网络出版时间：2016-04-22 10:11:26
网络出版地址：http://www.cnki.net/kcms/detail/11.1541.G2.20160422.1011.005.html
图书情报工作
第 60 卷 第 6 期 2016 年 3 月

引文文本分类与实现方法研究综述*

王文娟[1,2] 马建霞[1] 陈春[1] 张凌波[3]
[1]中国科学院兰州文献情报中心 兰州 730000 [2]中国科学院大学管理学院 北京 100049
[3]兰州大学附属中学 兰州 730000

摘要：[目的/意义]对引文文本分类的标准、实现方法和应用进行梳理，分析存在的问题，提出可改进的方向。[方法/过程]总结目前引文文本分类的几个重要角度，如基于引用功能、基于情感倾向、基于引文影响力等，对引文文本分类的实现方法进行比较，分析其优缺点。[结果/结论]目前引文文本没有统一的分类标准和实现方法，引文文本的获取较为困难，计算机分类算法准确率较低，中文引文文本分析文献少。未来研究思路和方向应该是：统一文本分类的标准，提高引文文本计算机处理技术的准确性，扩大应用范围。

关键词：引文文本分类 引文内容分析 引文分析
分类号：G250
DOI：10.13266/j.issn.0252-3116.2016.06.018

引文文本一般是指引文脚注所在的句子或与上下文句子的集合[1]，能提供施引文献与被引文献之间关系的重要信息，具有重要的研究价值和意义。目前围绕引文文本的研究主要有引文文本分类、引用动机调查、引文主题抽取3个方向[2-5]。相对而言，引文动机调查和主题抽取的理论和方法比较成熟，而关于引文文本分类的研究文献数量众多，分类标准和实现方法各有不同，没有统一的模式[6]。近几年来，国内关于引文文本的研究文献逐渐增多，如文献[6]对引文文本分析方法的主要步骤和相关研究进展进行了综述；文献[7]概述了引文文本类型识别的步骤和实践进展；文献[8]尝试从概念、研究范畴、步骤和功能系统地构建引用内容(文本)分析的理论。然而这些文献中较少对不同的引文文本分类标准的特点进行归纳，对不同的分类方法的优缺点也没有系统梳理和对比。因此，本文拟在前人研究的基础上对中英文相关文献进行研读、梳理，系统总结归纳现有引文文本的各种分类标准及特点，对比分析引文文本分类实现方法的优缺点，理清引文文本分类的主要应用领域，剖析当前存在的问题，并对未来的研究重点进行展望。

1 引文文本的分类标准

引文文本的分类标准，主要是指科学家对引文文本进行内容分类分析时采取的角度或维度。在对引文文本分类的相关研究中，1993年，M. X. Liu[2]将相关研究的分析目的归纳为3个方面：提高检索效率，研究引用功能，研究引用质量。2004年，H. D. White[9]从情报学和语言学的不同学科角度，探讨了不同学科对引文文本分类方法的不同研究角度。2013年，祝清松等[7]在对引文类型标注方法进行综述的基础上将分类标注定为引用功能和观点倾向两类。笔者通过对主要综述性文献的比较和其他相关文献的广泛研读，认为引文文本的分类方式主要可分为以下5种类型。

1.1 基于引用功能的分类

引用功能是指被引文献在施引文献中起到的作用和产生的意义，早期对引用功能的研究以描述性讨论为主，功能定义比较单一，主要观点认为引文是对前人研究工作价值的肯定，是对知识产权的维护[10-14]。其中，具有代表性的是M. J. Moravicsik与P. Murugesan[15]从4个不同维度对引文文本进行分类，特别是第一个维度——将引文文本分为概念性引用、操作性引用和其他功能的引用，超过一半的引用(53%)为概念的引用，仅有7%的引用属于其他功能的引用。其对引用功能的划分比较粗略，没有考虑到更为复杂的情况，其后有很多学者对其加以应用和改进。

*本文系国家自然科学基金项目“基于科学基金项目及知识产出的研究前沿探测”(项目编号：71373260)研究成果之一。
作者简介：王文娟(ORCID：0000-0001-5233-1387)，硕士研究生；马建霞(ORCID：0000-0002-5401-9992)，学科咨询部主任，研究馆员，硕士生导师，通讯作者，E-mail：majx@126.ac.cn；陈春(ORCID：0000-0003-4351-4696)，副研究馆员，硕士；张凌波，中学生。
收稿日期：2015-12-22 修回日期：2016-03-05 本文起止页码：118-127 本文责任编辑：刘远颖

C. Oppenheim 和 S. P. Renn[16]将引文句的功能划分为7个大类,以1930年以前出版的28篇论文的施引文献为例,发现出版很久的论文仍被高频引用主要是因为该文献提供了研究的历史背景,而不是具体内容被借鉴。S. Cole[17]为研究社会科学家 R. Merton 提出的"社会结构与失范"理论思想的价值,将 R. Merton 相关论文被引文献中的引文文本分为10类,认为其他作者引用 R. Merton 的论文可能是为施引文献提供论据。Y. W. Chang[18]研究 *Little Science*,*Big Science* 杂志对自然科学和社会科学的作用与影响,抽取该杂志的引文文本进行人工判断和统计分析,发现虽然在自然科学和社会科学所有文献中引用该杂志比例最高的两个功能都是提供证据、展示相关研究,但是功能所占比例及比例排序都有不同。

以上不同学者提出的分类研究系统有明显差异,但也存在一定的相似性,因此 M. Camacho-Minano[19]使用文献计量和内容分析方法,对127篇研究引用功能分类的文献进行聚类,最后将引用功能总结为九大类。M. Camacho 的结果与 M. J. Moravcsik[15]、D. E. Chubin 等[20]的分类类似,但是没有指出与其他分类之间的对应关系。国内学者陈晓丽[21]选取我国社会科学一级学科具有代表性的期刊,通过分析期刊中文献引文文本和注释,将引用功能分为6个大类、19个小类,并认为该分类更符合中文期刊中引用功能的情况。

