·理论研究·

基于非参数检验的文献漏引行为表现特征分析

杨文霞[1,2],邓三鸿[1,2],杨　杰[1,2],张　琪[1,2],王　昊[1,2]

(1.南京大学 信息管理学院,江苏 南京 210023; 2.江苏省数据工程和知识服务重点实验室,江苏 南京 210023)

摘　要:【目的/意义】文献引用中普遍存在漏引现象,分析文献漏引行为的表现特征,有助于揭示不合理引用行为背后的引用动机与情境,避免漏引对学术论文价值造成的不良影响。【方法/过程】构建文献漏引概念模型,以Web of Science核心合集数据库中自然科学、社会科学和人文科学共15个学科的抽样文献作为数据来源,通过文本相似度与引用耦合强度确定可供分析的数据,利用非参数检验的方法从文献题录、文献内容、作者和期刊四个方面探索漏引行为表现特征。【结果/结论】结果显示,漏引行为在语言、学科、文献类型、被引频次、作者自引、作者H指数、期刊影响因子上表现特征明显,在高被引作者上表现特征则视领域而定,在文本内容相似度上表现特征并不明显,并推测表现明显的特征可能一定程度上影响了潜在漏引行为。【创新/局限】本文定量捕捉漏引行为的表现特征,有助于探索定量化研究文献漏引的工作,为基于大规模文献数据的文献漏引识别提供了参考。

关键词:文献漏引;引用耦合强度;文本相似度;非参数检验;引用特征

中图分类号:G252　　**DOI**:10.13833/j.issn.1007-7634.2024.07.009

0 引　言

科学的发展与创新并非一蹴而就,而是建立在一代代学者的知识积累之上。新的知识往往自以前的知识产生,体现在文献上就是引用与被引用关系[1-2]。在科学计量学领域,引文关系为探索科学发展中知识的流动提供了直接的证据[1-2],随着研究的发展,探索引文关系的研究还发现了一些不规范的引用现象,如文献漏引但在探索文献间的引用关系时,也会出现各种问题,如文献漏引[3]。尽管现有研究对漏引的定义尚未统一,但对其本质的描述基本一致,即主要关注一篇文献的内容参考了某篇或多篇文献内容,却并未在参考文献中将其列出的行为。

Coles JR夫妇[4]早在1972年便指出漏引问题,有学者认为无意中疏忽和没必要引用共知的理论与思想会导致漏引[3,5]。《学术出版规范——期刊学术不端行为界定》中提到了漏引问题[6],中科院则将漏引归为学术不端[7]。一些学者利用现有文献,发现了目前漏引现象不容忽视。比如,Donner P研究发现有29.3%的论文存在一个或多个参考文献漏引[8],Jiang ZR等[9]发现在ACM数据库中,18.5%的出版物没有任何引用数据,55.6%的出版物没有任何被引数据。引文链接的准确性和完整性影响数据质量[10]。因此,文献漏引不但容易造成文献数据和观点失真,也会对学术论文价值的客观评估造成不良影响,使得被漏引文献学术价值难以被捕捉与量化,从而降低引文分析与科学评价的真实性与准确度。同时,漏引行为也是对潜在被漏引文献中作者知识贡献的忽视,不利于知识的传播。

漏引是应引未引,并不存在实际上的引用关系,因此很难从客观发生的文献数据集中发现这一现象。但是从海量文献数据中识别漏引行为,有助于学术论文价值的正确评价,避免不规范的引用行为。现有研究主要是通过人工识别[11-17]、举例[18-20]、分析现存参考文献链接[9,21]、耦合和共引[22-23]等方法识别漏引行为;也有研究两篇文章之间的参考文献重合的数量(耦合数量)测度两篇文章之间文本内容相似度,进而应用于漏引行为的识别[24]。但是,以上方法仍存在一些不足:①人工识别和案例分析方法不能大规模应用;②现有引文分析的研究大多只针对论文中的参考文献,无法捕捉参考文献之外的应引而未引的文献;③耦合和共引方法存在识别准确率低的潜在风险;④通过耦合数量衡量相似度进而识别漏引,虽然进一步改进了耦合和共引方法,但是直接测度两篇文章之间的相似度,然后与耦合数量相结合,从而基于文本相似度和耦合数量两个维度共同识别文献漏引的方法更为精确、有效。因此,为了弥补以上研究的不足,本文提出了一种二维文献漏引识别方法。通过构建文献漏引

收稿日期:2023-02-27

基金项目:国家社科基金项目"大数据环境下学术成果真实价值与影响的实时预测及长期评价研究"(19BTQ062);江苏省研究生科研创新计划项目"基于预训练模型的情报学领域术语抽取与应用研究"(KYCX22_0076)。

作者简介:杨文霞(1994-),女,山东聊城人,博士研究生,主要从事信息分析与科学评价研究;邓三鸿(1975-),男,江苏南京人,博士,教授,主要从事信息分析与科学评价研究;杨杰(2000-),男,四川广元人,博士研究生,主要从事网络分析与科学计量研究;张琪(1995-),女,山西晋城人,博士研究生,主要从事信息分析与数字人文研究;王昊(1981-),男,江苏南京人,博士,教授,主要从事信息分析与数字人文研究。

