

Google Scholar 和 CCD 的引文统计分析功能比较

——从期刊被引频次角度分析

邱均平, 温芳芳

(武汉大学 中国科学评价研究中心, 湖北 武汉 430072)

摘要:期刊被引频次是期刊评价中一个重要的计量指标,不同的引文数据库在用于被引频次统计时所得到的结果各不相同,这种统计结果的差异必将对期刊的排名结果造成较大的影响。文章选取“图书馆、情报与文献学”排名前10位的期刊,分别在 Google Scholar 和 CCD(中国引文数据库)中以期刊名为主题词进行精确检索,统计各个期刊在 2007-2008 两年间所刊载的论文的被引频次。通过统计结果的对比分析,我们认为 CNKI 是目前比较理想的中文期刊引文统计工具,而 Google Scholar 却在一定程度上代表着未来引文数据库的发展方向。

关键词:引文数据库;期刊评价;被引频次;引文分析

中图分类号:G237.5 **文献标志码:**A **文章编号:**1008-5831(2011)06-0084-06

一、引言

期刊被引频次在学术期刊评价中是一个重要的计量指标。学术期刊的被引频次能够比较客观地反映出期刊被利用的程度,在一定程度上可以作为学术期刊质量和影响力的评价依据。1955年美国情报学家加菲尔德创立了系统的引文索引理论^[1],1972年加菲尔德又把引文索引工具引入到期刊评价^[2]。在中文期刊的评价中,期刊被引频次也是一个重要的评价指标。无论是北京大学《中文核心期刊要目总览》,还是武汉大学中国科学评价研究中心的《中文期刊评价报告》,亦或是南京大学的 CSSCI 所遴选的核心期刊,都将期刊的被引频次作为基本的评价指标并赋予了较高的权重。但在进行期刊被引频次的统计时,每个评价机构所采用的统计工具各不相同。

目前,维普、CNKI、万方等中文期刊数据库和 Google Scholar 都开设了引文统计分析功能,每个数据库所得出的引文统计结果也各不相同,这种统计结果的差异必将对期刊的排名结果造成较大的影响。为了更加客观地反映出不同数据库在进行引文统计分析时的差异,我们将以 Google Scholar 和中国引文数据库(Chinese Citation Database, CCD)为例进行实证比较,以期能够为中文期刊评价机构和广大的科研工作者提供选择引文分析工具的标准和依据。

收稿日期:2011-10-13

基金项目:国家自然科学基金项目“基于作者学术关系的知识交流模式与规律研究”

作者简介:邱均平(1947-),男,湖南涟源人,武汉大学中国科学评价研究中心教授,博士研究生导师,主要从事信息计量与科学评价、知识管理与竞争情报研究;温芳芳(1984-),女,河南平顶山人,武汉大学博士研究生,主要从事信息管理与科学评价研究。

二、研究方法

在武汉大学中国科学评价研究中心所发布的《中国学术期刊评价研究报告》中,共有 73 种“图书馆、情报与文献学”学术期刊参与评价,我们选取其中排名前 10 位的学术期刊为研究对象:(1)中国图书馆学报;(2)情报学报;(3)图书情报工作;(4)大学图书馆学报;(5)情报科学;(6)图书情报知识;(7)图书馆;(8)图书馆论坛;(9)图书馆杂志;(10)情报理论与实践。

以这 10 种“图书馆、情报与文献学”学术期刊的期刊名称为检索词分别在 Google Scholar 和中国引文数据库中精确查找,统计 2007 - 2008 两年间

每种期刊的载文量以及每种期刊在这两年间所刊载的论文的被引频次。在 CNKI 所开发的中国引文数据库(CCD)中,输入期刊名称和出版时间进行精确匹配,可以直接统计出每种期刊的载文量和总被引频次。而在 Google Scholar 中输入期刊名称和出版时间,只能得出载文量和每篇文章的被引频次,期刊的被引频次需要将每篇文章的被引频次进行累加计算得出。

三、结果分析

课题组借助于 Google Scholar 和中国引文数据库分别对 10 种“图书馆、情报与文献学”期刊的载文量和被引频次进行统计和计算,统计结果如表 1。

表 1 10 种图书馆、情报与文献学期刊的载文量和被引频次统计结果(2007 - 2008)