上述各学者的观点总结见表1。从引文文本功能分析角度进行研究,发现文献、期刊、学科层面更深层次的引用规律,虽然基于引文功能的划分标准目前尚不统一,但这些信息是基于被引频次的引文分析方法无法识别的,不同研究的实验结果也都充分说明引文文本分析能够作为学术评价、提高检索准确性的重要参考。

表1　不同作者基于引用功能的分类观点

作者	研究目的	观点
C. Oppenheim 和 S. Renn[16](1978年)	研究特定文献的被引规律	①历史背景;②对相关著作的描述;③提供并不用于比较的信息或数据;④提供用于比较的信息或数据;⑤使用理论方程;⑥使用方法;⑦不可应用或不是最佳的理论或方法(意思是对理论或方法的否定)
S. Cole[17](1989年)	研究特定科学家的被引规律	①相关著作的一部分,在分析中不担当明确的角色;②支持作者的观点,合理化作者的观点和解释;③使用 R. Merton 提出的概念;④扩展或修改 R. Merton 的理论,或者作为作者理论的一部分使用;⑤解释引文研究的结论;⑥规划研究问题的步骤;⑦尝试验证一个衍生的理论;⑧尝试验证"社会结构与失范"理论的部分;⑨对"社会结构与失范"的批评;⑩其他
Y. Chang[18](2013年)	研究期刊的被引规律	①提供背景信息;②用于比较;③引用定义;④引用支持的证据;⑤引用图表;⑥提供扩展阅读;⑦引用使用过的方法;⑧列出相关研究;⑨提供补充或解释;⑩引用术语;⑪引用观点
M. Camacho[19](2009年)	研究引用功能的共性	①概念的;②操作的;③根本的;④敷衍的;⑤演绎的;⑥并列的;⑦肯定的;⑧否定的;⑨其他
陈晓丽[21](2003年)	理论探讨	①对论著支持者表示谢意;②列举课题参与者;③对原文进行补充说明(a. 译名的其他译法;b. 举出文中没有论述的其他观点;c. 补充限于篇幅而未展开的内容;d. 解释文中不易理解的词语;e. 进一步对论述进行解释;f. 举出文中叙述的实例;g. 提供原文的论证);④对原文提出辩证、商榷、考证等意见(a. 纠正引文中的错误;b. 对原文引用的外文中的中国地名、人名进行考证;c. 对原文引用观点提出商榷意见;d. 对沿用的名字译法提出异议;d. 对原文论述提出新的见解);⑤列出可供参考的文献;⑥标明来源(a. 指出原文中引文的版本;b. 指出统计样本的来源;c. 标明语词出处)

1.2　基于引用动机的分类

引用动机是指施引作者引用对应文献的心理内在和外在动因,即基于什么目的引用。引文功能从引用的结果对引文的作用、意义进行研究分析,而基于动机分类的引文文本分析,是从引用原因探究引文的作用和价值,两者分类的角度有所不同,但研究目的是一致的,甚至在很多文献中,将引文动机与引文功能的分类方法不加以区分[13,22－23]。早期的研究中,学者主要认为引用动机出于"说服"目的,施引为了提高文献的权威性[24－26],增强对读者的说服性。

1965年,E. Garfield[27]为了探究作者施引的原因,结合高能物理学的期刊文献中的具体引文文本案例,将引用动机归纳为15大类。其对引用动机的划分被 M. Weinstock[28]继承,并利用科学引文文本的实例进一步说明每一类别的具体意义,E. Garfield 的划分标准奠定了关于引用动机的研究基础。

不同于 E. Garfield 和 M. Weinstock,F. C. Thorne[29]认为施引作者的引用动机里会有一些"不公正"的因素,并随机选取了物理学主要期刊文献1945－1975年的引文文本案例,将引用动机划分为19个大类,考虑到作者作为社会人的属性,除了"公正"的动机外,还有自利、政治因素、内部争斗等"不公正"的动机。S. K. Sen[30]选用情报学的期刊文献对不同动机进行描述,认为自引、对重要同行的引用等行为是出于增

加重要性的动机,对于这种动机是否“公正”并没有做说明,态度倾向中立。

将以上作者的动机分类进行梳理,见表 2。使用引文文本对引用动机进行分类研究的文献并不多见,原因是很多学者研究发现使用问卷调查或访谈的方式更能准确了解施引作者的引用动机[31-33]。如 1985 年 T. A. Brooks[32] 访谈某大学的 26 位科研工作者,最终形成七大类动机描述,对比之下认为从引文文本获得的引用动机数据并不可靠,此后关于引用动机的研究文献(包括国内的研究文献),多摒弃引文文本分析方法而普遍采用问卷调查或访谈法[34-35]。

表 2 基于引用动机的引文文本分类观点

作者	动机倾向	观点
E. Garfield (1965/1977 年)、M. Weinstock(1971 年)	“公正”的动机	①对开拓者的尊重;②对有关著作的信任;③验证其所用的方法和仪器等;④提供背景材料阅读;⑤对自己的著作予以更正;⑥对别人的著作予以更正;⑦评价以前的著作;⑧对自己的论点寻求充分的论证;⑨提供研究者现有的著作;⑩对未被传播、很少被引用或未被引证的文献提供向导;⑪验证数据及物理常数等;⑫核实原始资料中某个观点或概念是否被讨论过;⑬核查原始材料或其他著作中的起因人物的某个概念或名词;⑭否定他人的著作或观点;⑮对他人的优先权提出异议
F. Thorne(1977 年)	“公正”与“不公正”	①连续出版;②多重出版;③象征性引用;④详细引用;⑤协作报告;⑥证据有效性;⑦自利的引用;⑧深思熟虑的预谋;⑨寻找基金支持;⑩出版物的基金支持;⑪编辑的偏好;⑫主观偏好;⑬互相照应;⑭迎合压力;⑮出版社的政策;⑯忽略新作者;⑰内部斗争;⑱过时的引用;⑲政治因素
S. Sen(1990 年)	中立	①基于被引文献产生的新想法;②实验辩护;③理论辩护;④反驳批评;⑤提供可选项;⑥历史背景回顾;⑦获得权威作者的支持;⑧列举相似相关文献;⑨做短暂评论;⑩使用结果总结;⑪常规名义上的讨论;⑫出于礼貌;⑬创造一致的观点;⑭增加自己或同行的重要性

