DOI:10.13530/j.cnki.jlis.190025

基于引文内容的中文图书被引行为研究*

章成志　李　卓　赵梦圆　柳嘉昊　周清清

摘　要　从引文内容角度对图书被引行为进行分析,可改善传统依靠被引频次、专家评论等数据进行图书评价的片面性,进一步提高图书评价结果的准确性和科学性。本研究从亚马逊中文网站上选取计算机、法律、医学、文学和体育五个学科领域的中文图书,通过人工采集方式获取图书在施引文献中的引文内容,由此构建包含2 288条引文内容的数据集;然后从引用位置、引用强度、引用长度以及引用情感等方面,分析中文图书被引行为,并比较不同学科领域之间的差异。实验结果表明:不同学科领域对中文图书的引用位置具有不同的分布特征,表现出明显的学科差异;引用强度主要在1—3次,文学领域的平均引用强度最高;引用句长度一般在20—160字之间;施引作者对图书的引用情感中,超过80%表现为中性,而含有感情色彩的引用中,正面引用明显多于负面引用。图5。表5。参考文献24。

关键词　图书被引行为　引文内容　学科差异　图书评价

分类号　G251

Citing Behavior of Chinese Books Based on Citation Content

ZHANG Chengzhi, LI Zhuo, ZHAO Mengyuan, LIU Jiahao & ZHOU Qingqing

ABSTRACT

How to accurately assess book impact from massive data collection with uneven qualities is a challenging problem for libraries and researchers. Traditional methods for book impact assessment are based on citation frequency, library holdings or book reviews. Since most citation frequency based methods fail to indicate in-depth citation motivation, they cannot support comprehensive assessment of book impact. Hence, this paper used citation contexts to figure out researchers' attitudes and behaviors on citations of Chinese books, so as to improve book impact assessment. Specifically, Chinese books from five disciplines were collected: computer science, law, medicine, literature and sport science from Amazon.cn. Then we extracted citation contexts about these Chinese books from each citing literature manually and built a corpus with 2 288 citation contexts. Then, we analyzed citation behaviors over these Chinese books by mining citation locations, citation intensities, citation lengths and citation sentiments. The experimental results showed that: 1) when citing Chinese books, authors from five disciplines had different preference on citation locations. For example, literatures in computer sciences cite books more in Methodology part, while more

* 本文系国家社会科学基金重大项目"情报学学科建设与情报工作未来发展路径研究"(编号:17ZDA291)的研究成果之一。(This article is an outcome of the key project "Research on Discipline Construction of Information Science and Future Development Path of Information Work" (No.17ZDA291) supported by National Social Science Foundation of China.)

通信作者:章成志, Email: zhangcz@ njust.edu.cn, ORCID:0000-0001-8121-4796 (Correspondence should be addressed to ZHANG Chengzhi, Email: zhangcz@ njust.edu.cn, ORCID:0000-0001-8121-4796)

than half of citations in medicine were in Discussion part. While citation proportions of Introduction were similar among all disciplines. 2) Citation intensities mainly ranged from 1 to 3. More than three-quarter citation intensities were 1 in computer sciences and sport science. Meanwhile, about half of citation intensities in law, medicine and literature were 1. In addition, citations in literature had more high citation intensities. 3) The citation lengths were concentrated between 20 and 160, and showed distribution similarities in some disciplines. Overall, most citation lengths were in the "40-60" interval, accounting for about 23%. 4) Regarding citation sentiments of Chinese books, more than 80% citations were neutral. Compared with negative citations, there were more positive ones. Meanwhile, proportion of negative citations in Law was higher than other disciplines. In summary, this paper analyzed researchers' behavior of Chinese book citation in different disciplines based on detailed and reliable data. The proposed framework can provide supports for book publishers, libraries, academics and research management departments when selecting high-impact books. Meanwhile, we explored new research objects when doing citation context analysis, which can also provide references for follow-up researches. In terms of limitation, as full texts of some citing literatures cannot be collected, the built corpus has limitation in integrity and scale. Addressing citation context analysis, we only conducted frequency statistics about citation locations, lengths, intensities, more sophisticated analytical procedures should be developed in the future. 5 figs. 5 tabs. 24 refs.

KEY WORDS

Citation behavior of Chinese book.　Citation content.　Subject difference.　Book assessment.

0　引言

图书是人类用以表达观点、保存信息、传播知识的载体,也是学术研究中十分重要的资源。如何从众多质量参差不齐的图书中快速选取高影响力的作品,是图书出版商、图书馆、学术界以及科研管理部门等共同面临的一个重要问题。

传统的图书评价方法,大多依据被引频次、馆藏数量、专家书评等单一的数据源,如:苏新宁依据"通过20%的图书被引数量来反映在学界产生的80%的学术影响"的准则挑选最具有学术影响力的图书[1];White等学者通过图书的馆藏量来评价图书的影响力[2];Zuccala等学者则利用专家书评信息来评估专著的影响力[3]。

随着学术论文全文本数据获取的日益便捷,基于全文本的科学计量研究引起了很多研究者的兴趣,尤其是科学计量学与自然语言处理、机器学习、信息检索等领域的交叉研究,学术界已在全文引用分析、文本自动综述、实体主题抽取等领域开展了一系列的探索和应用。Ding等人统计分析了 *Journal of the American Society for Information Science and Technology* 中866篇学术论文的不同位置引用情况,他们将引用位置分为摘要、引言、文献综述、方法、结果、结论或未来工作等6个部分,考虑到对参考文献存在多次引用的现象,通过两种不同的方式计算引用次数[4]。胡志刚则从引文在施引文献正文中的具体引用位置、引用次数和引用语境三个方面出发,通过实证分析,揭示了引用行为的一些现象和特征[5]。当前关于引文内容的分析,大多都从引用次数、引用位置、引用情感等角度出发。其中,利用引文内容信息,可以识别作者的引用态度,挖掘图书的被引行为,从而进一步提高图书评价的准确性与科学性。然而,现有的图书评价研究没有考虑引文内容信息的影响。此外,目前引文内容的相关研究,大多以英