	Google Scholar			CCD		
	载文量	被引频次	篇均被引频次	载文量	被引频次	篇均被引频次
中国图书馆学报	328	695	2.12	292	1 126	3.86
情报学报	297	429	1.44	295	512	1.74
图书情报工作	1 048	1 460	1.39	1 212	2 310	1.91
大学图书馆学报	306	576	1.88	328	922	2.81
情报科学	931	928	1.00	806	1 702	2.11
图书情报知识	335	353	1.05	318	700	2.20
图书馆	598	690	1.15	576	1 311	2.28
图书馆论坛	793	895	1.13	718	1 559	2.17
图书馆杂志	709	644	0.91	662	1 218	1.84
情报理论与实践	516	621	1.20	482	1 055	2.19
总计	5 861	7291	1.24	5 689	12 415	2.18

为了更加直观地比较 Google Scholar 和中国引文数据库统计结果的差异,用三个柱状图将载文量、被引频次和篇均被引频次分别进行对比。在图 1、图 2、图 3 中,横坐标轴中的阿拉伯数字分别对应“图书馆、情报与

文献学”排名前 10 的期刊:(1)中国图书馆学报;(2)情报学报;(3)图书情报工作;(4)大学图书馆学报;(5)情报科学;(6)图书情报知识;(7)图书馆;(8)图书馆论坛;(9)图书馆杂志;(10)情报理论与实践。

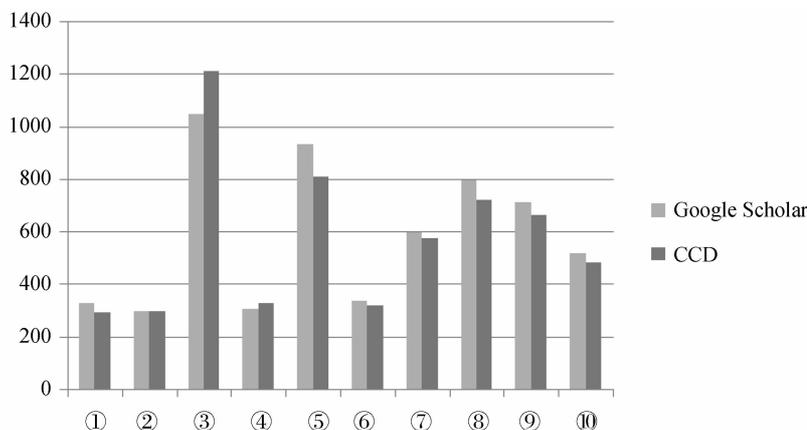


图 1 期刊载文量统计结果对比(2007 - 2008)

从表 1 和图 1 中的数据可以看到,利用 Google Scholar 和中国引文数据库所统计出的每种期刊的载文量各不相同。总体来看,利用 Google Scholar 检出的文献数量大于中国引文数据库的检索结果。但对

其检出结果进行核查之后发现,尽管 Google Scholar 检出的文献最多,但其中存在着许多错误。以《中国图书馆学报》为例,分别在 Google Scholar 和中国引文数据库中以“中国图书馆学报”为主题词做精确检

索,统计2007-2008两年间发表于《中国图书馆学报》的文章数量。Google Scholar 返回记录328条,中国引文数据库返回记录292条。经核查 Google Scholar 所返回的记录中有15篇文章重复,另有11篇文章并非来源于《中国图书馆学报》,检索记录错

误率高达8.23%。而中国引文数据库所返回的记录中没有重复或者错误记录。通过载文量统计可以发现,单从数量角度考虑,Google Scholar 要优于中国引文数据库,但从质量和数量两方面考虑,中国引文数据库要优于 Google Scholar。