1.3 基于情感倾向的分类

引文内容的情感倾向是指施引作者对于引文文献的态度——是支持,还是反对,或是保持中立。

M. J. Moravcsik 和 P. Murugesan[15] 较早对引文情感的不同进行研究,在分类系统中使用了正向与负向对引文进行分类,发现在高能物理学中约有 1/7 的文献是属于负向引用。他们还在后续研究中,使用了同样的分类系统分析物理学 4 本期刊之间的引用情感倾向,发现从总体上看每本期刊获得的负向引用比例并没有什么差别,但是 *Physics Review* 对 *High Energy Physics* 持有的反对意见是所有杂志中最多的,达到 14%,相比之下 *Nuclear Physics* 杂志对 *High Energy Physics* 杂志的反对意见仅有 2%[36]。

D. E. Chubin[20] 改进了 M. J. Moravcsik[15] 的分类系统,以分类树的形式对引文文本进行分类,分类树的第一层分为正向和负向两个子树,发现在高能物理学杂志中读者评论文章中的负向引用比期刊论文中的负向引用高很多。D. E. Chubin[20] 还发现被负向引用的论文在被证伪之后还会在一定的时间范围内继续被引用,但作者认为被负向引用的论文并不代表没有价值。

B. V. D. Martens 等[37] 则将引用内容的情感分析应用于交叉学科,发现绝大部分文献中的引文文本是正向的,中性文本的比例为 39%,负向文本所占比例仅有 4%。B. V. D. Martens 列举了有代表性的实例来说明不同情感倾向的表达方式,但是对处理方式的细节没有做过多的说明。

2014 年,C. A. Sula[38] 通过对情感倾向的表述规律进行分析,认为引文功能的分类表述可以与情感倾向对应。刘盛博等[39] 也采用类似的方式将引用功能与情感倾向进行对应,如认为对引文中方法、技术、数据等的引用以及对方法、结果的比较属于正向引用,对引文中观点的批评则属于负向引用。

通过对引文内容情感倾向的研究可以发现引文继承关系中的态度,进而可以分析学术期刊之间的亲近关系,还可以结合引文的其他属性,如位置信息,发现学术引用中更为潜在的引用规律。

1.4 基于影响力的分类

影响力是引文关系中最为直接的概念,考察引文文献对施引文献贡献的大小。同一篇引文对不同施引作者的影响力是不一样的,同一篇文献中不同引文的影响力也是不同的。

M. H. Macroberts 和 B. R. Macroberts[40] 在基因学领域选取若干篇文献的被引文献,比较引文文本与施引文献和被引文献内容的相关性并将其分为“有影响力的”和“无影响力的”两类。发现施引作者会存在一定的隐瞒行为,比如,施引作者会以平淡的语言描述对文献写作具有重要作用的文章,以掩盖被引文献的真正影响力。

S. Bonzi[41] 根据引文文本的表述方式和引文在施引文献中出现的次数,将被引文献的影响力程度按等级划分,发现大部分被引文献只是被作者在文中简单提及,而这些文献对确定施引文献的主题并没有帮助。

X. J. Wan[42]将实验数据的引文文本按影响力程度先分为5个等级,分值为1 - 5,并将其设为因变量,以引文出现的次数、出现的位置、引文发表时间、引文句子的平均长度、引文发生的平均密度、是否自引作为自变量进行回归分析。方程通过检验后,根据获得的方程系数计算其他文献、作者的影响力程度,结果显示84%的文献影响力值在2 - 3之间,即绝大部分引文是不重要的或重要性一般的,这与M. H. Macroberts等[43]、S. S. Teufel等[44]研究发现绝大多数的引用是冗余的结果具有一致性。

陈晓丽[45]从引用的力度和深度探讨引文对施引文献的影响力,引用的力度主要根据引文文本来进行判断,分为有力引用、适度引用、表面引用和无关引用。其中,有力引用是指"引文对文章立论、提供论据、继承发展学术成果起到较大的支持作用,文献对引文的依赖也较大",与引文的功能具有一定的关联性。

不同作者基于引文影响力的分类观点总结见表3。引文的影响力既可以从引文文本的定性角度去分析,也可以通过一定的处理方法设置定量计算的指标,分析的角度较为多元化。而直接获取影响力分组的研究相对较少,更多的研究是将引文文本作为计算影响力分类方式的一个辅助。

1.5 基于引用位置的分类

一般而言,科学文献的编纂有一定的组织结构和格式,如分为背景、方法、结果、讨论4个部分[46]。引文文本的位置即为所在的结构组织,而引文文本的位置蕴藏着丰富的信息,既能结合引文文本的其他属性研究引文的特性,也可以通过章节提高检索效率,另外,还能够通过引文文本的位置,发现学科之间的引用规律。

B. C. Peritz[22]研究引文文本所在的位置与引文功能之间是否具有一致性,将引文位置分为介绍、方法、结果、讨论和总结、附件5个部分,发现在不同的杂志中引文文本位置并不能代表引文功能,作者指出将功能与位置进行统一对应有时候会失灵。G. Herlach[47]通过药学期刊的少量的样本研究参考文献在文献结构中被多次提及的情况,主要分为4个部分:介绍、方法、结果、讨论,发现大约1/3的引文在全文中会被多次提及,被多次提及的引文出现在讨论章节中的次数明显比其他章节要高,总体来说,在介绍部分出现的引文数量最多。G. Halevi等[48]则通过引文内容的位置研究信息计量引用其他学科和引用本学科文献的情况,发现在引言部分,引用外部学科与内部学科的引文数量相差不大,说明信息计量学的研究视角是广泛的,有可能受其他学科的影响比较多。

