基于 Web of Science 的论文使用次数和被引频次的相关性分析

■丁佐奇

收稿日期:2017-07-12
修回日期:2017-09-14

中国药科大学《中国天然药物》编辑部,江苏省南京市童家巷 24 号 210009

摘 要 **【目的】**研究论文使用次数与被引频次的相关性,以判断使用次数是否可作为论文影响力评价的客观指标。**【方法】**利用 Web of Science(WoS)新推出的"文献级别用量指标",对《中国天然药物》高使用和高被引论文的使用次数和被引频次的相关性进行研究,并研究整合替代医学学科论文使用次数对被引频次的表征意义。**【结果】**被引频次、使用次数及使用次数(180 天)三者之间互相具有较好的相关性。2014 年和 2015 年高使用论文获得高引用,表明期刊可以利用论文发表 2 年内的使用次数预测其未来是否高被引;年代越久的高被引文献,也可能带来后续使用次数的二次提高。**【结论】**论文使用次数与被引频次的相关性研究未能形成定论,时间跨度的选择可能是造成矛盾结果的主要原因。论文使用次数与被引频次的峰值出现时间不同,但二者呈正相关,使用次数可作为论文影响力早期评价的客观指标。

关键词 使用次数;被引频次;相关性;预测性
DOI:10.11946/cjstp.201707120578

科学计量指标可以客观反映期刊论文的质量和影响,是期刊评价和科研管理工作的重要工具[1]。论文的被引频次和下载量是用于论文评价的金指标,被引频次是衡量学术质量以及学术影响力的重要评价指标,下载量则可测度论文的可见度及传播速率。因此,对论文及期刊进行评价,包括对核心期刊的确定,被引频次与下载量同时作为文献价值的表征而共同被纳入到评价指标体系。被引频次越高,学术影响力越大,Web 即年下载率越高,说明读者对期刊的兴趣越大[2]。

近年来,国内学者主要利用中国知网(CNKI)等引文数据库对论文的被引频次与下载量的相关性进行了研究。王丽[3]研究发现 CNKI 中医药卫生科技类文献的下载量与被引频次之间没有明显的相关性,无规律可循。张小强[4]对 CNKI、中国科学引文数据库(CSCD)和中国人文社会科学引文数据库(CHSSCD)来源期刊的下载量和被引频次的相关性进行了研究,发现两者之间具有高度正相关性。严美娟等[5]对 CNKI 中的 6 种儿科类期刊进行了研究,发现被引频次与下载量相关性的高低和不同栏目论文内容有关。国际上有不少学者利用各大国际数据库对论文的下载量与被引频次的相关性做了广泛的研究。Nieder 等[6]研究了 5 种肿瘤学领域的开放获取(OA)期刊的情况,下载量数据来源于这些期刊的网站,被引频次数据来源于 Scopus 数据库,结果发现下载量和被引频次的相关性较差。赵一权等[7]利用美国计算机协会数字图书馆数据库研究了计算机领域的 31 种期刊,发现无论是在期刊层面,还是在文献层面,下载量和被引频次都具有较强的正相关性。Jiang 等[8]对 *J Am Med Inform Assoc* 网站的论文下载量和其在 Web of Science(WoS)的被引频次进行了研究,发现两者之间存在较强的正相关性。Moed 等[9]根据 ScienceDirect 平台的下载量和 Scopus 数据库的被引频次发现所研究期刊的下载量和被引频次呈正相关,但高被引和高下载论文之间重合的论文较少。Guerrero-Bote 等[10]也利用 ScienceDirect 平台和 Scopus 数据库对期刊和论文两个层面下载量和被引频次的相关性进行研究,发现期刊层面的相关性优于论文层面。Jahandideh 等[11]则利用 ScienceDirect 平台和 Scopus 数据库,研究了高下载和低下载论文各 25 篇及其后续被引频次的相关情况,发现一定时段内的下载量可以较好地预测论文未来可能达到的被引频次。Subotic 等[12]同样利用 ScienceDirect 平台和 Scopus 数据库,研究了高下载及低下载论文各 129 篇及其后续被引频次的相关情况,发现先期的下载量和后续的被引频次呈