中国知网　https://www.cnki.net

概念模型，从耦合强度（Coupling Strength，CS）和文本内容相似度两个维度来获取文献漏引数据，分析文献漏引行为的表现特征，以期为基于大规模文献数据的文献漏引识别提供一些参考。

1 文献漏引相关研究

1.1 文献漏引原因相关研究

相对于国外学者注重于漏引行为的识别，国内学者更注重漏引原因及治理。大多学者认为作者主观原因为主要因素，戴泳等[25]、苏广利等[3]、陈敏[5]、郑美莺[26]认为主要是由于作者主观因素导致了漏引，如怕暴露抄袭行为，显示自己的创新性，获取私利，没必要引用共知的理论与思想，无意中疏忽等；唐虹等[27]认为刻意漏引包含着深度的学术不端，需要编辑、审稿专家和主编共同深入筛查；朱大明[28]则从审稿专家的角度着重分析了参考文献引证的鉴审，提出要注重引证的实效，避免不合理的引用行为；叶新明[29]和郭玲等[30]认为应当从学术、引文规范、编辑和作者等方面分析原因；徐英杰等[31]则进行实证分析，从经济学角度对比分析了漏引原因，认为作者通常在评估“用”且“引”和“用”而“不引”两种行为之后，为了预期效用最大而选择了前者。

1.2 文献漏引定义相关研究

虽然漏引现象由来已久，漏引原因也被广泛讨论，但对于漏引并没有统一的定义，马兰等[32]、陈敏[5]、MacRoberts M H[18]认为漏引是指文献内容涉及到了某篇或多篇文献内容，但却没有在文后参考文献中明确将其列出的行为；王志标[33]认为漏引指作者在论文中引用了他人的观点、表述、数据、图表、方法、公式等却未注明来源的行为；李锋[34]认为漏引是指作者出于某种目的在撰写论文时对为自己研究提供了重要事实、数据、研究方法的文献故意在参考文献中不列出来的行为；许雪梅[35]认为漏引是指漏列实际引用过参考文献的行为；王华菊[36]认为漏引是引用过别人的研究成果或调查结果，却未给出参考文献的行为。

以前的研究虽然对于漏引的定义并未统一，但其本质基本一致。因此本文对漏引的定义为：当一篇文献的内容直接或间接的受到一篇或多篇文献的影响，却并未在参考文献列表中将其列出的行为。对于如何确定该篇文献是否受到一篇或多篇文献的影响，本文受 HuangM H 等[22]、Chen D Z 等[23]和 Ciotti V 等[24]的方法的启发，提出“耦合强度”和“文本相似度”相结合的方法确定是否发生漏引行为。

1.3 文献漏引特征相关研究

有些研究中，由于时间久远以及数据库访问权限等因素的影响，作者可能获取不到文献全文，导致理解偏差，导致遗漏相关文献，导致漏引。也有些文献阅读过程中，作者受到语言阅读障碍、跨学科知识壁垒的影响，且有惰性心理，倾向于引用母语的文章以及本学科内的文献，Bookstein A&Yitzhaki M[37]的研究结果证实了该观点，即学者们更喜欢引用自己的母语文章；Shu F[38]通过研究中文引用偏好也得出了相同结论；MacRoberts M H[18]在讨论漏引的同时，同时指出文献的引用与否一般受到自引、出版物类型、时间和学科专业的影响；Tagliacozzo R[39]也指出在同一主题或背景下，相对于其他人的文献，作者更倾向于引用自己的成果。

此外，Teixeira D A[40]认为受到马太效应的影响，在相同研究背景下，作者由于崇拜学术权威和学术名望的心理，会更倾向于引用影响力更高、作者更权威、期刊影响因子更高的文章，以期获得认可。那么在相似背景下，作者引用了影响力更高、作者更权威、期刊影响因子更高的文献，势必会有影响力较低、期刊影响因子较低以及作者权威性较低的文献会被漏引，因此，漏引行为可能会在文献的影响力、所在期刊的影响因子以及作者的学术影响力的表现上存在差异。

因此，结合以上研究基础，本文将从文献题录、文献内容、作者和期刊四个方面分析漏引行为表现特征，具体指标为：语言、文献类型、学科、被引频次、文本内容相似度、作者自引、高被引作者、作者H指数和期刊影响因子等，以期为不规范的引用行为治理提供解决途径。

2 文献漏引概念模型与研究设计

2.1 文献漏引概念模型

本研究在现有研究基础之上，提出文献漏引概念模型，具体如图1所示。实线表示引用关系真实发生，虚线表示引用关系应当发生但未发生，即产生了漏引。引证文献C引用了焦点文献F，焦点文献F引用了参考文献R，引证文献C通过焦点文献F对参考文献R产生间接引用，但R未在引证文献C的参考文献列表中出现，即引证文献C漏引了参考文献R。