文文献为研究对象,针对中文文献,特别是中文图书的相关研究尚比较缺乏。

为此,本文以中文图书的施引文献为研究对象,并基于施引文献中对图书引用的引文内容,分析中文图书的被引行为。首先,以计算机、法律、医学、文学、体育五个学科领域为例,从中文亚马逊网站选取399本中文图书;然后通过人工采集方式获取图书施引文献的全文数据,并从中摘取图书在施引文献中的引文内容信息,得到2 288条引文内容数据;最后,从图书在施引文献中的引用位置、引用强度、引用长度以及引用情感等方面分析中文图书被引的行为特征,为后续的图书评价方法提供依据。

1 相关研究工作概述

引文内容信息包括被引文献在施引文献中的被引位置(简称"引用位置")、被引次数、引用内容的长度以及作者的引用态度等信息,这些信息都可用于图书的评价分析。本文从引文位置、引用强度、引文长度以及作者的引用情感等方面,对相关的研究工作进行概述。

1.1 引用位置相关研究概述

科学文献通常都有一定的章节结构,引文内容的位置即为引文内容在施引文献组织结构中的位置[6]。引用位置可以辅助人们揭示学科间的引用规律。Sombatsompop等学者将引用位置作为评价学术论文引用质量的一个重要因素[7]。他们将引用位置划分为4种类型,即"引言""实验与材料""结果与讨论""结论与其他",研究认为"结果与讨论"部分中的引用比"引言"部分的更重要[7]。Catalini等学者将文章结构分为"引言""材料与方法""结果与讨论""其他"等四个部分,通过对1998—2007年*Journal of Immunology*上的15 731篇全文中负面引用的分析,发现约84%的负面引用出现在"结果与讨论"部分[8]。Bertin和Atanassova选取PLoS七本学术期刊上发表的8万篇学术论文,基于线索词识别负面引用,结果表明约有72%的负面引用发生在"讨论"部分,其次是"结果"部分(约为14%)[9],这和Catalini等学者的研究结论比较一致。张梦莹等人在构建引文内容分析数据集时,将引用位置分为"引言""文献综述""方法""结果""讨论""结论"等六个部分[10]。他们抽取PLoS One期刊中6个不同学科2006—2015年间刊载的3 414篇学术论文,得到5 320条引文内容数据,分析发现:引文最集中的位置在"引言"部分[10]。Jurgens等学者将文章结构分为"引言""相关研究""动机""方法""评估""结果""讨论""结论"等八个部分,以分析每个部分中引用功能的分布特点,结果发现:"背景引用"功能在"引言""相关研究""动机"部分占比较多,在"方法"和"评估"部分中"使用"功能占比较大[11]。

从已有的研究可以看出,研究人员对引用位置的分类体系和分布情况有不同的研究结果。究其原因,主要包括两个方面:其一,不同研究人员对引用位置划分的标准和立论角度存在差异;其二,研究所使用的数据集在学科或规模上存在差异。本文综合考虑现有研究成果,将引用位置分为"引言""相关研究""数据""方法""实验""讨论""结论"等七个部分,并分析不同学科领域引用位置分布的差异性。

1.2 引用强度相关研究概述

一篇文献可能多次引用同一篇参考文献,同一引文内容中有可能标注多篇参考文献。参考文献的引用次数体现了不同参考文献的重要程度。胡志刚定义引文的引用强度为"引文在施引文献正文中被引用或提及的次数",他计算*Journal of Informetrics*上350篇学术论文中10 382条引文的引用强度,得到学术论文中引文篇数和引用强度呈幂律分布的规律,利用引文数和引用数的对比说明多次引用现象的普遍性[5]。Hassanl等学者随机选择自然语言处理领域的465篇学术论文,对得到的106 509条引文内容进行分析,指出,如果论文多次引用一篇参

考文献,则表明这篇参考文献对该论文更重要[12]。刘盛博和丁堃提出一个引文质量的评价指标,即:$Q_{引文}$=引文在文献中具体被引频次/引文在参考文献中出现频次,他们认为 $Q_{引文}$ 数值越大,表明这篇文章被引质量越高[13]。

本文在现有研究基础上,引入图书引用强度的概念,即为图书的被引次数与施引文献篇数的比值,用于分析不同学科领域的引用强度分布情况。

1.3　引文内容长度与引用情感相关研究

引用内容长度是指引文内容的字符串长度。探究引用内容长度的分布情况,可以展现每个学科领域的引用特色。章成志等人对 39 本英文专著的引用行为进行分析,发现引文及其上下文的平均长度主要分布在 100.00—200.00 字,较长的引文内容主要分布在专著的前半部分,而且不同学科的引文内容及其上下文的平均长度存在一定的差异[14]。

引用内容的引用情感表明施引文献作者对参考文献的引用态度。显然,支持态度和否定态度对图书影响力的评价具有不同的作用。刘盛博和丁堃将引用内容的引用情感分为正面引用、负面引用和中性引用 3 类,他们利用线索词对 *BMC_Bioinformatrics* 期刊中 147 817 条引用内容进行情感标注,结果显示 62.88%的引用为中性引用,只有 3.53%的引用是负面引用,表明研究者更倾向对所引用的文献持积极态度[13]。Athar 和 Teufel 选择 20 篇自然语言处理领域学术论文的 1 741 条引文内容,利用不同方法对引文内容的引用情感进行机器标注,结果发现:对于负面引用数量,考虑引文内容上下文比只考虑引文内容增加 3 倍,忽略引文内容上下文会丢失很多引用情感信息[15]。Abu-Jbara 等学者采用监督学习的方法,对自然语言处理领域 30 篇学术论文在施引文献中的 3 500 条引文内容及其上下文,进行情感标注,由于超过一半的引用是中性的,他们使用两阶段分类方法,第一步将引用分为主观和客观,第二步将主观引用分为正面引用和负面引用,结果发现这种分类方法更加直观[16]。Catalini 等学者对 *Journal of Immunology* 中的 15 731 篇学术论文全文进行分析,发现:4 888 篇包含负面的引用内容;15 731 篇论文对应的 762 355 条引文内容中,有18 304 条负面引文内容(占比 2.4%),对应的146 891 篇参考文献中,至少有一次被负面引用的占 7.1%[8]。