另外,将位置信息与引文的其他特征(如被引时间)相结合也能发现一定的引文规律。刘茜等[49]通过引文位置的时空序列变化,发现当一篇文献发表之后,若其观点在初期未被其他人认可,则对其文献的引用主要出现在背景章节和讨论章节中,直到其被广泛接受后,在讨论章节对其引用会逐渐增多;若其观点在初期即被认可,则会在背景章节中大量被引,在讨论章节

表3 不同作者基于影响力的分类观点

作者	分类	描述或示例
M. H. Macroberts等[40](1986年)	有影响力	施引文献内容与被引文献内容相似
	无影响力	施引文献内容与被引文献内容不相似
S. Bonzi[41](1982年)	没有特别提及	很多学者研究发现…
	简单提及	张三研究…
	一次引用或讨论	张三发现了…
	一次以上的引用或讨论	/
X. J. Wan[42](2014年)	非常不重要	引文与施引文献相关性极小,很少在文中被提及,移掉后对施引文献并没有影响
	不重要	引文与施引文献相关性较小,较少被提及,可以被其他文献替代
	一般	引文与施引文献相关,一般在一个句子中被提及,有一定用处但没有显著的影响。评价度量或一般工具的引用多属于此类
	重要	引文与施引文献高度相关,通常在文中被提及多次,文本的长度很长,有明显的影响力
	非常重要	引文与施引文献高度相关,多次被提及和强调。施引文献一般是被引文献的扩展,被引文献是施引文献的思想基础。有时候,被引文献是重要的用于比较的标杆
陈晓丽[45](2000年)	有力引用	如《中国社会科学》1998年第4期对《中国著作权法》以及香港、台湾、澳门等地著作权法的引用
	适度引用	如某个主题、某个观点、某个论断或某一段落有一定的支持作用
	表面引用	如《中国社会科学》1998年第4期第153页"堡垒"概念的引用
	无关引用	如只是为了弄清文献叙述中提及的无关紧要的某一词语、概念、事件等的来龙去脉或延伸意义

中出现的次数较少。

基于功能、动机、情感倾向等的分类标准是基于引文文本的语义信息，对比而言，引文位置并不关注文本的语义信息，是对引文文本的物理划分，在引文文本分类研究的文献中可结合其他属性探究引用规律，既可独立作为引文文本的分类标准，也可以辅助其他分类标准用于科学研究。

2 分类方法的实现

早期基于引文文本的分类研究处理的样本比较少，相关研究者一般根据专业知识背景和经验对文本进行分类。随着计算机技术的发展，分析处理样本量越来越多，方法也层出不穷，本章节主要对几种不同的分类方法进行简单介绍。

2.1 人工标注

早期利用引文文本进行分类，一般是从理论描述上根据经验总结进行分类，几乎没有验证的引文文本样本。如上文提到的 E. Garfield[27] 对引用功能的分类描述，是基于个别引文文本案例的总结，并没有用大量的样本数据进行验证，根据 M. Camacho-Minano[19] 的总结，早期分析处理的引文文本一般不超过 1 000 条，仅 P. Murugesan 等[50] 与 B. C. Peritz[22] 在研究中使用的引文文本稍微多一些，前者使用了交叉学科的 1 501 条引文文本，而 B. C. Peritz 处理的样本为 2 209 条。

由于人工标注是基于个人的经验进行的，比较主观的，为了避免这种主观性，学者会通过比较多个主体对引文文本的分类结果来验证分类标注的准确性。

2.2 计算机标注

随着计算机技术的快速发展，研究者开始考虑如何将分类标注这种庞复的工作交给计算机去处理，在探究的过程中主要形成了两种不同的计算机处理方式：一种是基于线索词规则判断，另一种是基于机器学习的分类方法。

2.2.1 基于结构线索词的方法 1986 年，J. Swales[51] 通过对引文文本的分析，认为科学论文中的引文一般遵循一定的修辞结构，S. S. Teufel 等[44] 也通过引文文本发现了这一规律，这一发现构成了基于线索词规则的引文文本分类标注基础。

M. Garzone 等[52] 使用 8 篇物理学文献和 6 篇生物文献中的所有引文文本，通过仔细阅读，抽取引文中的线索词和语法结构构建线索词词库，比如若引文文本出现在论文的结果章节且包含“假定”“阅读”“报道”等词汇，则将其归类为用于发展新的模型和假设的引用。M. Garzone 等的分类系统综合考虑线索词和结构信息，最终正确率为 78%，另外 11% 的引文分类部分正确（同一条引文可以归属不同分类）、11% 的引文分类错误。

H. Nanba 等[53] 则简化引用功能的分类系统，仅分成三大类处理 282 条引文内容，抽取出了 160 个线索词。其中 76 个线索词属于 B 类（基于他人的理论或方法的引用），如“主要基于”“被用于”“我们能够”等；84 个线索词属于 C 类（对相关研究进行比较或指出其问题的引用），如“然而”“尽管如此”“但是”等；B 类和 C 类无法识别的文本全部归为 O 类。处理过程为先抽取出引文文本的所有线索词，然后与不同类别的词库进行匹配，最后得到 76.9% 的准确率。

刘盛博等[54] 根据引文文本中的情感词汇，将引文的情感倾向主要分成正向、中性、负向三大类，考虑句子的主语，如正向引用中，主语为施引文献线索词则有“更好地”“应用”“使用”等，而当主语为引文时则线索词有“显著的”“基础性的”等。他研究了 *BMC Bioinformatric* 杂志被引的情感倾向，共处理了 150 661 条引文文本，数据量相当庞大，准确率为 95% 左右。

结构线索词方法的关键步骤在于区分引文文本的主语和谓语的时态，然后对文本句子进行一定的转换，线索词一般集中在谓语的词性和连接介词的词性上。

2.2.2 基于机器学习的方法 随着计算机技术的进步，机器学习技术越来越成熟而被人们广泛用于各种科研数据处理中，引文文本的分类算法也不例外。

较早的有 S. B. Pham 和 A. Hoffmann[55] 使用一种涟波下降规则（Ripple Down Rule），类似机器学习中的决策树分类算法。S. Pham 和 A. Hoffmann[55] 将引文文本基于情感倾向和功能主要分为 4 类：基础、支持、局限、比较，先使用 482 条引文文本构建基本特征向量，再使用 150 条引文文本对特征向量进行补充，最终获得的正确率可以达到 94% 以上，支持功能类型的引文识别正确率最高，达 97.3%。