基于文献漏引概念模型，本文以文献C引用文献F而漏引文献R为切入点，探究文献R和文献F之间是否存在明显差异，并推测这些差异性可能导致文献C引用文献F而漏引文献R。首先，以某篇文章为焦点文献，得到其参考文献和引证文献；其次，对参考文献和引证文献进行两两配对；之后，结合耦合强度和引文内容相似度抽取文献C引用文献F而漏引文献R的引用关系；最后，运用非参数检验的方法分析这些引用关系在文献题录、文献内容、作者和期刊中的差异性，探索这些差异性是否影响潜在引用行为。

2.2 研究设计

图2为本研究的研究流程。其中，数据获取、处理、分析分为四步：数据获取、数据预处理、目标数据集构建、数据分析。其中R-C表示文献C漏引文献R，F-C表示文献C引用文献F，R-F表示文献F引用文献R。

数据获取过程中，根据王忠武[41]和郑文涛[42]对自然科

中国知网 https://www.cnki.net

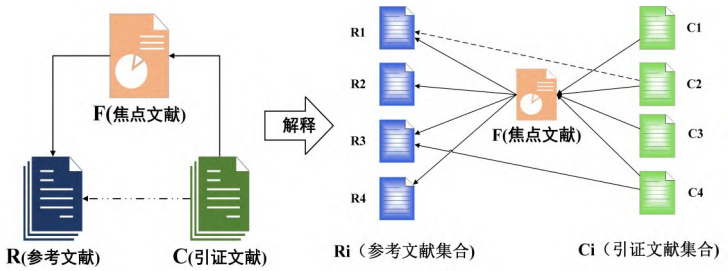


图1 文献漏引概念模型

Figure 1 Conceptual model of literature missing citations

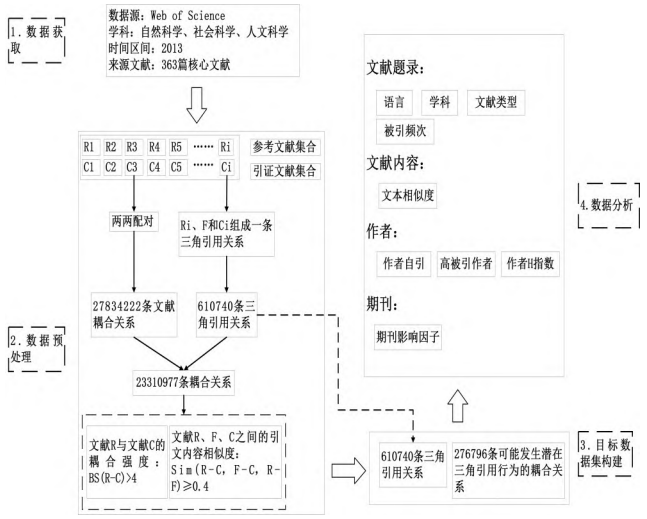


图2 研究流程图

Figure 2 Research process

学、社会科学、人文科学的定义,本文以Web of Science核心合集作为数据来源;根据WoS学科分类体系,选取法学、社会学、经济学、管理学、文学、语言、艺术、历史、哲学、数学、物理学、化学、生物学、天文学、地球科学15个学科,并根据学科文献量的大小和被引频次的高低分层抽样;其次,为保证数据样本多样性,文献类型同时包含Article、Review、Proceedings Paper;最后,共获得2013年的363篇样本文献作为焦点文献,共获取610 740条实际引用关系与27 834 222条可能发生漏引行为的引用关系数据。数据获取时间为2022年10月4日到2022年10月7日。

数据预处理中,首先,获取焦点文献F的参考文献R和引证文献C,对文献R和文献C两两配对,共27 834 222条文献耦合关系;其次,获取以文献F为起始文献的610 740条实际引用关系数据;之后,在已获得的耦合数据集中删除发生实际引用关系的数据;最后,保证每组可能发生漏引行为的数据中,文献R的发表时间要早于文献C,最终得到可供分析的文献数据共23 310 977条。

然后,构建两项指标测度可能发生的漏引。一方面,判断文献R、F和C之间的文本相似度,在引用结构中,R-C,F-C和R-F之间引文内容相似度越高,意味着在很大程度上文献C参考了文献F中关于R的引文内容,通过比较引文内容相似度的高低判断文献C是否存在漏引行为,相似度越高,则越倾向于发生了漏引行为。另一方面,耦合强度[15-17]指文献R和文献C参考文献重复的数量。重复数量越多,意味着尽管文献R与C之间没有实际引用关系,但两者具有密切的潜在关系,文献C在一定概率上参考了文献F中较多的参考文献。耦合强度越高,则越倾向于发生了漏引行为。根据耦合强度CS(R-C)、文本相似度Sim(R-C,F-C,R-F)的计算结果构建气泡图,见图3。