可以看出,基于引文内容及其上下文对引用情感进行标注,可以明显增加引用情感识别的准确性。此外,由于大部分引用属于中性引用,利用机器标注的方法可能会忽略占比较少的负面引用,因此,本文采取人工标注方法对引文内容的情感类别进行标注。

2　研究方法

2.1　基本思路

鉴于传统图书评价方法忽略引文内容信息的不足,本文利用中文图书在施引文献中的引文内容,分析不同学科领域图书的被引行为特征。研究思路为:首先,从在线图书电商网站采集计算机、法律、医学、文学、体育五个学科领域的图书元数据信息,然后依据学术搜索引擎采集图书的施引文献相关信息;其次,从文献全文数据库获取这五个学科领域图书的施引文献全文内容,并提取其中中文图书被引用的相关信息;接着,从引用位置、引用强度、引用长度、引用情感等方面对引文内容进行分析;最后,研究中文图书的被引行为和学科之间的差异性。具体研究流程如图 1 所示。

2.2　数据获取

本文研究的图书元数据信息和图书施引文献的元数据信息分别来源于亚马逊中文网站(https://www.amazon.cn/)和百度学术(http://xueshu.baidu.com/),在 2016 年 11 月完成数据采集工作。目前主流的中文全文数据库如知网、万方、维普等均存在部分文献未收录的情况,

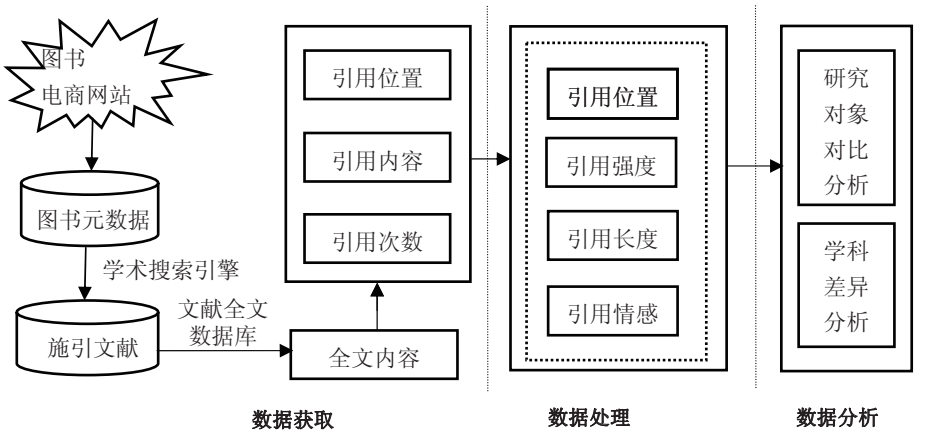

图 1　基于引文内容的中文图书被引行为研究框架

相比以单一的中文全文数据库作为检索入口，百度学术更可能覆盖所有施引文献的信息。为了尽可能地找到图书的所有施引文献信息，本研究以百度学术作为检索入口，以中文图书的元数据信息作为检索关键词，获取图书的施引文献信息。在选取图书的学科领域时，首先以亚马逊官方网站提供的中文图书商品分类的一级类目①为基础，同时考虑到一级类目与学科分类类目的匹配程度以及自然学科和人文社会学科的差异，最终以计算机、法律、文学、医学、体育五个学科领域为例，通过文献全文数据库获取图书引文内容数据集，通过以下两个数据获取的步骤，构建中文图书的引文内容语料库。

（1）根据图书在亚马逊中文网站上的评论数必须大于或等于 1、图书在百度学术上被引次数大于或等于 1 并且图书必须包含目录信息的原则，得到计算机、法律、文学、医学、体育五个学科领域的图书 6 006 种。

（2）为了保证引文内容的准确性，采用人工标注的方法获取这些图书被引的引文内容及其上下文（即引文内容的前两句和后两句）。但鉴于人工标注方法的工作量大、成本较高，且6 006 种图书的被引分布存在差异，如被引次数在[0-5]区间的较多，而被引大于 15 次的相对较少，为使获得的引文内容数据更具有代表性，根据图书被引分布，按比例（即 6 006 种图书在各被引区间的分布比例）抽取各个被引区间（在选取数据时，我们对引文的分布进行了分析，最后选取了 0—5，6—10，11—15，16—20，大于 20 作为被引区间）的图书共计 500 种。由于部分图书的被引文献全文无法获取或全文中没有参考文献标注痕迹，最终得到 399 种中文图书及其在施引文献中的引文内容与上下文。图书被引的具体分布情况如表 1 所示。

表 1　引文内容语料库中图书的领域分布情况

学科领域	图书数量（种）	被引文献数量（篇）	总被引数量（次）
计算机	69	262	284
法律	86	408	548
文学	82	395	614
医学	94	480	585
体育	68	198	257
总计	399	1 743	2 288

① https://www.amazon.cn/gp/book/all_category/ref=sv_b_1

2.3　数据标注与处理

在人工标注的基础上,依据全文数据得到引文内容的位置信息,通过引文内容语料库进一步获取引文内容的强度、长度以及引用情感等信息。具体标注方案如表 2 所示。

表 2　引文内容特征标注说明

特征属性	类别	属性描述
引用位置	引言	介绍该文的写作背景和目的的章节
	相关研究	概述与该文相关研究工作的章节
	方法	说明该文所使用方法或手段的章节
	数据	介绍该文数据来源的章节
	实验	描述该文实验过程的章节
	讨论	对该文实验结果进行解释与讨论的章节
	结论	总结该文研究结论的章节
引用强度	S≥1	图书被施引文献引用的总次数/图书的施引文献的总篇数
引用长度	—	施引文献中引文内容的字符串长度
引用情感	正面引用	引文内容反映出施引作者持有正面态度
	负面引用	引文内容反映出施引作者持有负面态度
	中性引用	引文内容不能反映施引作者的情感态度