机器学习中常用的分类算法还有支持向量机（SVM）和朴素贝叶斯，这两种方法也适用于引文文本分类，如 X. D. Zhu 等[56] 使用 SVM 将引文的影响力分成有影响和无影响两类，并将结果结合其他数据用于文献的影响力分析，R. Radoslav[57] 使用朴素贝叶斯改进 M. Garzone 等[52] 的规则算法，对 360 篇会议论文进行标注，以 2 829 条引文文本进行测试，小范围提高了

准确率。

2010 年,N. K. Agarwal 等[58]为了比较 SVM 与朴素贝叶斯两种机器学习算法在引文文本的分类效果上的优劣,选取了 1 710 条引文文本进行标注作为训练样本,再分别使用 SVM 和朴素贝叶斯算法对引文进行分类,得到的结果是 SVM 比朴素贝叶斯的准确率稍高一点。

另外,还有很多其他机器学习算法应用于引文文本分类,如 M. H. Le 等[59]的马尔科夫链的两种模型应用、S. M. Wang 等[60]的随机场应用等。这些算法的应用相对 SVM 和朴素贝叶斯使用率较低,准确性也并没有明显提高,实施过程比较复杂。

2.3 不同分类方法的比较

在方法的效率上,基于经验的人工标注方法需要耗费较多的时间和人力,因此在资本投入和样本量之间存在不可调和的矛盾,而计算机标注的算法效率比较高,尤其是前期规则提取的工作准备好以后,则全部由计算机自动完成,无需人力投入。

在方法的准确性上,基于经验的人工标注方法一般由具有学术背景的专家对引文文本逐条判断,准确度比较高。计算机标注中,基于结构和线索词的方法则完全依赖于结构和线索词规则提炼的完备性;基于机器学习的算法则依赖于训练样本的选取,样本的偏差性直接影响实验结果的准确度。

在标注要求上,人工标注和基于机器学习的算法要求标注者具备一定的学术背景,能够准确判读引文文本的意义,而基于结构和线索词的算法除此之外还需要标注者具备一定的语言学基础,能够对引文文本的结构进行切分和语法转换。

在方法的重合性上,基于机器学习的方法处理过程中需要人工对实验数据的训练样本进行标注,在算法特征的选择上,也可能会以结构和线索词的某些规则作为机器学习的一部分特征。

在样本规模上,人工标注的处理方式耗费较多精力,并且早期能获取的全文数据有限,处理的样本量较少,大多数样本量在几百条,少数会多至千条以上。当计算机技术引入之后,样本绝大多数在千位量级,甚至部分文献处理的样本量达到了 10 万条以上[39,61]。线索词和机器学习两者比较而言,由于方法本身的特点,机器学习算法普遍使用的训练样本要更多一些,在一定范围内,充足的样本量有助于提高准确率。

基于以上对引文文本分类实现方法的比较,对它们的优缺点进行梳理,结果如表 4 所示:

表 4 3 种引文文本分类实现方法的优缺点

算法 / 特点	基于经验	基于结构和线索词	基于机器学习
优点	准确率高	节省人力;便于梳理不同分类的对应规则;能为主题分析提供支持;处理样本容量高	具有可移植性、可扩展性;节省人力;样本容量高,可验证
缺点	需要投入大量人力和时间;样本量一般相对较少	对句子依赖性过高;无法穷尽所有规则;需要标注者有一定的语言学基础;准确率低	难以人工干预;准确率不高;依赖于训练样本的分布

3 引文文本分类的应用

引文文本来源于引文全文数据的获取,技术的进步给全文数据的获取提供了便利,也因此促进了引文文本分析的实际应用。引文文本分析的主要方面均有不同的应用前景,如引文文本的主题抽取可用于聚类,更进一步地分析某研究领域的主要思想的交互和演化;引用动机的调查可以探索研究人员的引用行为与引文之间的关系,进一步评估引文的重要性。相比之下,引文文本分类研究的应用范围有相似性,也有独特之处,主要分为以下 4 个方向。

3.1 用于科学评价

早期 M. J. Moravcsik 与 P. Murugesan[15]、J. Cole[17]、D. E. Chubin 等[20]均通过引文文本的分类调查研究某个领域内的引用冗余情况,进而说明被引次数难以代表真正被引文献的真正影响范围。而在基于情感倾向分析的研究文献中,则发现绝大部分引文是中性描述的,而即使是负向引用,在总体上只占很小的比例,仍然有其特有的规律,如被负向引用的文献在证伪后一段时间内仍能获得大量引用,之后则再难被引[62]。这些研究均反映出长期以来基于被引次数评价学术影响力的方式存在缺陷,而引文文本能够为学术评价提供定性角度的重要信息,可以作为定量评价方法的重要补充。

另外,我国的自然科学基金评价标准,提出需要指出引文中对该著作或个人的正面评价和负向评价,其中正负评价倾向必须从引文文本中提取。从不同的维度(如情感倾向、功能、被引力度等)对引文被引的情况进行详尽的分析,能够对机构、期刊、个人的学术水平做出更为清晰、公正的评价。

3.2 用于构建科学地图

引文文本往往能够反映从题目和摘要中难以识别出的信息,是施引文献和被引文献之间最为直接的联

系,是内容上的交集。

K. W. McCain 和 L. J. Salvucci[62]使用 1976 - 1999 年引用布鲁克斯的一篇文章的 497 篇论文、574 条引文文本,来分析"布鲁克斯规律"的传播,应用引文内容分析方法研究布鲁克斯的学术思想在学术交流中的扩散情况。C. A. Sula 和 M. Miller[38]应用计量学的方法,通过引文文本的联系来定义人类学期刊中各作者之间的关系地图,分析不同作者的差异性、相似性和继承性等。孙海生等[63]利用引文文本,采用思维导图的方式研究学科领域的发展轨迹,能够清楚地显示重要文献所起的作用。可见,基于动机或情感倾向的引文文本分类能够反映科学地图中的隐性信息,基于影响力的分类提供地图中的权重信息,引文文本中各种属性都能为构建科学地图提供更详细精确的路线。

