的引文分析，认为基于内容的引文分析是下一代的引文分析方向，并将其分为两个层次：一是语法层面，引文分布在文献中的不同语法结构中（出现在文中不同章节位置）；二是语义层面，引文具有不同的语义贡献（如重要或不太重要的贡献、肯定或否定型贡献）。

基于以上的研究背景，本文认为目前的引文分析方法应当进一步深入引文内容进行分析，通过引文内容分析来解决单纯利用被引用次数带来的只重数量不重内容的问题，将数量和质量评价有机融合，有效地揭示文献被引用的原因，更充分地挖掘施引文献和被引文献创新之间的关联，进而更有效地展现学科领域的演化过程，为面向科技创新的战略情报研究和服务提供新方法和新思路。

2 引文内容分析内涵

科技论文中大部分的创新并不是完全创新，而是在以往创新基础上的再创新，既有继承性又有变化性。科技论文的继承性主要是通过引用和被引用来体现的，并以参考文献的形式出现在科技论文的正文后面，且在正文中以特定的形式进行标记，本文称为引用标记，如 3、[19，24]、(15- 17) 等，将引用标记所在的句子称为引用句。同时，引用句上下文的句子通常在内容上也与被引文献相关，本文将与引用句在内容上相关联的上下文句子称为引文上下文。引用句是施引文献和被引文献的直接关联，引文上下文蕴含更加丰富的语义信息，引文内容分析指对引用句或引文上下文的内容分析。引文内容分析假设被引文献对施引文献创新的重要程度并不相同，其更加注重引文之间在内容层面上的语义关联，深入引文语篇内容来分析施引文献与被引文献之间的语义关联，揭示被引文献对施引文献创新所起的作用和贡献程度。引文内容分析主要包括数据预处理、引文内容抽取和内容深度分析三个步骤，具体流程如图 1 所示。

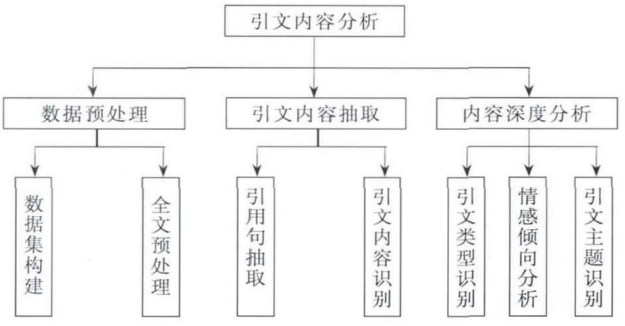

图 1 引文内容分析方法流程

(1) 数据预处理。数据预处理是引文内容分析的首要任务和重要保障，主要包括数据集构建和全文预处理。其中，数据集构建包括元数据下载、被引文献集构建和施引文献集构建，被引文献集由被引论文组成，施引文献集由引用被引文献集的论文组成。全文预处理主要针对施引文献集进行，包括全文下载、格式转换和数据清洗。下载全文可借助 EndNote 等工具进行批量处理。格式转换一般是将 PDF 格式的全文转换成易于处理的 TXT 格式。因为转换过程中会出现乱码等问题，因此需要对转换后的 TXT 数据进行适当的人工清洗。

(2) 引文内容抽取。引文内容抽取是引文内容分析的关键步骤，主要包括引用句抽取和引文内容识别。其中，引用句抽取是从施引文献集的全文文本中抽取引用标记所在的句子，首先要获得被引文献在施引文献全文中的位置，这需要利用参考文献序号和引用标记的对应来完成。另外，还可以利用语篇分析等获得引用句的相关上下文句子，利用引文上下文来进一步构建引文内容。引文内容识别是在引用句或引文上下文的基础上进行的，由于引用句中可能包含多个被引文献，因此需要对引用句进行处理，主要是根据引用标记来判定，获得所需被引文献对应的引文内容，一般包含以下四种情况：㈠如果引用句中只有一处含有参考文献序号且包含被引文献对应的参考文献序号，那么整个引用句作为引文内容；㈣如果引用句中有多处含有参考文献序号且有标点符号或并列、转折、比较等连接词（and、or、but、though、either …or、both …and、not only …but also、as、whereas、as well as、in contrast to、similar to 等）将引用句分隔成若干部分，那么保留包含被引文献对应的参考文献序号所在的部分；㈣如果引用句中有多处含有参考文献序号且作为句子的同一成分并列出现，通常以冒号或列举词（for instance、such as、including、by means of、because of、using、through、about、from、by、for、e. g. 等）开始，那么保留包含被引文献对应的参考文献序号所在的部分；¼ 如果引用句中有多处含有参考文献序号且分布包含多种情况，并将引用句分割成若干部分，那么以保留包含被引文献对应的参考文献序号所在的若干部分为首要原则。