(1)引文内容位置信息的标注

将引文内容位置(简称为引用位置)按照章节类型详细划分为“引言”“相关研究”“数据”“方法”“实验”“讨论”“结论”等七个部分。在标注过程中发现,文学和法律两个领域的施引文献中不同作者的写作风格具有较大差异,难以通过文献的章节标题来判断引用位置信息。相对而言,体育、医学、计算机三个领域的文章组织结构较为直观,作者写作风格较为接近。因此,笔者对计算机、医学和体育等三个领域的图书进行引用位置信息的标注,共得到 1 045 条包含引用位置信息的引用数据。

(2)引用强度的计算

由表 1 可以看出,每个学科领域图书的总被引数量都大于施引文献数量。有学者指出,一篇文献在施引文献中被引用的次数越多,表明其重要性越大[11-12]。为此,本文利用引用强度指标来度量给定文献的重要程度,计算公式如下:

$$S_{图书} = Q_{引用}/Q_{引文} \quad (1)$$

其中,$S_{图书}$表示图书的引用强度,$Q_{引用}$表示图书被施引文献引用的次数,$Q_{引文}$表示图书施引文献的篇数。$S_{图书}$值越大表明该图书的引用强度越高。

(3)引文内容长度计算与引用情感信息标注

笔者依据引文内容语料库,将每条引文内容的字符串长度,作为引文内容的长度。将引文内容的引用情感分为三类,即正面引用、中性引用和负面引用。采用人工标注方式得到图书被施引文献引用的引用情感信息。依据引文内容及其上下文,对引文内容的引用情感进行评断,最终给出情感类别。表 3 为情感信息标注的样例。

通过引文内容及其上下文判断施引作者对图书的引用情感,带有一定程度的主观色彩。为保证人工标注结果的准确性,由三位标注者独立完成引文内容的引用情感标注,并依据 *Kappa* 系数作为一致性评价指标[17],对三位标注者标注结果的一致性进行评估。*Kappa* 系数

表 3　引文内容的引用情感标注样例

标注类别	引文内容前的第二句	引文内容前的第一句	引文内容	引文内容后的第一句	引文内容后的第二句	领域
正面	在宪政国家,宪法规范所确认的公民基本权利,作为宪法基本价值目标和实现宪法基本价值目标的手段的统一,是宪法至上的基本内涵所在,是公正社会秩序的基础和基本保障。	它们本身具有“超级法”的规范效力,被视作法律的普遍原则的重要组成部分,约束作为民意机关的立法机关的具体立法活动。	诚如美国联邦最高法院杰克逊大法官所言:“权利法案的真正目的,是把某一些东西从政治冲突的此长彼消下解放出来,放置在一个民众之多数和官府都够不着的地方,把它们确定为法庭所依据的法律原则。”	因此,立法机关必须充分接受这些基本权利的约束,不享有和不得行使凌驾于这些基本权利之上的权力,这也是立法机关的基本宪法义务所在。	这就要求宪法应以明确的方式承认,宪法规定的基本权利可以约束立法机构。	法律
负面	由于风速的不确定性及其变化较大,导致风能转换系统是一个具有参数不确定性的系统。	根据这个缺点,有学者建立了风能转换系统线性参数变化(Linear Parameter Varying, LPV)模型。	然而,基于 LPV 模型增益调度控制的设计需要求解大量的线性矩阵不等式(Linear Matrix Inequality, LMI)组,增加了计算量。	Munteanu 等根据风速的多时间尺度特性,提出风能转换系统的双频模型。	该模型不仅可以保证控制精度,而且又避免了在求解 LPV增益调度控制器的大量计算。	计算机
中性	与社会管理结构系统相比,政府管理结构系统改革更多属于政府职能转变的范畴,而作为社会事业,体育管理体制中的社会管理结构系统改革才是当前改革的重点。	我国体育事业社会管理结构系统包括三个子机构:中华全国体育总会、中国体育科学学会和中国奥林匹克委员会。	这些社会管理机构尽管组织严密、分工细致,在相关职责范围内发挥了一定的作用,但结构性管理仍然有进一步完善的空间。	例如,中华全国体育总会将管理任务定为“指导国家业余体育运动”,其具体任务由成立之初的八项缩减为目前的三项。	国家体育事业日新月异的实践发展经验表明,中华全国体育总会变更其结构性管理内涵、缩短其管理任务外延,却令其遭遇尴尬,例如管理内涵与其他机构重合,管理外延又覆盖不到其应涉及的领域。	体育

计算公式如下:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \qquad (2)$$

其中,$P(A)$代表标注结果一致性的实际观测值,$P(E)$表示标注结果一致性的期望值。若 $K \geqslant 0.8$,则说明标注结果很可靠,若 $0.8>K>0.67$,则表明标注结果较为可靠[18]。对三位标注者标注结果两两之间进行 *Kappa* 值计算,每个学科领域最高的一致性结果分别为:计算机 0.633,文学 0.726,法律 0.723,医学 0.827,体育 0.805。可

以看出,不同学科领域标注的一致性结果不同,高于可信一致性标准 $K = 0.69$ 的有四个领域,而计算机领域一致性只有 0.633,这可能是由计算机领域引文内容数量较少导致的。因为 *Kappa* 系数对于标注频率低的对象更为敏感,所以少量标注结果不一致也会降低标注结果的一致性[19]。