基金项目:中国高校科技期刊研究会专项基金资助(CUJS 2017-001)。

作者简介:丁佐奇(ORCID:0000-0003-0957-4193),博士,副编审,硕士生导师,E-mail:zqding1028@163.com。

正相关。Lippi 等[13]利用 ScienceDirect 平台的下载量和 SciVerse 上的被引频次,发现一定时段内的高下载能够带来后两年的高引用。

从以上文献综述可以看出,由于 ScienceDirect 平台能够提供下载量、而 Scopus 数据库能够提供被引频次,国际上利用其对下载量和被引频次的相关性做了大量研究。WoS 虽然有强大的文献检索和分析功能,但一直没有文章下载量这个指标,直到 2015 年 9 月 26 日,WoS 数据库平台升级至 5.19 版,才新增了"文献级别用量指标",即使用次数。"文献级别用量指标"反映了某篇论文满足用户信息需要的次数,具体表现为用户点击了指向出版商处全文的链接(直接链接或开放链接),或是对论文进行了保存以便在题录管理工具中使用(通过直接导出或另存为将论文重新导到其他格式)。使用次数每天更新一次,使用次数是从 2013 年 2 月 1 日开始某条记录的全文得到访问或是对记录进行保存的次数,该次数可能会逐渐增长或保持不变;使用次数(180 天),是最近 180 天内通过某条记录的全文得到访问或是对记录进行保存的次数,该计数会随着固定时段结束日期的推进而上升或下降[14]。

鉴于目前利用 WoS 数据库进行下载量和被引频次相关性研究的文献较少[15-16],本文利用 WoS 数据库对《中国天然药物》(CJNM)高被引论文和高使用论文各自的使用次数和被引频次的相关性进行研究。有大量的研究发现,科技期刊论文被引量的峰值年在论文发表后的第 6~8 年,而论文下载量的峰值年在论文发表后第 2 年[2,17-19]。由于高被引和高下载论文的峰值年相差较远,同一发表年份的高被引和高下载论文重合率的研究价值不大,本研究还对 CJNM 及其所在学科整合替代医学论文同一年份的高使用和低使用论文的后续被引频次进行比较,分析 2014 年和 2015 年高使用和低使用论文对应的被引频次是否有差异。

1 研究对象与方法

1.1 研究对象

(1) 研究 CJNM 2013—2017 年在 WoS 核心合集中总被引频次(T_C)排名前 50(TOP50)论文、使用次数(180 天)(U_1)TOP50 论文、使用次数(U_2)TOP50 论文的具体情况,分析被引频次 TOP50 论文 T_C 和 U_1、U_2 的相关性,使用次数(180 天)TOP50 论文 U_1 和 T_C、U_2 的相关性,使用次数 TOP50 论文 U_2 和 T_C、U_1 的相关性。(2)研究 CJNM 及整合替代医学学科的论文 2014 年和 2015 年单年的使用次数 TOP50 论文和排名最后 50(LOWER50)论文,分析其后续的被引频次是否有显著性差异。CJNM 数据检索日期为 2017 年 4 月 20 日,整合替代医学学科数据检索日期为 2017 年 9 月 10 日。

1.2 研究方法

采用 GraphPad Prism 6.01 软件分析数据,Pearson 相关性分析,双尾,置信区间为 95%,显著性系数 $P<0.05$ 时差异有统计学意义;t 检验 $P<0.05$ 时差异有统计学意义。

2 研究结果

2.1 相关性分析

对 CJNM 2013—2017 年在 WoS 核心合集中被引频次、使用次数(180 天)、使用次数 TOP50 论文的发表年份进行统计,结果如表 1 所示。被引频次 TOP50 论文主要集中在 2013—2015 年,其中以 2013 年居首(24 篇,占 48%);使用次数 TOP50 论文同样集中在 2013—2015 年,但以 2014 年居首(18 篇,占 36%);使用次数(180 天)TOP50 论文集中在 2016 年(30 篇,占 60%)。

表 1 CJNM 被引频次、使用次数、使用次数(180 天)TOP50 论文数量分布 (篇)

指标	2013 年	2014 年	2015 年	2016 年	2017 年
被引频次 TOP50	24	19	6	1	0
使用次数 TOP50	15	18	12	5	0
使用次数(180 天)TOP50	4	9	6	30	1