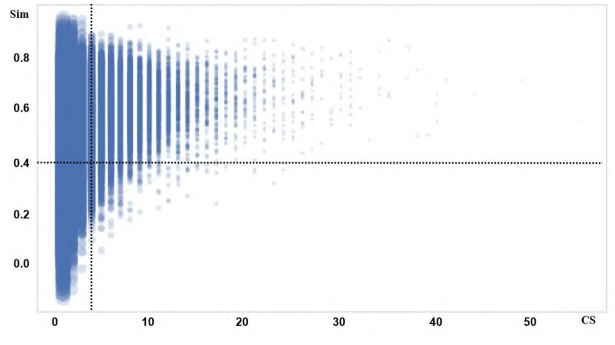


图3 耦合强度与引文内容相似度分布气泡图

Figure 3 Bubble diagram of coupling strength and citation content similarity

其中,横坐标表示耦合强度值,纵坐标表示文本相似度值,气泡大小表示对应位置的耦合关系数量。当耦合强度在4以上,同时三对文本相似度大于0.4时,引用关系的分布变

中国知网 https://www.cnki.net

化较为显著。Chen D Z等[23]的研究结果表明耦合强度达到3.14时,存在漏引关系。因此,本文猜测当CS(R-C)>4且Sim(R-C,F-C,R-F)>0.4时,耦合关系可能发生了漏引行为。

目标数据集构建中,耦合强度和文本相似度共同得到的耦合关系为329 384条,经过去除没有期刊因子、没有作者、空值及其他不符合条件的数据之后,剩余276 796条数据作为研究对象。数据分析过程中,主要运用非参数检验的方法,即卡方检验和K个独立样本检验,从文献题录、文献内容、作者和期刊四个方面分析漏引行为表现特征。

3 漏引行为表现特征分析

3.1 文献题录表现特征分析

3.1.1 语言、文献类型和学科表现特征分析

在文献题录中,由于语言、文献类型、学科和被引频次的数据特征不同,因此采用了不同的检验方法。由于文献C在所有数据中都引用文献F,因此无法做差异性对比,所以从文献C引用文献R和文献C漏引文献R的角度出发,判断文献C引用文献R是否受到语言、文献类型、学科的影响。对比文献C引用文献R和文献C漏引文献R在三种题录类型上的差异属于对比定类数据与定类数据之间的差异性,因此采取卡方检验,结果见表1。其中有效个案数是指样本数量;卡方值是指观察值与理论值之间的偏离程度,观察值与理论值越接近,卡方值越小,反之越大;渐进显著性小于0.05时,具有统计学意义,存在显著差异性。

表1 不同学科中语言、文献类型和学科卡方检验结果

Table 1 Chi-square test results of language, literature types and disciplines in different disciplines

学科	文献R-C题录类型	有效个案数	皮尔逊卡方值	渐进显著性(双侧)
自然科学	语言	838730	139.363	0.000
	文献类型	838730	316.143	0.000
	学科	838730	59393.308	0.000
人文科学	语言	6609	23.983	0.000
	文献类型	6609	22.143	0.000
	学科	6609	33.243	0.000
社会科学	语言	44210	4.347	0.037
	文献类型	44210	7.005	0.008
	学科	44210	936.567	0.000

从卡方检验显著性结果来看,三种大类学科的文献中,语言、文献类型与学科都存在显著差异。对于自然科学与人文学科两个大类而言,三种文献R-C的题录类型的显著性都小于0.001,因此可以认为在这两个大类学科中,题录类型与漏引行为存在显著差异。在社会科学的文献中,语言(P=0.037)和文献类型(P=0.008)的P值也小于0.05。因此,总体上来看,可以认为文献C引用文献R和文献C漏引文献R在语言、文献类型、学科的表现上具有显著差异性,文献C引用文献R可能受到语言、文献类型、学科的影响。

3.1.2 被引频次表现特征分析

文献F和文献R被引频次差值的数据特征符合正态检验,但存在方差不齐,因此采用于非参数检验中K个独立样本检验中的中位数检验。中位数检验是利用卡方分布统计量对多组独立被引频次的分析,检验他们来自总体的中位数是否存在显著差异。由于该检验不能对各领域单独检验,因此只有所有领域综合的检验结果。此外为了更形象地了解文献F-R被引频次的差值在三个领域中的数据特征,利用对数化后的被引频次做了箱线图,见图4(a)。

通过对比文献R和文献F在被引频次上是否存在明显差异。对文献F-R被引频次的差值进行中位数检验,得到中位数2.9090,卡方值4288.489(P<0.000)。可以看出,文献F-R被引频次的差值具有显著性,且中位数和卡方值都较大,说明显著性较强。说明文献F与文献R的被引频次的差值存在明显差异。

在(a)图中,可以较为明显地看出三个领域中被引频次(R-F)的最值、中值和四分位数差异较大。另外由于自然科学和社会科学领域数据量大,因此有较多的异常值。可以看出三个领域的被引频次(R-F)差值的中值都大于零,尤其是自然科学和社会科学,这意味着文献F的被引频次总体上要高于文献R。Price[43]曾指出:相对于被引量较少的文献,被引量较高的文献更容易获得再次引用。因此,由于文献R、F在相关研究领域内的影响力和权威性,文献C的作者更倾向于引用影响力较高的F,而漏引影响力较低的文献R。因此猜测,文献R、F的被引频次大小及其差距可能是文献C漏引文献R的重要影响因素。