3.3 用于信息检索

早期的相关研究明确提出对引文文本分类可以用于提高信息检索的效率[64]。常规的检索系统只能根据引文是否被引(或者共被引关系等)来提供检索文献,检索结果中往往包含很多无关的文献,需要科研人员对文献的具体内容进行判读。引文文本分类是对引文关系的细分,提供的引文被引位置、情感倾向、所起作用等信息便于科研人员快速定位到目的文献,免去了对不相干文献进行排查的麻烦。

CiteSeer[65]是比较成熟的基于引文文本的搜索系统,但是只支持计算机科学类的文献,范围还比较小。另外,越来越多的研究者探索基于本体方式实现引文文本的搜索引擎,主要根据文本分类的关系确定引文三元组关系描述,提供搜索引擎,如 T. Fujiwara 等[66]的 Coil 系统、D. Shotton 等[67]创建的语义出版和参考本体系统。

3.4 用于信息推荐

研究者在论文写作之前都需要搜集尽可能多的相关研究资料,并在著作中正确引用,根据引文文本获得的引文之间的关系则能大大减少研究过程中资料搜集和标引的劳动量。刘盛博等[68]研究发现基于引文文本的检索结果的准确性高于基于主题在 SCI 和 Google Scholar 中检索的准确率,为基于引文文本的推荐系统提供了理论证明。早在 1981 年,J. Duncan 等[69]就针对如何构建推荐系统做出了尝试——提供一个在线的引文推荐系统,主要是在教育学领域内,该系统虽只是实验性的,但其效果可观。Q. P. J. He 等[70]、F. Zarrinkalam[71]也在其他领域对推荐系统的构建做出了尝试,并取得了很好的结果。虽然基于引文文本的推荐系统,其应用范围还存在局限,但其实际意义不可忽略。

4 存在的问题

目前为止,关于引文文本分类的研究较多,不同学者应用引文文本分类的研究方法在很多学科中都取得了一定的研究成果,然而仍然存在一系列问题。

4.1 分类标准不统一

目前引文文本的分类标准多是基于研究者个人的研究目的而定的,缺乏统一的认识和理解。对类似的文本,有的学者在引文功能中区别引文文本之肯定和批评意见,而有的学者则认为都是对前人工作的阐述,不必加以区别。另外,引文文本的划分特征数量也不一致,不同的分类系统之间无法实现一一对应。

4.2 分类样本不充分

引文文本分类研究的前提是获取全文样本,再从中抽取引文文本句子。然而由于多种原因,比如版权保护、语言障碍等,使得所有学者研究中所获取的全文文本数量有限。在可获得的全文文本中,绝大部分全文数据库提供的格式为 PDF,而不同的数据库 PDF 编译的格式又不一样,这也给计算机抽取引文文本带来了挑战。

另外,使用计算机抽取引文文本,则很可能存在一定范围的抽漏、抽错等现象,目前相关研究中尚未有抽取过程准确率的评估研究。

4.3 准确性有待提高

准确性是指计算机引文文本分类结果与人工分类结果的相对准确率,目前计算机技术已经相对比较成熟,然而在相关的研究案例中也可以发现,分类结果的准确性各有不同,绝大多数研究结果的分类准确率在 80% - 90% 之间,准确率低使得引文文本难以投入到实际应用中。在将来的技术处理中,提高准确率是亟待解决的一大难题。

4.4 中文为实验数据的研究文献少

目前关于引文文本分析的研究文献中使用的研究数据以英文为主,仅有少量其他语种的文献,在国内的研究文献中以中文引文文本为研究对象或实验样本的研究文献数量很少。主要原因在于中文全文数据获取并不方便,常见的中文数据库如 CNKI、维普、万方等提供的 PDF 格式标准不一,且容易出现乱码。相比之下,英文数据库如 Citeseer、PubMed 等全文数据则能提供结构比较完整的引文文本和格式。另外,中文的表述格式较为复杂,中文引文文本分析的难度较大,对中文引文文本的分析需要依赖分词技术和语义分析技术

的发展,但就目前而言,中文的分词技术还有待改进。

5 结语

引文文本分类研究从更深层次对引文关系进行研究,弥补了从引文被引频次来评价引文价值的评价方法中将所有引文价值等同的不足,更科学地反映了引文的真正作用和重要程度,笔者认为未来可以从以下几个方面进行深入分析:

5.1 确定分类标准的规范

不同的学者在研究中所采用的分类标准不同,判断规则也不明确,使得不同的研究结果无法进行整合,没有可比性,因此需要对分类标准进行统一。分类的标准决定了分析结果的有效性,规则的确定能有效避免研究者的主观性对分类结果造成的影响。只有实现标准规范化,才能使得分类分析的结果更为严谨。

5.2 提升引文文本处理技术

引文文本分析的准确度取决于文本处理技术的发展,其中包括两个方面:一是引文文本的抽取,二是引文文本的语义分析。引文文本是否能够被有效、完整地识别是后续分析的基础,引文文本的语义分析则是判断的直接依据,引文文本分析的进一步发展必须以克服对于人工手动处理的依赖为前提,只有当引文文本处理的技术提升了,引文文本分析才能得到更大范围的使用。

5.3 增加引文文本分类的实际应用

引文文本分析能够为科研人员的学术水平评价、科研人员的职称评定、科研基金的分配、核心期刊的评定、研究机构创造力的提升等提供重要的参考信息,在实际应用方面具有广阔的前景。目前的研究主要集中在理论探讨,尚未应用到实际的学术评价活动中,在信息检索和推荐方面也仅有少部分的尝试。只有增加引文文本分类在实际中的应用,才能对其理论研究加以检验并改进,反过来更快地促进应用条件的完善,最终能将引文文本的内在价值在实际的科研活动中体现出来。

参考文献:

[1] 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.