表 2 显示被引频次 TOP50 论文的 T_C 和 U_1、U_2 均呈正相关性,其中 T_C 和 U_2 的相关性较强(相关系数 $r=0.802$,$P<0.0001$)。表 3 显示使用次数(180 天)TOP50 论文的 U_1 和 T_C、U_2 均呈正相关性,其中 U_1 和 U_2 的相关性较强($r=0.646$,$P<0.0001$)。表 4 显示使用次数 TOP50 论文的 U_2 和 T_C、U_1 均呈正相关性,其中 U_2 和 T_C 的相关性较强($r=0.843$,$P<0.0001$)。

表 2 CJNM 被引频次 TOP50 论文的 T_C 和 U_1、U_2 相关性

指标	范围/次	均值	r	P
U_1	0~25	4.74	0.649	<0.0001
U_2	4~135	27.16	0.802	<0.0001

表 3 CJNM 使用次数(180 天)TOP50 论文的 U_1 和 T_C、U_2 相关性

指标	范围/次	均值	r	P
T_C	0~84	5.64	0.489	0.0003
U_2	6~135	24.62	0.646	<0.0001

表 4 CJNM 使用次数 TOP50 论文的 U_2 和 T_C、U_1 相关性

指标	范围/次	均值	r	P
T_C	0~84	8.88	0.843	<0.0001
U_1	1~28	7.12	0.545	<0.0001

2.2 预测性分析

CJNM 2014 年和 2015 年单年的使用次数 TOP50 论文和 LOWER50 论文的 T_C 差异性见表 5，2014 年和 2015 年使用次数 TOP50 论文和 LOWER50 论文的 T_C 差异均有显著性，2014 年和 2015 年使用次数 TOP50 论文的篇均被引频次分别是使用次数 LOWER50 论文的 2 倍和 11 倍。

表 5 CJNM 使用次数 TOP50 论文和 LOWER50 论文的 T_C 差异性

年份	指标	范围/次	均值	P
2014	TOP50	0~22	4.40	<0.0001
	LOWER50	0~10	2.20	
2015	TOP50	0~13	2.90	<0.0001
	LOWER50	0~2	0.26	

为了探讨上述个刊案例结果是否可以推及到整个学科，继续分析了整合替代医学学科 2014 年和 2015 年单年的使用次数 TOP50 论文和 LOWER50 论文(表 6)，发现 2014 年和 2015 年使用次数 TOP50 论文和 LOWER50 论文的 T_C 差异均有显著性，2014 年和 2015 年使用次数 TOP50 论文的篇均被引频次分别是使用次数 LOWER50 论文的 31 倍和 16 倍。

表 6 整合替代医学学科使用次数 TOP50 论文和 LOWER50 论文的 T_C 差异性

年份	指标	范围/次	均值	P
2014	TOP50	1~64	18.10	<0.0001
	LOWER50	0~8	0.58	
2015	TOP50	3~44	12.64	<0.0001
	LOWER50	0~7	0.78	

3 分析与讨论

由于大多数情况下文献的被引频次与该文献质量呈高度正相关，使得引文分析作为科学评价的方法具备一定的可行性，但是作者引用的文献往往仅占其在研究工作中所阅读过的文献的一部分，那么其中未被引用的文献的价值该如何去体现？下载量成为一个日渐公认的评估指标，它在直观上能够与该文献的被阅读次数相对应。相对于被引频次来说，下载量能够较早地反映文章的受关注度。文章的下载量与被引频次是否呈正相关？普遍认为：文章被阅读的次数越多，被引用的可能性越大。事实上，下载量虽然反映了文章被社会关注的程度，但它仅是文章被引的前奏，并不是所有的下载都会带来引用。Carey[20] 认为下载量较于被引频次能够更好地反映期刊的真实影响力。也有专家认为，下载量只是反映了论文的可获得性及可见度，而被引频次才能真正反映论文的学术质量。目前尚未有论文下载量与被引频次相关性的定论。

WoS 数据库作为国际上公认的权威数据库，一直没有下载量指标是一个比较大的遗憾，且该数据库文献的入库速度比较缓慢，当然这和该数据库严谨的收录流程有关，不过文献信息的实时公开同样是非常重要的。所幸的是，WoS 数据库于 2015 年下半年增加了“文献级别用量指标”，即使用次数指标。本研究对 CJNM 及其所在整合替代医学类论文进行了分析，得出以下几点结论：