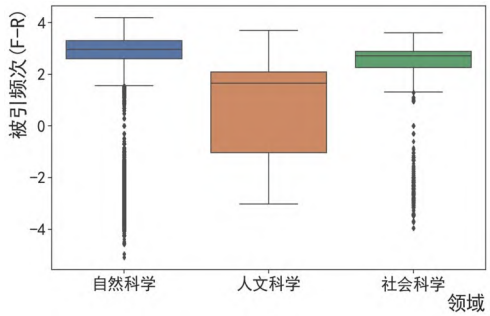
综上所述,文献C引用文献R与文献C漏引文献R受语言、学科、文献类型的影响,文献C引用文献F而漏引文献R的漏引行为在被引频次上表现特征明显,怀疑被引频次可能一定程度上影响了潜在漏引行为。

3.2 文献内容表现特征分析

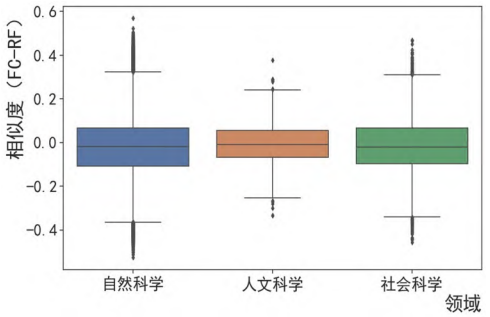
跟作者研究主题越接近的文献越倾向于被引用。因此,通过对比文献F-C和文献R-F的文本内容相似度,探索文献R和文献F在文本内容相似度上是否存在明显差异。文献R和文献F在文本内容相似度差值的数据特征符合正态检验,但是方差不齐,因此也采用中位数检验。得到中位数为-0.0179,卡方值2.245(P<0.326)。可以看到,文献F-C和文献R-F的文本内容相似度差异并不显著,且中位数和卡方值都非常小,也佐证了该结果。

为了更形象地了解不同领域中文献F-C和文献R-F的文本内容相似度差异,也做了箱线图,见图4中(b)。可以看到三个领域虽然由于数据量的差异,极端值的表现不同,但是其最值、中值和四分位数都比较一致,从侧面证明了中位数检验结果。因此猜测即使在不同领域,文献F-C和文献R-F的文本内容相似度差异并不是导致文献C引用文献F

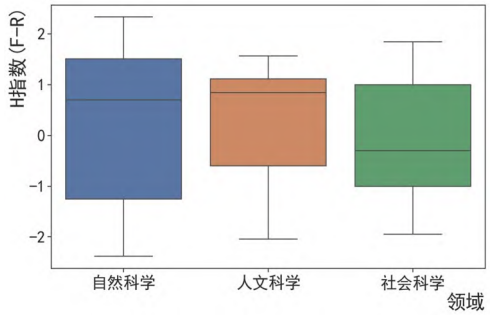
中国知网 https://www.cnki.net



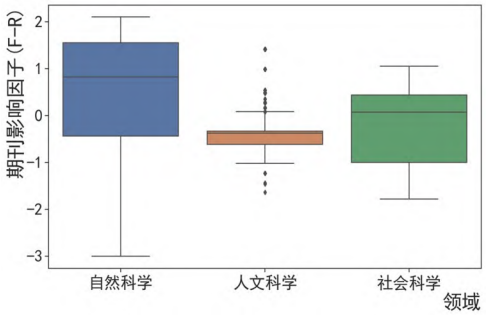
(a) 不同领域中被引频次(R-F)的差值对比



(b) 不同领域中文本相似度(RC-FC)的差值对比



(c) 不同领域中作者H指数(R-F)的差值对比



(d) 不同领域中期刊影响因子(R-F)的差值对比

图4 不同领域中各类表现特征差异对比

Figure 4 Comparison of performance characteristics in different disciplines

不引用文献R的原因。

3.3 作者表现特征分析

3.3.1 作者自引和高被引作者分析

学者的研究通常具有连续性,可能会出现自引行为。自引被认为是对前期研究成果进一步延伸与扩展[44]。由于一篇文章有多个作者,本文中只要文献C引用了包含文献R或者文献F中任意一个作者的论文,两篇文章之间就形成了一条作者自引。通过对文献R-C和文献F-C之间的作者自引进行卡方检验,探索两者之间的作者自引是否存在明显差异,见表2。

当作者的学术影响力不低于世界平均水平,或者至少有一篇引起了学术界的集中关注时,该作者可能会成为高被引作者。高被引作者数据来源于WOS数据库。笔者对文献R和文献F的通讯作者或第一作者的高被引作者进行了卡方检验,见表2。对比文献F-C和文献R-C在作者自引和高被引作者中的差异属于对比定类数据与定类数据之间的差异性,因此采取卡方检验。此外,为了对比文献R-C和文献F-C在作者自引数量上的差异,文献R和文献F的通讯作者或第一作者的高被引作者数量上的差异,本研究计算其数量如表3所示。

在表2中,可以比较明显地看到在三种学科中文献R-C和文献F-C之间的作者自引存在显著差异,文献R和文献F之间的高被引作者存在显著差异。进一步结果见表3。

表2 不同学科中作者自引和高被引作者卡方检验结果

Table 2 Chi-square test results of author self-citation and highly cited author in different disciplines