本文选取每个学科领域中 *Kappa* 值最高的两名标注者的标注结果,对不一致的标注进行共同讨论分析,最终得到引文内容的引用情感标注数据。

3　结果分析

3.1　引用位置的结果分析

笔者对包含 1 045 条引用位置信息的引用数据进行统计,得到不同领域中文图书在其施引文献中的引用位置分布情况,结果如图 2 所示。计算机领域 43.97%的引文内容出现在“方法”部分,其次是“引言”和“相关研究”部分。医学领域“讨论”部分的引文内容占比达到 51.42%,出现在“引言”部分占比为 21.89%。

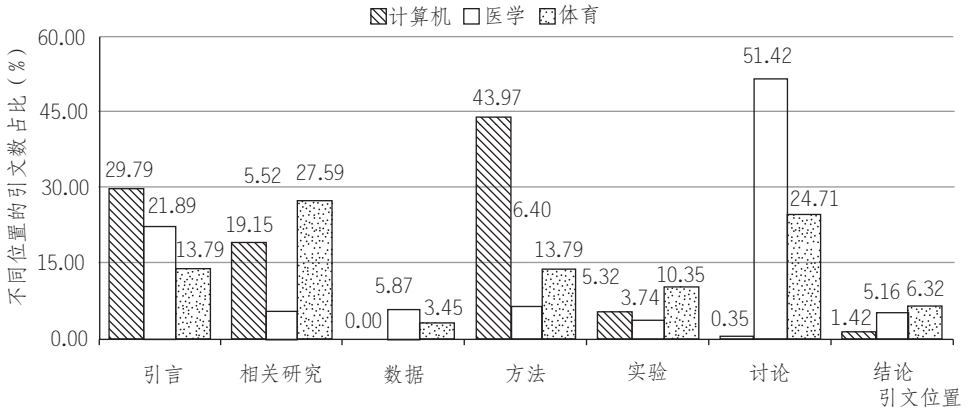

图 2　不同领域的图书被引位置分布

而在体育领域,引文内容在“相关研究”和“讨论”部分的比重较大,均超过 20%。

从总体的引用位置分布情况来看,“引言”部分包含的引文内容占全部引文内容的比例相对较高,而在“数据”“实验”“结论”等三个章节,每个领域的占比均较低。由于不同章节位置所表达的引用功能不同,有的章节偏向研究背景描述或方法论研究,而有的章节注重实证分析,因此造成引文内容在不同章节分布的不均衡。另外,还可看出不同学科的“讨论”部分的引文内容占比差异较大,尤其是医学和计算机这两个领域的差异尤为突出。

3.2　引用强度的结果分析

笔者对不同学科领域中文图书的引用强度进行统计,并将其划分为 4 个区间,结果如图 3 所示,其中“1”表示图书在施引文献中平均被引 1 次,而“1—2”则表示引用强度大于 1 小于等于 2,以此类推。

由图 3 可以看出:计算机领域引用强度为 1 的达到 91.30%,体育领域达到 77.94%,说明这两个领域的图书在施引文献中绝大部分只被引用 1 次;而在法律和医学领域,引用强度在 1—2 之间相对较多。相比较而言,文学领域出现较多高引用强度的现象。

3.3　引文内容长度的结果分析

对不同领域中文图书的引文内容长度,按照从高到低的顺序排序,并对每篇论文进行编号,得到图书被引长度分布情况。五个领域的 399

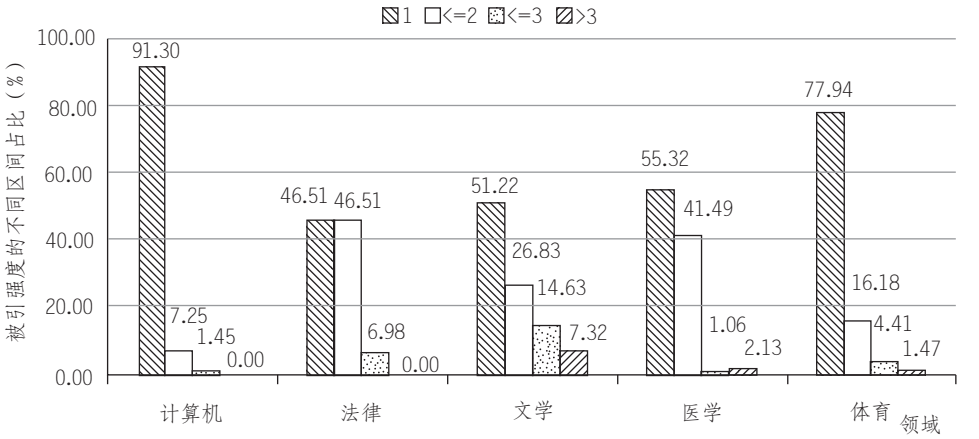

图 3　不同领域的图书被引强度分布

本图书对应的全部引文内容中，仅有 1.28% 的引文内容长度分布在 360—600 字之间，为了使图表的对比结果更加直观，图 4 仅选取 360 字以内的引文内容长度进行展示。针对不同领域的引文数量不对等问题，在图 4 结果的基础上，将每个领域的引用长度按区间进行比例计算，结果如图 5 所示（图中横坐标 20 表示区间 (0,20]，以此类推）。