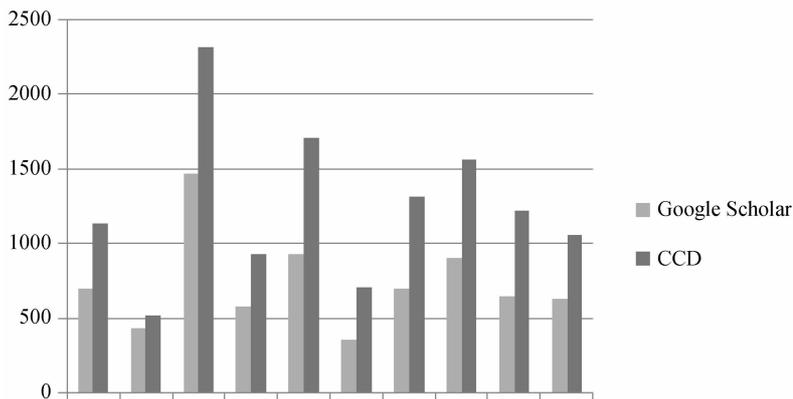


图2 期刊被引频次统计结果对比(2007-2008)

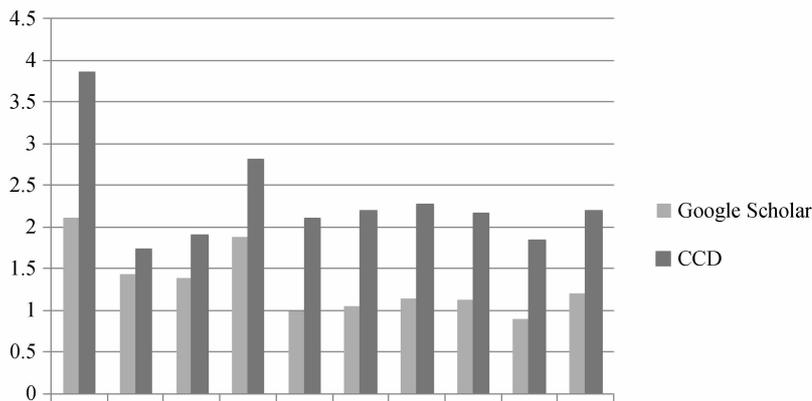


图3 期刊篇均被引频次统计结果对比(2007-2008)

从表1、图2和图3中被引频次和篇均被引频次的统计结果看,Google Scholar 和中国引文数据库在引文统计上存在着更大的差异,从CNKI中检索得出的被引频次远远高于由Google Scholar 检索得出的被引频次。尽管从Google Scholar 中所检出的期刊载文量大于CNKI,但总被引频次和篇均被引频次却远远低于CNKI。由此说明,Google Scholar 对于中文期刊及其引文数据的收录远不及CNKI全面。

2008年上海交通大学情报科技研究所的陶慧卿等人^[3]选择信息科学领域的权威期刊《美国信息科学与技术学会杂志》(Journal of The American Society for Information Science and Technology, JASIST)作为文献源,以2003年至2007年该期刊所刊载的文章作为统计分析对象,分别统计了JASIST在Web of Science 和 Google Scholar 中的被引频次。结果发现无论是总被引频次,还是平均被引频次和最大被引频次,Google Scholar 都要远远高于Web of Science。陶慧卿等人认为,由于在Google Scholar 中得到的引

用文献既包括传统学术期刊,也包括Web 网页、预印本和电子期刊等,所以它不仅能够反映学术期刊之间的引用关系,也能广泛反映学术期刊、Web 网页、预印本、电子期刊之间的引用关系。因此,Google Scholar 的引文数据要比传统文献数据库丰富和全面。但在我们的统计结果中,Google Scholar 的引文数据却不及传统文献数据库丰富和全面。这一差异恰恰反映出 Google Scholar 在中外文学学术期刊引文数据收录上的差异。

在英文环境下,Google Scholar 所收录的引文数据不仅包括传统的学术期刊,而且包含丰富的非期刊类学术文献,其中包括网络学术文献。因此,Google Scholar 不仅能够反映出期刊文献之间的引用关系,而且能够反映出期刊文献被Web 网页、预印本等其他类型文献引用的情况,从而能够更加真实全面地反映出期刊被利用的程度。但在中文环境下,Google Scholar 所收录的中文学术期刊文献及其引文关系基本上都来自于CNKI、维普、万方等传统

的期刊文献数据库。Google Scholar 的引文数据库的全面性、准确性和规范性都不及传统的期刊数据库,因此,目前还不能用作中文期刊评价时引文数据的统计工具。笔者认为,综合考虑各方面因素,CNKI 所开发的中文引文数据库是目前中文期刊评价时比较理想的引文数据统计工具。当然,在期刊评价过程中,为了保证评价结果具有更高的客观性和准确度,也可以同时采用多个引文数据库进行综合检索和统计。