学科	类型	有效个案数	卡方值	渐进显著性(双侧)
人文科学	自引	412	22.266	0.000
	高被引	412	-	-
社会科学	自引	9791	316.532	0.000
	高被引	9791	544.815	0.000
自然科学	自引	266593	13463.801	0.000
	高被引	266593	4581.258	0.000

注:根据本文计算方法,在人文科学领域中没有出现高被引数据,因此无检验结果。

表3 不同学科中作者自引和高被引作者统计

Table 3 Statistics of self-cited and highly cited authors in different disciplines

学科	作者自引占比		高被引作者占比	
	文献F-C	文献R-C	文献F	文献R
自然科学	7.20%	3.15%	14.56%	15.61%
人文科学	4.96%	1.60%	0.00%	0.97%
社会科学	3.46%	0.65%	5.48%	2.25%

从作者自引结果来看,三个领域中文献F-C的自引占比要明显高于文献R-C,说明文献C引用文献F时更倾向于作者自引,在文献C漏引文献F时,作者自引相对不明显。所以,作者自引可能影响潜在漏引行为,导致文献C漏引文献R。

中国知网 https://www.cnki.net

从高被引作者结果来看,在自然科学和人文科学领域,文献F-C与文献R-C的高被引作者存在显著差异,文献R的高被引作者数量要高于文献F;在社会科学领域中,文献F的高被引作者占比则是文献R的2倍。在文献R的高被引作者数量高于文献F的情况下,文献C引用文献F而漏引文献R,说明至少在自然科学和人文科学领域,高被引作者不影响潜在漏引行为。但在社会学领域,高被引作者不影响潜在漏引行为。因此高被引作者是否影响潜在引用行为应当视具体领域而定。

3.3.2 作者H指数分析

学者的学术影响力在很大程度上能体现其在学术共同体中的地位,作者出于权威和崇拜等因素更倾向于引用学术影响力更高的作者,其经典评价指标通常是H指数[45]。因此,为了研究文献C引用文献F而漏引文献R在H指数的表现特征,且文献R和文献F在H指数差值的数据特征符合正态检验,但是方差不齐,因此也采用中位数检验。得到中位数为0.6021,卡方值255.189(P<0.000)。可以看到在总体性上,H指数(R-F)的差值具有显著性,但相对于被引频次的检验结果,中位数和卡方值都并不大,说明显著性不强。意味着文献F与文献R作者的H指数差值存在差异,但是不明显。

为了进一步了解H指数(R-F)的差异特征,接下来在对不同领域中H指数(R-F)进行了对数化的基础上做了箱线图,见图4(c)。可以较为明显地看出三个领域中作者H指数(R-F)的最值、中值和四分位数差异较大。可以看出自然科学和人文科学的H指数(R-F)差值的中值都大于零,而社会科学的中值小于零。人文科学的H指数(R-F)差值总体上大于零,自然科学和社会科学的H指数(R-F)差值在正值和负值之间的分布较为均衡。这意味着在人文科学中文献F的作者H指数要高于文献R,但在自然科学和社会科学中,文献F和文献R的作者H指数比较则较为不明显。结合检验结果推测,作者学术影响力可能一定程度上影响了潜在漏引行为,导致文献C引用文献F而漏引文献R,但是这种影响并不强烈。

漏引行为在作者自引、作者H指数上差异化明显,在社会学领域的高被引作者上差异化明显。因此作者自引、作者H指数可能一定程度上影响了潜在漏引行为,高被引作者在社会科学领域可能一定程度上影响潜在引用行为,在其他两个领域可能并不影响潜在引用行为。

3.4 期刊表现特征分析

影响因子更高的期刊,影响力更大,作者为了获得认可更倾向于引用影响因子更高的期刊[39]。因此,为了研究文献C引用文献F而漏引文献R在期刊影响因子上的表现特征,且文献R和文献F在期刊影响因子差值的数据特征符合正态检验,但是方差不齐,因此也采用中位数检验。得到中位数为0.7292,卡方值6060.061(P<0.000)。可以看到在总体性上,期刊影响因子(R-F)的差值具有显著性,相对于作者H指数的检验结果,中位数和卡方值都较大,说明显著性较强。意味着文献F与文献R作者的H指数差值存在较为明显的差异。

为了进一步了解期刊影响因子(R-F)的差异特征,接下来也对不同领域中期刊影响因子(R-F)做了对数化并做了箱线图,见图4(d)。可以看出三个领域的差异非常明显,在极端值表现上,只有人文科学存在极端值,但是该领域数据主体的变化范围较小,且基本处于负值范围,意味着文献F的期刊影响因子要低于文献R,该领域数据体量非常小,不易对综合数据产生影响;自然科学数据主体的变化范围较大,且大部分处于正值范围,意味着文献F的期刊影响因子要高于文献R,且该领域数据体量较大,较易影响综合数据;相对于前两种领域,社会科学数据主体的变化范围处于中间位置,但大部分为负值,意味着大部分文献F的期刊影响因子要低于文献R,但该领域数据体量较小,也不易对综合数据产生影响。结合期刊影响因子的检验结果推测,期刊影响因子更倾向于在自然科学中影响潜在引用行为,导致文献C引用文献F而漏引文献R。