[2] LIU M X. Progress in documentation - the complexities of citation practice - a review of crtation studies[J]. Journal of documentation,1993,49(4):370 - 408.

[3] NICOLAISEN J. Citation analysis[J]. Annual review of information science and technology,2007,41:609 - 641.

[4] RITCHIE A. Citation context analysis for information retrieval[D]. Cambridge:University of Cambridge,2009.

[5] 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.

[6] 祝清松, 冷伏海. 引文内容分析方法研究综述[J]. 情报资料工作,2013(5):39 - 43.

[7] 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.

[8] 刘盛博, 丁堃, 唐德龙. 引用内容分析的理论与方法[J]. 情报理论与实践,2015(10):27 - 32.

[9] WHITE H D. Citation analysis and discourse analysis revisited[J]. Applied linguistics,2004,25(1):89 - 116.

[10] MERTON R K. The Matthew effect in science[J]. Science,1968,3810(159):59 - 63.

[11] GARFIELD E. Citation indexes for science - new dimension in documentation through association of ideas[J]. Science,1955,122(3159):108 - 111.

[12] LIPETZ B A. Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators[J]. American documentation,1965,16(2):81 - 90.

[13] FROST C O. Use of citations in literary research - preliminary classification of citation functions[J]. Library quarterly,1979,49(4):399 - 414.

[14] CRONIN B. Norms and functions in citation - the view of journal editors and referees in psychology[J]. Social science information studies,1982,2(2):65 - 78.

[15] MORAVCSIK M J, MURUGESAN P. Some results on function and quality of citations[J]. Social studies of science,1975,5(1):86 - 92.

[16] OPPENHEIM C, RENN S P. Highly cited old papers and reasons why they continue to be cited[J]. Journal of the American Society for Information Science,1978,29(5):225 - 231.

[17] COLE S. Citations and the evaluation of individual scientists[J]. Trends in biochemical dciences,1989,14(1):9 - 14.

[18] CHANG Y W. A comparison of citation contexts between natural sciences and social sciences and humanities[J]. Scientometrics,2013,96(2):535 - 553.

[19] CAMACHO - MINANO M D M, NUNEZ - NICKEL M. The multilayered nature of reference selection[J]. Journal of the American Society for Information Science and Technology,2009,60(4):754 - 777.

[20] CHUBIN D E, MOITRA S D. Content - analysis of references - adjunct or alternative to citation counting[J]. Social studies of science,1975,5(4):423 - 441.

[21] 陈晓丽. 引文类型比较分析[J]. 图书与情报,1998(4):51 - 54.

[22] PERITZ B C. A classification of citation roles for the soclal - sci-

ences and related fields[J]. Scientometrics,1983,5(5):303 - 312.

[23] COZZENS S E. Comparing the sciences - citation context analysis of papers from neuropharmacology and the sociology of science[J]. Scientometrics,1983,5(1):127 - 153.

[24] GILBERT G N. Referencing as persuasion[J]. Social studies of science,1977,7(1):113 - 122.

[25] BONZI S, SNYDER H W. Motivations for citation - a comparison of self citation and citation to others[J]. Scientometrics,1991,21(2):245 - 254.

[26] ROUSSEAU R. The gozinto theorem - using citations to determine influences on a scientific publication[J]. Scientometrics,1987,11(3/4):217 - 229.

[27] GARFIELDS E. Can citation indexing be automated? Essays of an Information Scientist[M]. Philadelphia: ISI Press,1977:84 - 90.

[28] WEINSTOCK M. Citation index[M]. Philadephia Institute for scientific Information,1971.

[29] THORNE F C. Citation index - another case of spurious validity [J]. Journal of clinical psychology,1977,33(4):1157 - 1161.

[30] SEN S K. A theoretical glance at citation process[J]. International forum on information and documentation,1990,15(1):1 - 7.

[31] VINKLER P. Comparative investigation of frequency and strength of motives toward referencing, the reference threshold model - comments on theories of citation? [J]. Scientometrics,1998,43(1):107 - 127.

[32] BROOKS T A. Private acts and public objects - an investigation of citer motivations[J]. Journal of the American Society for Information Science,1985,36(4):223 - 229.

[33] CASE D O, MILLER J B. Do bibliometricians cite differently from other scholars? [J]. Journal of the American Society for Information Science and Technology,2011,62(3):421 - 432.

[34] 邱均平, 陈晓宇, 何文静. 科研人员论文引用动机及相互影响关系研究[J]. 图书情报工作,2015,59(9):36 - 44.

[35] 马凤, 武夷山. 关于论文引用动机的问卷调查研究——以中国期刊研究界和情报学界为例[J]. 情报杂志,2009(6):9 - 14,18.

[36] MORAVCSIK M J, MURUGESAN P. Citation patterns in scientific revolutions[J]. Scientometrics,1979,1(2):161 - 169.

[37] MARTENS B V D, GOODRUM A A. The diffusion of theories: a functional approach[J]. Journal of the American Society for Information Science and Technology,2006,57(3):330 - 341.

[38] SULA C A, MILLER M. Citations, contexts, and humanistic discourse: toward automatic extraction and classification[J]. Literary and linguistic computing,2014,29(3):452 - 464.

[39] 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(3):77 - 81.

[40] MACROBERTS M H, MACROBERTS B R. Quantitative measures of communication in science - a study of the formal level[J]. Social studies of science,1986,16(1):151 - 172.

[41] BONZI S. Characteristics of a literature as predictors of relatedness between cited and citing works[J]. Journal of the American Society for Information Science,1982,33(4):208 - 216.

[42] WAN X J, LIU F. Are all literature citations equally important? Automatic citation strength estimation and its applications[J]. Journal of the Association for Information Science and Technology, 2014,65(9):1929 - 1938.

[43] MACROBERTS M H, MACROBERTS B R. Citation content analysis of a botany journal[J]. Journal of the American Society for Information Science,1997,48(3):274 - 275.

[44] TEUFEL S, SIDDHARTHAN A, TIDHAR D. Automatic classification of citation function[C]//EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22 - 23 July 2006, Sydney, Australia. Stroudsburg: Association for Computational Linguistics,2006:73 - 92.