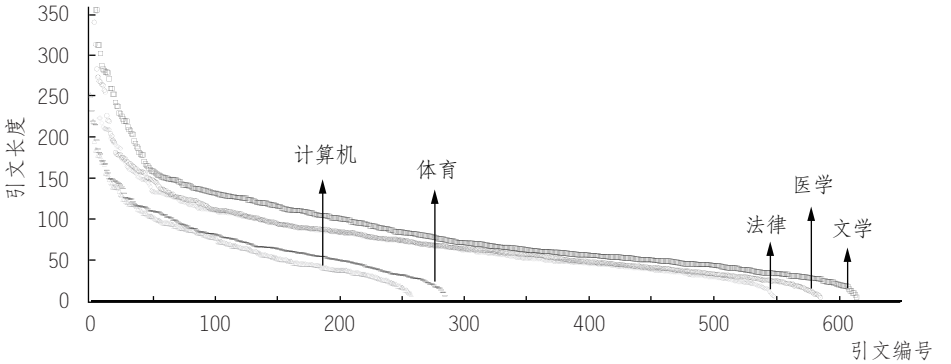

图 4　不同领域的图书被引长度分布

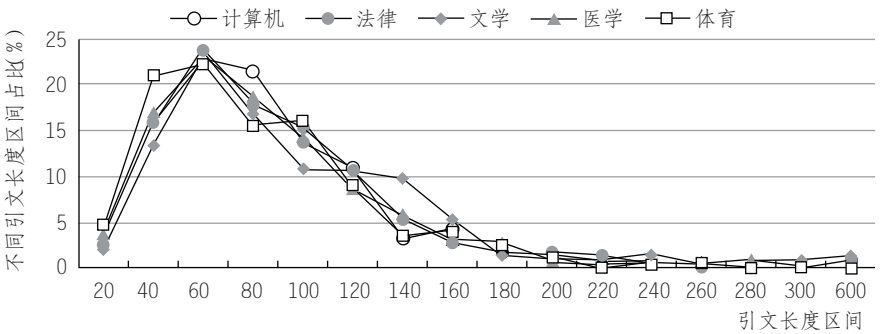

图 5　不同领域的图书被引长度区间分布

由图4可以看出:法律与医学、计算机与体育,两两之间的引文内容长度较为相近,而文学领域的引文长度相比其他领域则更大。图5显示五个领域在"40—60"字区间内均达到峰值,占比为23%左右,超过峰值后,引文内容数量逐渐下降,整体的趋势显示绝大部分引文内容的长度分布在"20—160"字的区间内。

3.4　引文内容引用情感的结果分析

本研究对引文内容引用情感结果统计发现:不同情感类别的数据比例存在较大差异(绝大部分引用情感为中性),这与Abu-Jbara等学者研究中的情感类别分布情况较为相似,因此本文借鉴Abu-Jbara等学者对引用极性的分析方法[16],对引文内容引用情感按照两步分类的方法展开分析。首先,对中性引用和带有情感的引用内容进行统计分析,结果如表4所示;其次,将带有情感的引用内容进一步细分为正面和负面引用,然后进行统计分析,结果如表5所示。

表4　情感标注比例分布(%)

领域	计算机	法律	文学	医学	体育
中性比例	84.2	83.6	84.9	85.1	93.4
情感比例	15.8	16.4	15.1	14.9	6.6

表5　正面情感与负面情感比例(%)

领域	计算机	法律	文学	医学	体育
正面比例	91.1	76.7	91.4	87.4	94.1
负面比例	8.9	23.3	8.6	12.6	5.9

体育领域中性比例相对较高,达到93.4%,其他领域带有感情色彩的引文数据仅占15%左右,这与Athar针对自然语言处理领域310篇论文的引文内容统计结果较为接近(占比14%)[20]。从表5可以看出,计算机、文学、医学和体育领域,引文内容引用情感正面的比例约为90%,法律领域负面引用达到23.3%。总体而言,作者在引用中文图书时更倾向于表达正面情感。

4　讨论

图书作为学术研究中重要的交流资源,研究其被引行为有重要的理论与现实意义。对图书的被引行为进行研究,不仅可以用于评价图书的学术影响力,还可以探索不同学科被引行为的差异性,从而为不同学科的图书评价实践提供参考依据。当前,关于学术论文中引文内容的相关研究已引起学术界的广泛关注[4,5,13,19]。然而,目前学术界关于引文内容分析的研究,尚未区分引文内容对应的参考文献类型(如期刊论文、会议论文、图书等),笔者将这些参考文献统称为非特定类型参考文献。本文针对图书与非特定类型参考文献在施引文献中的引文内容,进行特征分布差异的分析。

4.1　图书与非特定类型参考文献被引内容的特征分布差异

(1)引用位置

从图书在施引文献中出现的位置分布情况来看,计算机、医学及体育三个学科领域在"引言"部分占比相对较高(分别为29.79%、21.89%、13.79%)。张梦莹等对PLoS One期刊中6个不同学科的5 320条引文内容数据进行位置统计,结果表明位于"引言"部分的引文共2 318条,占比43.57%[10]。这表明,作者在"引言"部分,都较多地引用参考文献,但图书与非特定类型参考文献的引用位置存在一定的差异。Sombatsompop等学者对300篇临床医学领域论文的引用位置进行标注,他们发现51.8%的引用出现在"结果和讨论"部分,其次为"引言"部分[7](占比42.7%,实际上,Sombatsompop在该文中对"引言"部分的定义涵盖本文所标注的"引言"与"相关研究"部分)。本文对94种医学领域图书被引的统计结果表明,引用位置占比最高的为"讨论"部分(占比51.42%),其次为"引言"与"相关工作"部分(占比27.41%)。这表明,在医学领域,图书与非特定类型参考文献

在被引文献中引用位置方面,具有一定程度上的相似性。

(2)引用强度

胡志刚统计分析 350 篇 JOI 期刊论文的引用强度,结果表明:论文的引用强度分布呈负幂律分布,期刊论文中平均每篇引文的引用强度为 1.5 左右[5]。图书被施引文献引用的强度大多数分布在 1—2 之间,而高引用强度的分布,不同学科存在着较为明显的差别。章成志等以生物医学与生命科学、工程与计算机科学两个大类的 39 种学术专著全文中包含的引文内容为数据源,从引文内容特征角度研究学术专著引用行为,发现多数学术专著中,引文内容的引用次数和参考文献的比值集中在 1—1.5 之间[14],这与本文的研究结果较为接近。

(3)引用长度

通过引文内容长度的统计分析,发现五个学科领域图书对应的引文内容长度集中在 20—160 之间。虽然不同学科作者写作习惯不同,但学科之间的引文内容长度的分布比较接近。卢超等针对 600 篇图书情报领域的学术论文引文内容数据,进行引文内容长度的统计,结果发现学术论文的平均引文长度主要分布在 50—200 之间[21],这一结果与本文研究结果较为相近。