4 结论与展望

文献漏引是普遍存在且相对隐蔽的现象,既属于学术规范问题,又属于学术道德问题。本文从文献漏引的角度出发,分析了文献漏引结构,并搜集了自然科学、社会科学和人文科学的15个学科363篇文献及610 740条实际引用关系与23 310 977条可能发生漏引行为的引用关系数据,之后通过文本相似度与耦合强度确定可供分析的276 796条耦合关系数据,最后从文献题录、文献内容、作者和期刊四个方面分析漏引行为表现特征。结果显示,漏引行为在语言、学科、文献类型、被引频次、作者自引、作者H指数、期刊影响因子上表现特征明显,在高被引作者上表现特征则视领域而定,在文本内容相似度上表现特征并不明显,并推测表现明显的特征可能一定程度上影响了潜在漏引行为。

本文也存在一定的不足之处,虽然本文的数据量大,但是对于精细分析每一个学科的数据尚存欠缺;本文仅对文献数据进行了分析,对于多个因素之间的相互影响并未做分析;本文未考虑数据库障碍的不可获得性,时间久远的不可获得性,共知的理论和思想是否作为可引用未作讨论。因此为了解决以上问题,有必要对以上数据进行进一步分析。

在未来研究中,将会以解决不同因素之间的相互影响为主要研究工作。另外,为了识别文献不合理引用,有必要在文献漏引的基础上做进一步拓展——文献不相关引用,以多个学科的全文本数据为研究对象。首先,从文本相似度和耦合引用强度两方面确定非相关引用是否存在。其次,从引用强度、引用位置、引用动机、引用情感、作者自引等方面分析非相关引用的特点,从而规避非相关引用的不合理引用行为。

中国知网 https://www.cnki.net

参考文献

[1] 赵冬生.知识引用新论[J].情报杂志,2002(2):19-21.

[2] 周九常,蔡丽萍.论知识引用[J].情报资料工作,2001(4):8-10.

[3] 苏广利,许新军.社科论著参考文献引用中的七种不良行为[J].图书馆工作与研究,2002(2):15-17.

[4] JONATHAN R C, STEPHEN C. The Ortega Hypothesis[J].Science,1972,178:368-75.

[5] 陈敏.试析伪引文的形成原因与对策[J].科技信息(科学教研),2008(7):220-229.

[6] 国家新闻出版署.《学术出版规范——期刊学术不端行为界定(CY/T174— 2019)》[EB/OL].[2023-01-02]. https://www. nppa. gov. cn/nppa/contents/805/102815.shtml.

[7] 杨卫平.20种常见科研不端行为及其认定要点[EB/OL]. [2023-01-02]. http://jianshen. cas. cn//kyddwyh/tszs/202005/t20200522_4747248.html.

[8] DONNER P. Missing citations due to exact reference matching: Analysis of a random sample from WoS.Are publications from peripheral countries disadvantaged? [C]//21st International Conference on Science and Technology Indicators(STI)-Peripheries,Frontiers and Beyond,Valencia, 2016:85-89.

[9] JIANG ZR, LIU XZ. Recovering Missing Citations in a Scholarly Network A 2-Step Citation Analysis to Estimate Publication Importance[C]//13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL),Indianapolis,IN,2013:419-420.

[10] MOED HF.Citation Analysis in Research Evaluation (Information Science & Knowledge Management) [C]//Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics,2005:437-441.

[11] FAUCHER G. Missing Citations[J]. American Journal of Physics,1987,55(6):490.

[12] LEUNG PT. MISSING CITATIONS - RESPONSE[J]. American Journal of Physics,1987,55(11):696-970.

[13] FELTRACCO P, ORI C. Missed citation[J]. Acta Anaesthesiologica Scandinavica,2013,57(5):674-674.

[14] BELEÑA JM, NUÑEZ M. Reply to "missed citation?"[J]. Acta Anaesthesiologica Scandinavica,2013,57(5):674.

[15] MCDONELL JR.Missing Citation[J].Social Work,1993,38(6):780.

[16] FRICKE B. Missing citation[J]. Chemical & Engineering News ,2003,81(7):8.

[17] JAMA INTERN MED. Missing Reference Citation [J]. JAMA Internal Medicine,2016,176(10):1579-1579.

[18] 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.

[19] MACROBERTS M H, MACROBERTS B R. Problems of citation analysis: A critical review[J]. Journal of the American Society for Information Science, 1989, 40(5): 342-349.

[20] 赵伯兴.论引文行为中非道德风险的防范与规避[J].情报科学,2005(6):849-852.

[21] LIN J,YU Y,SONG J. et al. Detecting and analyzing missing citations to published scientific entities[J]. Scientometrics ,2022,127:2395 - 2412.

[22] HUANG M H, CHEN D Z, DONG HR. Identify Technology Main Paths by Adding Missing Citations Using Bibliographic Coupling and Co-citation Methods in Photovoltaics[C]//2011 Proceedings of PICMET ´11: Technology Management in the Energy Smart World (PICMET), 2011:1-6.