3.1 高使用和高被引论文发表年份不一致

分析 CJNM 2013—2017 年发表的论文，发现被引频次 TOP50 论文主要集中发表在 2013 年，使用次数 TOP50 论文主要集中发表在 2014 年，使用次数(180 天)论文主要集中发表在 2016 年。毕竟从引用到发表还有一定的时滞，而使用次数相对及时，本研究结果再次验证了高被引论文是较老的文献，而高使用论文是较新的文献，尤以使用次数(180 天)高的论文为最新。

3.2 使用次数和被引频次具有较好的相关性

本研究中，被引频次和使用次数及使用次数(180 天)三者之间互相具有较好的相关性。多位专家发现论文在发表当年及第二年达到下载高峰，而引用高峰则需要几年的时间才能达到。目前论文下载量与被引频次的相关性未能形成定论，其中时间跨度的选择可能是造成矛盾结果的主要原因。因此，在论文下载量与被引频次相关性的研究中，研究对象的发表年份的选择很重要。例如，Coats[21] 研究了 *Int J Cardiol* 同一年份内的 TOP10 被引和下载论文，发现两者没有重合的论文；还有一些文献选择 5 年甚至 10 年前的论文进行研究，得到下载量与被引频次无相关性的结论，这有待商榷，毕竟 5 年前的论文早已达到被引高峰，而高下载的论文集中在近 2 年，高被引和高下载论文之间无重合也情有可原。

3.3 论文高被引可能带来二次高使用

本研究发现，被引频次 TOP50 论文的被引频次和使用次数及使用次数(180 天)均呈正相关。这和 Wang 等[16] 的研究结果一致，他们利用 WoS 的使用次数指标分析发现，研究人员整体上倾向于使用较新的文献，但被引频次高的老文献有助于其后续使

用次数的二次提高。毕竟,历年发表的文献多如恒河沙数,研究人员可能通过被引频次的降序排列以尽快获得需要的论文。

3.4 论文使用次数可用来预测后续被引频次

本研究还发现,单刊使用次数 TOP50 论文的使用次数和被引频次的相关性较强,单刊和整个学科论文 2014 年和 2015 年的高使用论文和低使用论文分别对应后续的高被引和低被引论文,说明可以通过论文发表后 2 年内的使用次数来预测未来可能达到的被引频次。

此外,影响论文下载量和被引频次的因素还有很多。Timsit 等[22] 利用 WoS 的被引频次和 SpringerLink 的下载量研究了影响 *Intensive Care Medicine* 下载和引用的因素,发现每年第二季度下载量较多,综述下载量较多,且最后一位作者的 h 指数越高,下载越多;而会议报道和专论被引较多,且第一位作者的 h 指数越高,被引频次越高。Wang 等[23] 研究了 *Nature Communications* OA 论文和非 OA 论文,发现 OA 论文较非 OA 论文在引用和下载方面均有优势,且 OA 论文不仅能增加总下载量,还能延长被下载的时限。

综上所述,期刊可以利用论文发表后 2 年内的使用次数对其未来是否为高被引进行预测;年代悠久的高被引文献,也可能带来后续使用次数的二次提高。WoS 文献级别用量指标能够在论文发表后的短时间内,对其影响力进行快速评估,提示该指标可作为论文影响力早期评价的客观指标,期刊可以借助这个指标来选择论文进行及时宣传和推广。

参考文献

[1] 马峥. 通过计量指标分析发现操纵期刊评价结果的行为[J]. 编辑学报,2016,28(6):608-611.

[2] 丁佐奇,郑晓南,吴晓明. 科技论文被引频次与下载频次的相关性分析[J]. 中国科技期刊研究,2010,21(4):467-470.

[3] 王丽. 中国知网数据库中高被引文献与高下载文献类型分析——以医药卫生科技类文献为例[J]. 编辑学报,2015,27(5):503-506.

[4] 张小强. 期刊下载频次与被引频次及影响因子相关性——以中国知网 CSCD 与 CHSSCD 刊物为样本的计量分析[J]. 情报理论与实践,2011,34(8):36-40.

[5] 严美娟,肖丽娟,张莉,等. 儿科类期刊栏目设置与论文被引频次和下载频次之间的关系[J]. 编辑学报,2012,24(4):399-401.

[6] Nieder C, Dalhaug A, Aandahl G. Correlation between article download and citation figures for highly accessed articles from five open access oncology journals[J]. *Springerplus*, 2013, 2(1): 261.