(4)引用情感

根据图书对应引用情感的统计结果可知,对中文图书引用的情感中超过 83% 的为“中性引用”;在带有情感的引用中,施引作者更倾向于“正面引用”。刘盛博等对 *BMC_Bioinformatrics* 期刊中 147 817 条引用内容进行分析,发现 62.88% 的引用没有表现出明显的情感倾向,被标注为“中性引用”,“正面引用”与“负面引用”分别为 33.59% 和 3.53%[13]。陆伟等对“主题模型”相关领域的 20 篇文章的 673 条引文记录进行标注,发现 96.14% 的引用为中性,“正面引用”与“负面引用”分别为 1.49% 和 2.38%[19]。张梦莹等对 PLoS One 中 5 320 条引文内容数据进行情感标注,结果表明 96.64% 的引文内容对应的情感分析结果为中性,“正面引用”与“负面引用”占比分别为 2.12% 和 1.24%[10]。这表明,在引文内容的引用情感方面,图书与非特定类型参考文献存在共同之处,即绝大部分的引用为“中性引用”。多数研究结果表明,在带情感的引用中,图书与非特定类型参考文献相似,都倾向于“正面引用”。

综上,图书与非特定类型参考文献在被引文献中引文内容特征分布虽然存在细微差异,但整体分布较为接近。当然,由于本文调研数据仅限于计算机、文学、法律、医学、体育五个领域的 399 种中文图书,选取样本数据的数据量、学科领域、语言因素都可能会导致结论上的部分差异,因此以上分析结果可能存在一定的局限性。由于本文的理论和方法不依赖于特定语种、特定学科或特定文献类型,所以也适用于其他语种或其他学科的图书、学术论文等文献类型。因此,今后调研还需在领域或学科类别、语言种类、图书数量等方面进一步扩展或增加。

4.2 图书被引行为在不同学科领域的差异

(1)引用位置

不同学科领域的引用位置分布存在差异,这表明每个学科中文图书对应的引用位置分布有各自明显的倾向。与“方法”“讨论”两部分相比,各领域在“引言”部分占比相对较高,且差异相对较小。计算机领域更多在“方法”部分引用图书。计算机科学是系统研究信息与计算的理论基础及其在计算机系统中实现与应用的学科[22],因此,计算机学科更偏向引用图书中相关的方法、模型或算法等。而医学领域有超过半数的引用集中在“讨论”部分,有研究表明医学论文的“讨论”是论文的重要组成部分[23],这与 Bertin 和 Atanassova 针对 80 000 篇生物与医学等领域学术论文的引文内容的统计结果[9]较为相近,说明相同学科领域的位置分布较为一致。

(2)引用强度

计算机与体育领域均有超过 3/4 的引用强度值为 1,而在法律、文学与医学领域,引用强度

为1的引用数量占比在50%左右。引用强度值在1—2的,法律和医学领域相对较多,达到40%左右。相比而言,文学领域高引用强度比重较大,这与文学领域施引作者倾向于引用较多的图书有关,由此导致图书在被引论文中的引用次数比其他领域更多。

(3)引用长度

本文研究发现法律与医学领域、计算机和体育领域,两两之间的引文内容长度分布比较接近,而文学领域的引用长度明显偏高。相对而言,法律和医学这两个领域,会出现较多的案例分析,在引用内容中会占用较多篇幅。在文学领域,施引作者倾向引用图书中的相关片段进行详细解读,因此引文内容长度较其他领域更长。

(4)引用情感

虽然每个学科领域大部分的引用情感为"中性引用",但体育领域的"中性引用"达到93.4%,明显比其他领域略高。在带有情感的引用中,每个学科对图书的引用都偏向"正面引用",而法律领域的"负面引用"占比相对较大,导致这种现象的原因可能为:在法律领域,施引作者常通过对各种案例的研究分析,总结经验教训、发现问题与不足,因此在与不同作者的观点进行对比时,更容易表现出否定的意见,从而产生负面引用。

综上,中文图书被引行为在不同学科领域存在明显的差异。这也表明,在中文图书被引行为基础上的相关研究,如图书的科学评价,也应考虑不同学科或领域之间的差异。如前文所述,本文调研数据仅限于五个领域的中文图书,图书被引行为的学科领域差异分析结果亦可能存在局限性。随着开放获取更加容易,学术图书将更多地被下载、讨论和引用[24],因此在今后拟扩展更多的学科或领域,获取规模更大的图书被引行为方面的数据,从而得到更加全面的学科差异分析结果。

5　结语

本文以中文图书的施引文献为研究对象,从引文内容的引用位置、引用强度、引用长度以及引用情感等角度,研究中文图书被引行为,并分析不同学科领域之间的差异性。统计分析结果表明:中文图书对应的引文内容中,"引言"和"相关研究"部分的引文所占比率较高,医学注重在"讨论"部分的引用,而计算机注重在"方法"部分的引用;中文图书对应的引文内容长度集中分布在20—160字之间,不同学科领域分布有一定的相似性;中文图书被引的强度主要在3次以内;中文图书的被引情感主要为中性,在对图书带有感情色彩的引用中,施引作者更倾向于表达正面情感。

当然,中文图书被引行为的研究还存在一些不足之处。在数据采集的过程,由于部分施引文献无法获取全文,在语料库构建的完整性与规模上存在一定的局限。此外,在引文内容的数据分析方面,目前仅对引文内容的位置、长度、引用强度进行频次统计,分析方法较为浅层。在后续的研究中,将扩大领域或学科类别、增加图书数量,借助机器学习、自然语言处理等技术,结合科学计量理论与方法,进一步深入分析引文内容的情感、引用功能、引用动机等语义层面的引文信息,对图书的被引行为进行更加广泛而深入的探索,以期为图书的科学评价提供参考。

参考文献

[1] 苏新宁. 我国人文社会科学图书被引概况分析——基于CSSCI数据库[J]. 东岳论丛,2009,30(7):5-13.
(Su Xinning. Analysis of the citation of humanities and social sciences books in China——based on CSSCI database[J]. Dongyue Tribune, 2009,30(7):5-13.)