[23] CHEN D Z,HUANG M H,HSIEH HC,et al. Identifying missing relevant patent citation links by using bibliographic coupling in LED illuminating technology [J].Journal of Informetrics,2011,5(3): 400-412.

[24] CIOTTI V,BONAVENTURA M,NICOSIA V,et al.Homophily and missing links in citation networks [J]. EPJ Data Science,2016,5(1):7-20.

[25] 戴泳,张四新.学术论文故意漏引现象分析[J].现代情报,2005(4):48-50.

[26] 郑美莺.剖析参考文献的虚假和不当引用[C]//华东地区高等院校自然科学学报编辑协会.学报编辑论丛(第十四集).上海:上海交通大学出版社,2006:103-105.

[27] 唐虹,朱银周.刻意漏引重要参考文献稿件的学术不端审查[J].云南开放大学学报,2022,24(4):113-119.

[28] 朱大明.同行审稿应注意对参考文献引证的鉴审[J].科技导报,2015,33(12):4,125.

[29] 叶新明.引文中伪引和漏引的机制分析[J].中国图书馆学报,1998(1):72-74.

[30] 郭玲,陈燕.参考文献著录中的学术道德缺失现象及其防范[J].编辑学报,2007(1):8-10.

[31] 许英杰,谢双玉.学术论文参考文献"用"而"不引"的经济学分析及对策[J].情报科学,2008(4):572-575.

[32] 马兰,赵新力,孙晓艳.科技论文中参考文献的故意漏引现象探析[J].编辑学报,2005(3):188-189.

[33] 王志标.学术期刊论文引用失范表现、原因及治理[J].中国出版,2020(21):46-50.

[34] 李锋.参考文献著录中的不良行为及其防范[J].湖南广播电视大学学报,2017(3):93-96.

[35] 许雪梅,阳丽霞,刘吉元.科技期刊参考文献内容真实性

中国知网 https://www.cnki.net

存在的问题及预防对策[J].长沙铁道学院学报(社会科学版),2005(1):162-169.

[36] 王华菊,金丹,陈竹,等.科技论文参考文献著录的常见错误分析[J].编辑学报,2014,26(S1):112-113.

[37] BOOKSTEIN A,YITZHAKI M.Own-language preference: A new measure of "relative language self-citation" [J].Scientometrics,1999,46(2):337 - 348.

[38] SHU F.Chinese-language articles are biased in citations[J].Journal of Informetrics,2015,9:526-528.

[39] TAGLIACOZZO R.Self-Citations in Scientific Literature[J].Journal of Documentation,1997,33(4):251-265.

[40] TEIXEIRA D A,SILVA J A.The Matthew effect impacts science and academic publishing by preferentially amplifying citations,metrics and status[J].Scientometrics,2021,126(6): 5373-5377.

[41] 王忠武.论自然科学、社会科学、人文科学的三位一体关系[J].科学学研究,1999(3):3-9,111.

[42] 郑文涛.人文社会科学若干概念辨析[J].首都师范大学学报(社会科学版),2008(3):141-148.

[43] PRICE DJD.Citation classic-Little science,big science[J].Current Contents,1983,29:18.

[44] 刘桂琴.基于作者自引的知识扩散分析[J].情报杂志,2018,37(7):146-149.

[45] 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,2009,28(6):8-14.

(责任编辑:毛秀梅)

Analysis of the Performance Characteristics of Literature Missing Citations Behavior Based on Non-Parametric Test

YANG Wenxia[1,2], *DENG Sanhong*[1,2], *YANG Jie*[1,2], *ZHANG Qi*[1,2], *WANG Hao*[1,2]

(1.School of information management, Nanjing University Nanjing 210033*,China;*

2.Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023*,China)*

Abstract:【Purpose/Significance】The phenomenon of missing citations is common in document citation. Analyzing the characteristics of missing citations can help reveal the motivation and context behind unreasonable citation behavior, and avoid the negative impact of missing citations on the value of academic papers.【Method/Process】Construct a conceptual model of literature missing citation, use the sample literature from 15 disciplines in the Web of Science core collection database, including nature science, social science, and human science as the data source, determine the data available for analysis through text similarity and citation coupling strength, and explore the performance characteristics of omission behavior from four aspects: literature titles, literature content, authors, and journals using non-parametric testing.【Results/Conclusion】The results show that missing citations behavior has significant characteristics in language, discipline, document type, citation frequency, author self citation, author H index and journal impact factors. For highly cited authors, the performance characteristics depend on the domain. It is not obvious in text content similarity. It is suspected that obvious characteristics may affect potential missing citations behavior to some extent.【Innovation/Limitation】This paper quantitatively captures the performance characteristics of missing citations behavior, which is helpful to explore the work of quantitative research literature missing citations, and provides a reference for literature missing citations identification based on large-scale literature data.

Keywords: Literature Missing Citations; Citation Coupling Strength; Text Similarity; Non-parametric Test; Citation Feature

中国知网 https://www.cnki.net