(3) 内容深度分析。内容深度分析是引文内容分析的目标，是对引文内容的进一步挖掘和分析，其分析方法的广度和深度直接影响引文内容分析的应用价值，目前相关研究主要包括引文类型识别、情感倾

向分析、引文主题识别等。其中,引文类型识别是通过对施引文献引用被引文献的文本内容进行分析来对引文进行分类,是在引用动机研究的基础上,通过对被引文献在施引文献中的引文内容、引文位置和引文范围等的分析,对引文类型进行标引,进而揭示引文之间的语义关联。情感倾向分析是利用引文内容来挖掘施引文献作者对被引文献的各种观点、态度或立场,主要分为肯定、中立和否定三种基本类型。引文主题识别主要是指利用引文内容来识别代表被引文献的主题词,相对于关键词或从标题、摘要和全文抽取的主题词更加客观地反映被引文献的被引原因,更加直观地揭示被引文献的研究内容。

3 相关研究进展

引文内容分析的早期研究主要是人工对引文文本内容进行判读和总结,主要分析和发现各种引用现象和引用规律等[10]。Moravcsik 等[11]首次进行了广泛的引文文本分析,调查了理论高能物理领域 30 篇文献的引文,结果显示大部分引文是肯定引用,但是还有很大一部分引文是随意引用的,对引文数量作为评价指标提出质疑。Frost[12]对 74 篇文献的 828 篇引文进行内容分析,结果发现引用最多的为研究方法或结果,自引和证明引文的相对较少。Maricic 等[13]对多学科 357 篇文献的引文进行分析,包括引文位置(介绍、方法、结论和讨论)和引文水平,结果发现一般引文主要出现在介绍部分,重要引文主要出现在方法、结论和讨论部分。Hanneyet 等[14]通过引文文本内容分析对糖尿病和心脏病学研究的第一代文献对第二代文献的影响进行调查,结果发现被引用次数不能被用来代表被引文献的重要性,并且在该领域早期研究对后期的影响很小。

随着文本挖掘技术的提升以及全文本获取的可行性,对文献全文的分析越来越多,大量的全文本信息为情报研究提供了丰富的语料,如何从中挖掘出有价值的内容成为以内容为主情报研究的关键。引文分析也因此获得了新的研究思路,引文文本内容能够更加直接地反映引用的内容和关系等,利用引文文本内容来增强传统引文分析成为重要的研究方向,主要包括引文类型识别、情感倾向分析、引文主题识别等。

引文类型识别的任务是对引文进行分类,主要包括三个步骤:引文类型定义、引文内容抽取和引文类型标引。引文类型定义是对引文分类的标准,一般从引用功能和观点倾向两个角度出发;引文内容是识别对象,抽取的范围在狭义上仅指包含引文的句子,广义上还包括跟引文句子相关的其他上下文句子;引文类型标引是根据引文类型定义对抽取的引文内容进行引文类型标引。引文类型识别大部分的研究都从引用功能(被引文献对施引文献的作用,如 Pham[15]将引文分为基础、支持、局限、比较四类)和观点倾向(施引文献对被引文献的态度或立场,如 Shotton[16]将引文情感关系分为肯定、否定、中立三类)两种角度出发来对引文进行分类。引文类型识别为引文分析的语义增强提供了思路,但是还存在很多问题,如目前还没有形成统一的评价标准和测评语料,已有的研究基本都是对各自方法的评价或与以前工作的比较,结果验证的客观性和可靠性得不到保证。

情感倾向分析从一定程度上来说属于引文类型识别的一种,是对引文感情的分类,但两者又有不同,引文类型识别更多关注引用功能,揭示被引文献对施引文献的作用,而情感倾向分析更多关注主观情感,揭示施引文献对被引文献的态度。情感倾向分析以往都是对主观性文本信息的挖掘,主要来源于用户生成内容,如对产品信息的评论等[17]。引文内容是施引文献作者对被引文献相关内容的重新组织,属于带有主观情感的用户生成内容,情感倾向分析在引文内容的基础上挖掘施引文献对被引文献的观点倾向,为引文的质量判断提供有效的定性指标。许德山[18]对科技论文引用中的观点倾向进行分析,将语言学领域的主位推进理论和修辞结构理论运用到引用内容的识别和评价倾向的判断中,通过构建引用对象的语篇链路以及话题信息传递模式和观点表达模式的分析,确保评价信息与引用对象紧密相关,并设计情态链路的探测方法,识别评价人的意图走向和态度变化。