[2] White H D, Boell S K, Yu H R, et al. Libcitations: a measure for comparative assessment of book publications in

the humanities and social sciences[J]. Journal of the Association for Information Science & Technology, 2009, 60(6):1083-1096.

[3] Zuccala A, van Someren M, van Bellen M. A machine-learning approach to coding book reviews as quality indicators: toward a theory of mega citation[J]. Journal of the Association for Information Science & Technology, 2014,65(11):2248-2260.

[4] Ding Y,Zhang G, Tamy C, 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.

[5] 胡志刚. 全文引文分析方法与应用[D].大连:大连理工大学,2014. (Hu Zhigang. Full-text citation analysis and applications [D]. Dalian: Dalian University of Technology, 2014.)

[6] 王文娟,马建霞,陈春,等. 引文文本分类与实现方法研究综述[J]. 图书情报工作,2016,60(6):118-127. (Wang Wenjuan, Ma Jianxia, Chen Chun, et al. A review of citation context classifications and implementation methods[J]. Library and Information Service,2016,60(6):118-127.)

[7] 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.

[8] Catalini C, Lacetera L, Oettl A. The incidence and role of negative citations in science[J]. Proceedings of the National Academy of Sciences, 2015,112(45):13823-13826.

[9] Bertin M, Atanassova I. Weak links and strong meaning: the complex phenomenon of negational citations[C]// Proceedings of BIR 2016 Workshop on Bibliometric-enhanced Information Retrieval. Newark, New Jersey, USA, 2016:14-25.

[10] 张梦莹,卢超,郑茹佳,等. 用于引文内容分析的标准化数据集构建[J]. 图书馆论坛,2016,36(8):48-53. (Zhang Mengying, Lu Chao, Zheng Rujia, et al.Construction and analysis of the standard data set of citation content[J]. Library Tribune, 2016,36(8):48-53.)

[11] Jurgens D, Kumar S, Hoover R, et al. Measuring the evolution of a scientific field through citation frames[J]. Transactions of the Association for Computational Linguistics, 2018,(6):391-406.

[12] Hassan S, Akram A, Haddawy P. Identifying important citations using contextual information from full text [C]// Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL2017). Toronto, Ontario, Canada, 2017:41-48.

[13] 刘盛博,丁堃. 基于引用内容的引文评价分析[C]//第九届中国科技政策与管理学术年会论文集,2013:1-7. (Liu Shengbo, Ding Kun. Citation evaluation analysis based on citation context[C]// Proceedings of the 9th Annual Meeting of Chinese Association for Science of Science and S&T Policy,2013:1-7.)

[14] 章成志,王玉琢,卢超. 学术专著引用行为研究——基于引文内容特征分析的视角[J]. 情报学报,2017,36(3):319-330. (Zhang Chengzhi, Wang Yuzhuo, Lu Chao. Citing behavior of academic monographs:perspective based on character analysis of citation content [J]. Journal of the China Society for Scientific and Technical Information, 2017,36(3):319-330.)

[15] Athar A,Teufel S. Context-enhanced citation sentiment detection [C]// Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2012). Montreal, Canada, 2012:597-601.

[16] Abu-Jbara A,Ezra J, Radev D. Purpose and polarity of citation: towards NLP-based bibliometrics [C]//Proceed-

ings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013). Atlanta, Georgia, 2013: 596-606.

[17] Warrens M J. Chance-corrected measures for 2×2 tables that coincide with weighted Kappa[J]. The British Journal of Mathematical and Statistical Psychology, 2011,64(2):355-365.

[18] Carletta J. Assessing Agreement on classification tasks: the kappa statistic[J]. Computational Linguistics, 1996, 22(21):249-254.

[19] 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):93-104. (Lu Wei, Meng Rui, Liu Xingbang. A deep scientific literature mining-oriented framework for citation content annotation[J]. Journal of Library Science in China, 2014,40(6):93-104.)

[20] Athar A. Sentiment analysis of citations using sentence structure-based features[C]//Proceedings of the Conference of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA, 2011:81-87.

[21] 卢超,章成志. 基于引文内容的单篇学术论文参考文献网络结构研究[J]. 现代图书情报技术,2014,30(10):33-41. (Lu Chao, Zhang Chengzhi. Study on the reference network of single academic article based on citation content [J]. New Technology of Library and Information Service, 2014,30(10):33-41.)

[22] 张艺蔓,马秀峰,程结晶. 融合引文内容和全文本引文分析的知识流动研究[J]. 情报杂志,2015,34(11):50-54, 49.(Zhang Yiman, Ma Xiufeng, Cheng Jiejing. Research of knowledge flows based on citation content analysis[J]. Journal of Intelligence, 2015,34(11):50-54, 49.)

[23] 《临床合理用药杂志》编辑部. 医学论文的基础结构[J].临床合理用药杂志,2009,3(1):37.(Editorial Office of *Chinese Journal of Clinical Rational Drug Use*. Basic structure of articles in medical domain[J]. Chinese Journal of Clinical Rational Drug Use, 2009,3(1):37.)

[24] Emery C, Lucraft M, Morka A, et al. The OA effect: how does open access affect the usage of scholarly books? [R], 2017:11.

章成志　南京理工大学经济管理学院教授,博士生导师。江苏 南京 210094。
李　卓　南京理工大学经济管理学院本科生。江苏 南京 210094。
赵梦圆　南京理工大学经济管理学院本科生。江苏 南京 210094。
柳嘉昊　南京理工大学经济管理学院本科生。江苏 南京 210094。
周清清　南京理工大学经济管理学院博士研究生。江苏 南京 210094。

(收稿日期:2018-08-01;修回日期:2019-01-07)