[45] 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.

[46] BERTIN M, ATANASSOVA I, GINGRAS Y, et al. The invariant distribution of references in scientific articles[J]. Journal of the Association for Information Science and Technology,2016,67(1):164 - 177.

[47] HERLACH G. Can retrieval of information from citation indexes be simplified - multiple mention of a reference as a characteristic of link between cited and citing article[J]. Journal of the American Society for Information Science,1978,29(6):308 - 310.

[48] HALEVI G, MOED H F. The thematic and conceptual flow of disciplinary research: a citation context analysis of the journal of informetrics, 2007[J]. Journal of the American Society for Information Science and Technology,2013,64(9):1903 - 1913.

[49] 刘茜, 王健, 王剑. 基于引文动机的文献老化研究[J]. 情报探索,2015(10):1 - 4.

[50] MURUGESAN P, MORAVCSIK M J. Variation of nature of citation measures with journals and scientific speclalties[J]. Journal of the American Society for Information Science,1978,29(3):141 - 147.

[51] SWALES J. Citation analysis and discourse analysis[J]. Applied linguistics,1986,7(1):39 - 56.

[52] GARZONE M, MERCER R E. Towards an automated citation classifier[C]//13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000 Montéal, Quebec, Canada. Berlin: Springer,2000:337 - 346.

[53] NANBA H, OKUMURA M. Towards multi - paper Summarization Using Reference Information[C]//Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI 99). San Francisco: Morgan Kaufmann Publishers Inc., 1999:926 - 931.

[54] 刘盛博, 丁堃. 基于引用内容的引文评价分析[C]//中国科技政策与管理学术年会. 北京: 中国科学学与科技政策研究会, 2013.

[55] PHAM S B, HOFFMANN A. A new approach for scientific citation classification using cue phrases[J]. Advances in artificial intelligence,2003,2903:759 - 771.

[56] ZHU X D,TURNEY P, LEMIRE D,et al. Measuring academic influence:not all citations are equal[J]. Journal of the Association for Information Science and Technology,2015,66(2):408 - 427.

[57] Radoslav R. Exploring automatic citation classification[D]. Ontario:University of Waterloo, 2008.

[58] AGARWAL N K, XU Y J, POO D C C. A context - based investigation into source use by information seekers[J]. Journal of the American Society for Information Science and Technology,2011,62(6):1087 - 1104.

[59] LE M H, HO T B, NAKAMORI Y. Detecting citation types using finite - state machines[C]//10th Pacific - Asia Conference, PAKDD 2006. Singapore:Springer Berlin Heidelberg,2006: 265 - 274.

[60] WANG S M, YANG W P, CHOU H P, et al. Automatic bibliographic component extraction using conditional random fields[J]. Journal of Internet Technology,2012,13(5):737 - 747.

[61] CHRISTIAN C, NICOLA L, ALEXANDER O. The incidence and role of negative citations in science[J]. Proceedings of the National Academy of Sciences, 2015, 112(45):13823 - 13826.

[62] McCAIN K W, SALVUCCI L J. How influential is Brooks' law? a iongitudinal citation context analysis of Frederick Brooks' The Mythical Man - Month[J]. Journal of information science,2006, 32(3):277 - 295.

[63] 孙海生, 黄燕. 引用内容分析法在领域发展轨迹研究中的应用[J]. 情报探索,2015(9):56 - 60.

[64] COZZENS S E. Split citation identity - a case - study from economics[J]. Journal of the American Society for Information Science,1982,33(4):233 - 236.

[65] Giles L,Mitra P, WU J. CiteSeerX[EB/OL]. [2015 - 12 - 21]. http://citeseer. ist. psu. edu/imdex.

[66] FUJIWARA T, YAMAMOTO Y. Colil: a database and search service for citation contexts in the life sciences domain[J]. Journal of biomedical semantics, 2015, 6(1):1 - 11.

[67] SHOTTON D, PORTWIN K, KLYNE G, et al. Adventures in semantic publishing:exemplar semantic enhancements of a research article[J]. PLOS computational biology,2009,5(4):17.

[68] LIU S B, CHEN C M, DING K, et al. Literature retrieval based on citation context[J]. Scientometrics,2014,101(2):1293 - 1307.

[69] DUNCAN J. Computer - assisted production of bibliographic databases in history[J]. Indexer,1981,12(3):131 - 139.

[70] HE Q,PEI J,KIFER D,et al. Context - aware citation recommendation[C]//Proceedings of the 19th International Conference on World Wide Web, WWW 2010. Raleigh, North Carolina, USA. New York:ACM,2010:421 - 430.

[71] ZARRINKALAM F, KAHANI M. SemCiR A citation recommendation system based on a novel semantic distance measure[J]. Program - electronic library and information systems,2013,47(1):92 - 112.

作者贡献说明:

王文娟:提出论文思路,收集数据,撰写论文并修改;

马建霞:提出论文思路及修改意见,修订最终版本;

陈春:提出论文思路及修改意见;

张凌波:承担部分资料的翻译工作。

A Review of Citation Context Classifications and Implementation Methods

Wang Wenjuan[1,2] Ma Jianxia[1] Chen Chun[1] Zhang Lingbo[3]

[1]Lanzhou Library of Chinese Academy of Science, Lanzhou 730000

[2]University of Chinese Academy of Sciences, Beijing 100049

[3]Lanzhou University High School,Lanzhou 730000

Abstract: [**Purpose/significance**]This paper systematically summarizes the citation context classification standards and compares the implement methods, concludes the existing sides of application, to analyze the exiting problems and propose directions for improvement. [**Method/process**]It introduces several current points of classifying the citation context, such as citation functions, attitudes toward, citation influence, and compares the different implementations to analyze its advantages and disadvantages. [**Result/conclusion**]The current classification standards has no uniform rules, the access of all citation context is difficult and the computer processing is low accuracy and rarely no experiments of Chinese citation contexts. Future research directions are integrating classification rules, improving the accuracy of computer technology of citation context analysis, and increasing the range of applications.

Keywords: citation context classification citation content analysis citation analysis