引文主题识别相对于引文类型识别和情感倾向分析更加注重引文内容的内部特征,研究目标是利用引文文本来增强引文的语义表达,主要方式是从引文文本抽取代表引文的主题词,利用主题分布来揭示被引文献对施引文献的主要作用或贡献。Small 将引用内容作为观点表达的概念符号,认为将共被引聚类和引文内容分析结合起来能够更好地揭示研究领域的知识基础[19-20]。引文文本周围的词语提供了主题动机的语义线索,分布在核心主题的引文应该得到更高的转变概率。或者说文献可以用很多主题词来表示,引文在文献中分布在不同位置,引文位置相邻文本的主题词在一定程度上表示了引文的主题分布,引文并不跟文献所有的主题词都相关,通过引文分布可以增强引文信息,提高引文分析的精度。Small 利用这种方法对重组 DNA 领域进行了分析,首先利用共被引聚类方法来追溯重组 DNA 领域的演化历史,然后利用引文内容分析来揭示聚类之间主题的变化,并将其引申到共被引内容分析来进一步揭示文献概念之间的关系。该方法的关键是用引文文本内容中出现频次最高的词或短语来表示引文,将引文标签化,在一定程度上对演变进程有了更好的解释。Liu 等[21]提出了全文本引文分析的方法,研究使用了有指导主题模型算法 LLDA(标签化 LDA),LLDA 用于表示文献和引文主题

分布，每一个主题是词语的概率分布，主题标签是一个作者赋予的文献主题词，文献和引文主题概率分布可以被转换成顶点(文献)和边(引文)的转变概率分布，从而来提高引文网络 PageRank。

4 应用展望

目前针对引文内容分析方法开展的相关研究还不是很多，主要影响因素是全文的可获取性和全文文本的识别率等，这相对于传统基于被引用次数的引文分析增加了很多困难，但是其理论意义和应用价值是无可替代的。引文内容分析将成为引文分析未来发展的重要方向，其将外在指标测度同语义内容挖掘的有机融合将会进一步推动引文分析的发展，进一步提升引文分析的学科地位和实践水平。

(1)学术质量及影响力评价。引文分析是进行学术质量及影响力评价的重要方法，主要围绕着论文被引用次数及其改进指标进行测度，另外还可以从引文网络的角度进行中心度等评价。这些方法都是通过计量学指标或社会网络分析指标等来代替论文本身来进行学术质量及影响力的评价，而引用行为极其复杂，引用动机难以度量，存在伪引、漏引等各种不规范或虚假引用现象，因此单纯利用这些数量指标或统计指标等论文外在特征进行评价有一定的不足。引文内容分析作为基于内容的引文分析方法，从论文内在特征出发进行质量测度，能够更加客观地评价引文，并且通过与传统数量指标的结合，实现数量和质量的综合测度，能够更好地用于学术质量及影响力评价。

(2)高被引文献被引原因揭示。高被引文献通常被作为学科领域发展方向的重要里程碑文献和权威文献，并利用高被引文献的研究主题来代表该学科领域的发展方向，这对于学科领域的知识结构演化、关键路径识别、发展趋势探测等具有重要意义。通常利用高被引文献自身的关键词或从标题、摘要或全文抽取的主题词来表征高被引文献研究内容，但关键词或主题词仅仅代表了该论文自身的内容，无法揭示其高被引的原因或一直被引用的原因，而这是研究者更加关注的地方或论文价值体现的地方。引文内容分析将这种从被引文献测度转向了从施引文献测度的思路，通过从施引文献中挖掘代表被引文献的引文内容，不仅能够很好地揭示其被引用的原因，而且基于引文内容识别的主题词能够更加客观地表征高被引文献的研究主题。

(3)科学与技术的关联探测。科学与技术之间是相互促进、相互推动的关系，这种互动关系对于科技政策制定者和科技活动管理者等具有指导意义，直接关系到学科领域的布局和投入等决策[22]，因此科学与技术的关联探测具有重要的实践意义。解析科学论文与技术专利文献之间的共词关系和引用关系是探测科学与技术关联的两种主要途径，其中引用关系被视为隐性知识流动的显性化，可以定量观测科学与技术之间的知识关联[23]。但是目前论文引用专利主要是基于专利号进行识别，而没有对其引用内容或引用态度进行判断，这会导致关联探测的准确度和可信度下降。引文内容分析为科学与技术的关联探测提供了新角度，不仅揭示了科学与技术的互动关系，而且能够揭示两者在哪些方面进行了什么层面的互动，进一步促进其决策支撑作用。