四、研究结论

通过 Google Scholar 和中国引文数据库引文统计结果的对比分析,我们认为中国引文数据库是目前中文期刊评价时比较理想的引文统计工具。中国引文数据库收录了中国学术期刊(光盘版)电子杂志社出版的所有源数据库产品的参考文献,并揭示各种类型文献之间的相互引证关系。它不仅可以为科学研究提供新的交流模式,同时也可以作为一种有效的科学管理及评价工具。中国引文数据库收录 CNKI 所有源数据库产品的参考文献,源数据库包括:中国期刊全文数据库、中国优秀博士学位论文全文数据库、中国重要会议论文全文数据库、中国重要报纸全文数据库、中国图书全文数据库、中国年鉴全文数据库等^[4]。引文数据库内容覆盖自然科学、工程技术、农业、哲学、医学、人文社会科学各个领域,共分 168 个专题。中国引文数据库支持用户从作者、机构、期刊、专题、基金、出版社等多个角度进行引文统计分析。在中国引文数据库中,能够直接检索得到每种中文期刊的载文量、下载量、被引频次、H 指数等指标。CNKI 对其收录的文献进行了规范的标引,保证了检索的质量。因此,从全面性、准确性、易用性等多个方面综合考虑,CNKI 是目前进行中文学术期刊评价时相对比较理想的引文统计工具。

中国引文数据库也并非完美无缺。利用中国引文数据库统计出的期刊被引频次,只能反映学术期刊被其他期刊、学位论文、图书等传统文献引用的情况,并不能反映期刊被学术博文、预印本等纯网络学术文献所利用的情况。而事实上,学术交流正在朝着多元化的方向发展,特别是随着网络技术的发展和普及,纯网络学术文献大量出现,成为广大科研工作者传播和交流知识资源的重要途径。事实上,网络学术文献的研究和创作过程中也大量使用了学术期刊文献,如果不能反映这部分网络学术文献与期刊文献之间的引用关系,那么这种引文数据库就是不全面的。中国引文数据库只是 CNKI 基于中国期

刊全文数据库、中国优秀博士学位论文全文数据库等传统文献数据库开发出的引文分析工具,如果从整个学术交流的全面性考虑,由 CNKI 统计得出的被引频次是单一和片面的,并不能真实准确地反映出学术期刊在整个学术交流中的价值和影响力。

尽管 Google Scholar 对于中文学术文献的收录数量非常有限,检索的质量也不理想,目前尚不能成为中文期刊评价时的引文统计分析工具。但从长远发展来看,Google Scholar 却代表着未来引文数据库的发展方向。对于英文文献来说,Google Scholar 是强大的学术搜索工具和引文统计分析工具。Google Scholar 与 Nature、OCLC、ACM 等多家学术出版商进行了合作。通过这种广泛的合作,Google Scholar 所涵盖的范围无限扩大。通过 Google Scholar,读者可以从学术出版者、专业团体、预印本库以及从网络上获得各种类型的学术文献,包括学术期刊、文摘、同行评议论文、学位论文、图书、预印本、PPT 介绍和技术报告等^[5]。Google Scholar 不仅收录传统的期刊文献,而且大量收录纯网络学术文献。因此,Google Scholar 不仅可以免费搜索跟踪同行评议过的文献的引证文献,还能搜索跟踪会议录、学位论文、预印本和在版文献,以及其他非传统媒体形式文献的引证文献^[6]。Google Scholar 对其所收录的学术文献的引用关系加以整理,能够更加全面真实地反映出各种类型的学术文献之间的相互利用关系,Google Scholar 对于英文文献的引文检索范围要远大于传统的引文数据库。因此,在英文环境下 Google Scholar 已经为引文统计分析提供了一种新选择。