[7] 赵一权,王振民,熊文炳,等. 科学论文的下载与引用关系研究:以 ACM 数字图书馆为例[J]. 中国科技期刊研究,2014,25(6):818-823.

[8] Jiang X, Tse K, Wang S, *et al*. Recent trends in biomedical informatics: a study based on JAMIA articles[J]. *Journal of the American Medical Informatics Association*, 2013, 20 (e2): e198-e205.

[9] Moed H F, Halevi G. On full text download and citation distributions in scientific-scholarly journals[J]. *Journal of the Association for Information Science and Technology*, 2015, 67(2): 412-431.

[10] Guerrero-Bote V P, Moya-Anegón F. Relationship between downloads and citations at journal and paper levels, and the influence of language [J]. *Scientometrics*, 2014, 101 (2): 1043-1065.

[11] Jahandideh S, Abdolmaleki P, Asadabadi E B. Prediction of future citations of a research paper from number of its internet downloads[J]. *Medical Hypotheses*, 2007, 69(2): 458-459.

[12] Subotic S, Mukherjee B. Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles[J]. *Journal of Information Science*, 2014, 40(1): 115-124.

[13] Lippi G, Favaloro E J. Article downloads and citations: Is there any relationship? [J]. *Clinica Chimica Acta*, 2013, 415: 195.

[14] 孙学军. SCI 新增功能"文献级别用量指标"是个什么东东[EB/OL]. [2015-10-10]. http://blog.sciencenet.cn/blog-41174-926981.html.

[15] 付中静. WoS 数据库收录论文文献级别用量指标与被引频次的相关性[J]. 中国科技期刊研究,2017,28(1):68-73.

[16] Wang X W, Fang Z C, Sun X L. Usage patterns of scholarly articles on Web of Science: A study on Web of Science usage count[J]. *Scientometrics*, 2016, 109(2): 917-926.

[17] 方红玲. 我国科技期刊论文被引量和下载量峰值年代——多学科比较研究[J]. 中国科技期刊研究, 2011, 22(5): 708-710.

[18] 丁佐奇,郑晓南,吴晓明. SCI 药学论文被引峰值研究及国别比较[J]. 科技与出版,2012 (8):114-116.

[19] Schlögl C, Gorraiz J, Gumpenberger C, *et al*. Comparison of downloads, citations and readership data for two information systems journals[J]. *Scientometrics*, 2014, 101(2): 1113-1128.

[20] Carey R M. Quantifying scientific merit: Is it time to transform the impact factor? [J]. *Circulation Research*, 2016, 119(12): 1273-1275.

[21] Coats A J. Top of the charts: Download versus citations in the International Journal of Cardiology[J]. *International Journal of Cardiology*, 2005, 105(2): 123-125.

[22] Timsit J F, Citerio G, Lavilloniere M, *et al*. Determinants of downloads and citations for articles published in intensive care medicine[J]. *Intensive Care Medicine Experimental*, 2015, 3(1): A863.

[23] Wang X W, Liu C, Mao W L, *et al*. The open access advantage considering citation, article usage and social media attention[J]. *Scientometrics*, 2015, 103(3): 1149.

Correlation analysis between usage count and citation frequencies based on Web of Science

DING Zuoqi

Editorial Office of *Chinese Journal of Natural Medicines*, China Pharmaceutical University, 24 Tongjia Lane, Nanjing 210009, China

Abstract: [**Purposes**] This paper aims to validate the correlation between usage count and citation frequencies for papers, and to judge whether usage count is a good indicator for evaluating the influence of papers. [**Methods**] Taking *Chinese Journal of Natural Medicines* as an example, we studied the correlation between usage count and citation frequencies based on the Web of Science (WoS) database. Furthermore, we also studied whether the usage count can be used to predict the citation performances in the field of integrative complementary medicine. [**Findings**] There are good correlations among citation frequencies, usage count, and usage count within the recent 180 days. The higher used papers in 2014 and 2015 reached the higher citations, suggesting that the usage count can be used to predict the citation performances of papers. Moreover, the high-citation papers may lead to the second increase in usage. [**Conclusions**] Controversial results have been published; different criteria for the selection of time range is probably the major cause. Although the climax is different for the performance of usage and citation, they are positively correlated. Thus, our results suggest that the usage count is a good indicator for early evaluating the paper influence.

Keywords: Usage count; Citation frequency; Correlation; Prediction

(本文责编:梁永霞)