5 结语

引文分析从知识流动的角度揭示科技文献之间的创新关系，具有科学知识评价、科学发展模式揭示和科学前沿探测等功能[24]，具有重要的理论意义和应用价值。但是，传统的引文分析更多注重论文的外在特征，以被引用次数等数量指标来代替论文的内在特征进行价值和质量判断，有一定的不足。引文内容分析是从论文的内在特征出发，从施引文献的客观文本中抽取被引文献的主题内容，有利于揭示施引文献和被引文献之间的创新关系以及被引用原因。引文内容分析与被引用次数等指标的结合，有利于进一步实现数量和质量测度相结合，有利于促进引文分析的深入发展。

引文内容分析通过对引用句或引文上下文的文本内容分析来实现，主要包括数据预处理、引文内容抽取和内容深度分析三个步骤。目前的研究主要集中在引文类型识别、情感倾向分析、引文主题识别等方面，引文类型识别和情感倾向分析更注重引文功能或情感的分类，引文主题识别更注重语义内容的抽取。随着全文可获取性和全文文本识别率的提高，引文内容分析将会进一步地提升其应用水平，包括学术质量及影响力评价、高被引文献被引原因揭示、科学与技术的关联探测等方面。

参考文献

[1] 邱均平. 文献信息引证规律和引文分析法[J]. 情报理论与实践, 2001, 24(3): 236-240.

[2] Small H. Update on science mapping: Creating large document spaces[J]. Scientometrics, 1997, 38(2): 275-293.

[3] 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.

[4] 潘教峰, 张晓林, 等译. 第四范式: 数据密集型科学发现[M]. 北京: 科学出版社, 2012: 199.

[5] 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.

[6] 吴志荣. 对引文分析法方法论地位的重新思考[J]. 图书馆杂志, 2012, 31(5): 11-13.

[7] Garfield E. Can citation indexing be automated?[C]. Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington, 1964: 189-192.

[8] Weinstock M. Encyclopedia of Library and Information Science[M]. New York: Marcel Dekker, 1971(5): 16-40.
[9] Ding Y. Content-based citation analysis: The next generation in citation analysis [EB/OL]. [2012-11-14]. http://www.lis.illinois.edu/events/2012/09/26/content-based-citation-analysis-next-generation-citation-analysis.
[10] Bornmann L, Daniel H D. What do citation counts measure? A review of studies on citing behavior[J]. Journal of Documentation, 2008, 64(1): 45-80.
[11] Moravcsik M J, Murugesan P. Some results on the function and quality of citations [J]. Social Studies of Science, 1975(5): 86-92.
[12] Frost C O. The literature of online public access catalogs, 1980-85: An analysis of citation patterns[J]. Library Resources and Technical Services, 1989(33): 344-357.
[13] Maricic S, Spaventi J, Pavicic L, et al. Citation context versus the frequency counts of citation histories[J]. Journal of the American Society for Information Science, 1998(49): 530-540.
[14] Hanney S, Frame I, Grant J, et al. From bench to bedside: Tracing the payback forwards from basic or early clinical research-A preliminary exercise and proposals for a future study[R]. HERG Research Report No. 31, Health Economics Research Group, Uxbridge: Brunel University, 2003.
[15] Pham S B, Hoffmann A. A new approach for scientific citation classification using cue phrases[C]. Proceedings of Australian Joint Conference in Artificial Intelligence, 2003: 759-771.
[16] Shotton D. CiTO, the Citation Typing Ontology, and its use for annotation of reference lists and visualization of citation networks[J]. Journal of Biomedical Semantics, 2010, 1(Suppl 1): S6.
[17] 黄萱菁, 张 奇, 吴苑斌. 文本情感倾向分析[J]. 中文信息学报, 2011, 25(6): 118-126.
[18] 许德山. 科技论文引用中的观点倾向分析[D]. 北京: 中国科学院文献情报中心, 2012.
[19] Small H G. Cited documents as concept symbols[J]. Social Studies of Science, 1978, 8(3): 327-340.
[20] Small H G, Greenlee E. Citation context analysis of a co-citation cluster: Recombinant-DNA[J]. Scientometrics, 1980, 2(4): 277-301.
[21] Liu X, Zhang J, Guo C. Full-text citation analysis: A new method to enhance scholarly network[J]. Journal of the American Society for Information Science and Technology, 2012.
[22] 李 睿, 孟连生. 论基于专利引文的科学——技术关联探测方法中存在的问题[J]. 情报理论与实践, 2010, 33(3): 87-90.
[23] 李 睿. 基于专利引文分析的科学——技术关联探测模型改进[D]. 北京: 中国科学院文献情报中心, 2011.
[24] 梁永霞. 引文分析学知识图谱[M]. 大连: 大连理工大学出版社, 2012.

[作者简介] 祝清松, 男, 1985年生, 中国科学院国家科学图书馆博士研究生。
冷伏海, 男, 1963年生, 中国科学院国家科学图书馆研究员、博士生导师, 情报研究部主任。
收稿日期: 2013-03-02