对于中文文献来说,情况并非如此。Google Scholar 对于中文非期刊文献的收录范围和数量还非常有限,也没有对非期刊类学术文献之间的引用关系加以整理。从本次的统计结果可以看出,利用 Google Scholar 统计得出的期刊被引频次远远低于利用中国引文数据库统计得出的结果。因此,目前 Google Scholar 尚不能承担中文期刊引文统计分析的重任。但是 Google Scholar 在中文期刊的引文检索方面却具有较大的潜力,从 Google Scholar 在英文环境下的发展现状来看,Google Scholar 代表着未来中文期刊引文检索和统计分析的发展趋势。Google Scholar 跨越了学科、语言和地域障碍,如果 Google Scholar 能够扩大其对中文非期刊文献的收录范围,并且能够对各种类型的中文学术文献建立知识链接,那么从长远看,Google Scholar 在引文检索和统计分析方面的优势要明显地优于 CNKI、万方等传统的期刊数据库。毕竟 Google Scholar 具有强大的后台

数据库支撑,在纯网络文献的收集方面又具有得天独厚的条件,而且最为重要的一点是 Google Scholar 完全免费。

实际上,中国引文数据库代表着传统的引文统计分析工具,而 Google Scholar 则代表着新兴的引文统计分析工具。从 Google scholar 与中国引文数据库之间的对比中,可以看到新旧引文统计分析工具之间的差别。对于中文学术期刊的引文统计分析来说,新旧工具各有利弊。传统的引文数据库具有较高的准确性、权威性和规范性,但是涵盖的文献类型有限,不能反映期刊被学术博文、预印本等纯网络学术文献引用的情况。因此,由传统期刊引文数据库统计得出的期刊被引频次并不能真实反映期刊文献在整个学术交流系统中的被利用程度和学术影响力。与之相反,Google Scholar 不仅能够反映学术期刊之间的引用关系,也能广泛反映学术期刊、Web 网页、预印本、电子期刊之间的引用关系。尽管目前这一功能还只是一个理论构想,但相信随着 Google Scholar 在中文文献服务领域的不断发展和完善,Google Scholar 对中文学术文献引用情况的覆盖率将会远远高于传统引文数据库。从本次统计的过程和结果来看,Google Scholar 目前存在着诸多的不足,Google Scholar 的引文检索质量不高,其权威性、准确性和规范性也不及传统的引文数据库。即使在英文环境中,Google Scholar 目前也仍然不能替代传统的引文数据库而成为真正意义上的引文统计分析工具。

通过本次的实证研究和对比分析,同时结合引文统计分析的新需要,笔者认为在期刊评价时选择引文统计分析工具,应该遵从以下几个标准。

其一,全面性。引文数据库的收录量要大,并且能够从期刊、文献、作者等多个角度反映引用关系。笔者认为全面性应该主要从两个方面理解:首先,引文数据库所包含的学科领域要广泛。各个学科之间的交叉融合日益加深,许多研究主题跨越多个学科领域,需要参考借鉴多个学科领域的理论、思想和方法,不同学科之间的交叉引用越来越普遍。如果要真实全面地反映这种跨越学科的引用关系,则必须要建立一个同时涵盖人文社会科学、自然科学、工程技术科学的综合文献数据库。其次,引文数据库所涵盖的文献类型要丰富。学术交流呈现多元化,非期刊类学术文献大量出现,尤其是在网络技术的推动下,网络文献数量持续增长,期刊文献与非期刊类学术文献之间存在着广泛的引用和被引用关系。引文数据库只有涵盖更多的文献类型,才能更加全面真

实地反映期刊文献被利用的程度。

其二,准确性。对数据库中所收录的文献及其引用关系进行准确的整理和统计,保证引文检索的质量,提供一致性较好的数据结果。引文统计分析是一项非常严谨的科学活动,特别是当引文统计分析被用于期刊评价、人才评价等领域时,更是要求引文统计结果具有较高的精确度和稳定性。当某一引文数据库所返回的检索结果中存在重复、遗漏或错误等现象时,无论它的数据库规模有多大,其权威性都要受到质疑。例如,尽管 Google Scholar 所收录的文献数量很大,但在引文检索和分析时所出现的错误却使得其可信度和权威性大打折扣。

其三,及时性。数据库的更新要及时,这样才能更加真实地反映出期刊被利用的情况。特别是在网络环境下,学术文献的出版周期大大缩短,从其出版到被其他文献引用的时间也进一步缩短,随之用户对于引文分析的及时性提出了更高的要求。引文数据库只有提高更新速度,才能及时跟踪文献之间的引用关系。

其四,易用性。引文数据库要具备能够从期刊、论文、作者等多个方面实现多角度、全方位的引文检索和统计分析功能,并且其检索和统计功能要易于操作。在进行期刊评价时,一般要求能够以期刊名称直接检索到该种期刊的总载文量、总被引频次、各年载文量和被引频次等指标。

五、研究的局限性

本研究选择 10 种“图书馆学、情报学和文献学”学术期刊作为研究对象,分别以 Google Scholar 和中国引文数据库作为引文统计工具,统计了这 10 种期刊在 2007 - 2008 年两年间所发表的论文总数和这些论文的总被引频次。基于统计结果的对比,从被引频次的角度出发对 Google Scholar 和中国引文数据库的引文统计分析性能作了一个初步的探讨,以期能够在一定程度上,为日后期刊评价选择引文统计分析工具提供参考依据和选择标准。但由于研究能力和研究条件所限,本次实证研究还存在诸多的局限和不足,使得分析不够全面和深入,结论也缺乏说服力。首先,本次研究只选取了“图书馆、情报和文献学”一个学科领域的期刊作为研究对象,样本的代表性非常有限,难以反映其他学科领域的情况;其次,鉴于条件和精力所限,本次研究只统计了 10 种期刊在两年间的载文量和被引频次,并且只是将两种数据库的统计结果进行对比分析,样本量较小,研究结论的普遍性和说服力不强;再次,只是从被引频次一个角度,对 Google Scholar 和中国引文数据库的

引文统计功能进行了简单的对比,分析深度不够,研究角度也比较单一。

在后续研究中可以从以下几个方面对研究方法和研究过程加以改进。

第一,选择不同学科领域的期刊进行统计和对比,可以分别选择人文社会科学、自然科学、工程技术等多个领域的期刊进行对比分析,以验证本文的研究结论是否适用于其他学科领域。

第二,选择不同的时间段,对不同时间段内的期刊在不同引文数据库中被引频次的统计结果进行历时分析,以验证时间因素是否对研究结论产生影响。

第三,扩大样本数量,选择更多种类的期刊,统计其在更长的时间段内的被引频次,以增强研究结论的说服力。

第四,采用多种统计分析方法,多角度地对统计数据进行了更加深入细致的比较和分析,以寻找新的

或者更加深刻的研究结论。

参考文献:

- [1] GARFIELD E. Citation Index for Science [J]. Science, 1955, 122: 108 - 111.
- [2] GARFIELD E. Citation Analysis as a Tool in Journal Evaluation [J]. Science, 1972, 178: 471 - 479.
- [3] 陶慧卿, 庄琦, 潘卫. 关于 Google Scholar 与 Web of Science 引文分析的实证研究 [J]. 中国索引, 2008(4): 57 - 62.
- [4] 中国引文数据库简介 [EB/OL]. [2010 - 03 - 12]. <http://www.cnki.net/jianjie/jj7.htm>.
- [5] 许涛, 吴淑燕. Google 搜索引擎及其技术简介 [J]. 现代图书情报技术, 2003(4): 58 - 61.
- [6] 化柏林. Google 搜索引擎技术实现探究 [J]. 现代图书情报技术, 2004(3): 40 - 43.

Comparison of the Citations Statistic Analysis Functions between Google Scholar and CCD: Analysis Based on the Journal Citations

QIU Jun-ping, WEN Fang-fang

(Research Center for China Science Education, Wuhan University, Wuhan 430072, P. R. China)

Abstract: Journal Citation is an important index in the journals evaluation. Different citations databases always give different retrieval results. Such difference of citations statistic certainly has considerable influence on the journals evaluation. In this paper, the top 10 journals in the field of "Library, Information and Documentation science" were selected as research subjects. Then we searched each journal's citations during 2007 - 2008 from Google Scholar and CCD (Chinese Citation Database). Based on the comparison of their results, it was found that presently CCD is much better than the Google Scholar for Chinese journals citations statistic but Google Scholar may represent the development trend of citation databases.

Key words: citations database; journals evaluation; citations; citations analysis

(责任编辑 彭建国)