DOI：10.13998/j.cnki.issn1002-1248.2019.05.19-0315

基于中文期刊高被引论文的 Altmetrics 指标评价体系研究

张　瑶

（天津师范大学图书馆，天津　300387）

摘　要：[目的]为科学评价学术论文的社会影响力，以对学术论文影响力进行全面的评价，探索中文环境下 Altmetrics 指标评价体系；[方法]笔者以图书情报与数字图书馆研究领域的文献为研究对象，利用 Python 语言编写网络爬虫程序，追溯相关文献在社交网络平台来源，获取相关数据开展指标研究；[结果/结论]分析并探索了 Altmetrics 指标的数据来源与以及具体指标，初步构建了中文环境下图书情报与数字图书馆研究领域 Altmetrcis 指标评价体系，但由笔者研究学科的限制，针对该指标评价体系对其他学术领域的适用性问题，今后还需进一步开展相关研究来验证和比较。

关键词：高被引论文；Altmetrics；网络爬虫；指标评价体系

中图分类号：G250.73　　**文献标识码：**B　　**文章编号：**1002-1248（2019）05-0037-06

引用本文：张瑶. 基于中文期刊高被引论文的 Altmetrics 指标评价体系研究[J]. 农业图书情报, 2019, 31(5):37-42.

Research on Altmetric Indicators Evaluation system Model Based on Highly Cited Papers in Chinese Journals

ZHANG Yao

(Library of Tianjin Normal University, Tianjin 300387, China)

Abstract: [Purpose] In order to scientifically evaluate the social influence of academic papers and comprehensively evaluate the influence of academic papers, this paper explores the Altmetrics indicator evaluation system under Chinese environment. [Method] This paper takes the literature in the field of Library information and digital library as the research object, uses Python language to compile Internet worm program, traces the origin of relevant literature in social network platform, obtains relevant data and carries out indicator research.[Result/conclusion] This paper analyses and explores the data sources and specific indicators of Altmetrics, and preliminarily constructs an indica-

收稿日期：2019-04-17

基金项目：天津师范大学教育基金项目“中文环境下高被引学术论文替代计量指标评价模型研究”（项目编号：043-135202WT1704）

作者简介：张瑶（ORCID：0000-0002-9215-2866），天津师范大学图书馆，助理馆员，硕士，E-mail：514309080@qq.com。

DOI：10.13998/j.cnki.issn1002-1248.2019.05.19-0315

tor evaluation system of Altmetrics in the field of library information and digital library under the Chinese environment. However, due to the subject limitations, the applicability of the indicator evaluation system to other academic fields still needs further research to verify and compare.

Keywords: high cited paper; Altmetrics; internet worm; indicator evaluation system

学术成果的评价和计量指标问题一直是学术界争论的热点，传统的文献计量指标如论文被引频次、期刊影响因子、个体研究者和科研机构的 h 指数等都是基于引用行为的，将引文分析用于学术成果评价的这类评价方式被认为存在诸多缺陷，如引用时间滞后、引用动机不明（伪引、漏引、自引或互引）等，导致引文缺乏规范性、客观性等各种问题[1]。随着互联网的快速发展，网络出版和网络传播涌入其中，学术成果的类型也日益丰富，一些极具科研价值的学术成果游走于网络之中被广泛传播，因此，这些学术成果的影响力不能仅仅依靠传统的学术评价标准来评判，也需要考虑其社会影响力，该如何对这些学术成果的社会影响力进行科学评价，以弥补学术影响力评价的局部性和片面性，已成为现下的研究热点与难点，所以对传统文献计量指标进行更新、改造势在必行，Altmetrics 也由此诞生。

1 Altmetrics 及其发展状况

Altmetrics 一词最早于 2010 年由北卡罗来纳大学博士生 Jason Priem 提出，他最先在自己的 Twitter 上使用了“Altmetrics”这个单词，随后发表“Altmetrics：a manifesto”，将“Altmetrics”作为正式术语[2]。2012 年，中国医科大学图书馆的刘春丽率先发表《Web 2.0 环境下的科学计量学：选择性计量学》一文，将“Altmetrics”引入中国，该词本是 Alternative Metrics 的缩写，后被译为“选择性计量学”[3]、“替代计量学”[4]或“补充计量学”[5]。而目前“替代计量学”或“补充计量学”这两个译名被应用的更多[6]。自此，国内 Altmetrics 相关研究陆续延展，主要包括：（1）Altmetrics 工具研究：对常用的 Altmetric.com、ImpactStory、Plum Analytics 和 PLoS ALMs 这 4 种常用工具进行全面的分析和比较研究[7]。（2）Altmetrics 应用研究：应用研究主要体现在图书馆服务方面[8-9]与机构知识库的应用方面[10]。（3）Altmetrics 指标研究：如探究 Altmetrics 综合性指标与引用指标之间的相关性与一致性问题[11]等，传统评价指标与 Altmetrics 指标间固然存在一定的相关性与一致性，但程度相对不高；随后，刘晓娟等人基于 PLOS ALMS 数据，对 Altmetrics 指标的可用性进行深入的分析研究，得出 Altmetrics 指标形式多样，还需根据指标自身的特点、适用性等方面来考虑指标的可用性，建立适当的指标评价体系或指标模型[12]；紧接着，一些学者开展 Altmetrics 指标分层研究[13-14]，为 Altmetrics 指标模型的建立提供了前提与基础；赵蓉英等虽设计和构建了基于 Altmetrics 的学术论文影响力评价框架及模型[15]，但该模型还是基于英文环境的或者说是英文社交媒体平台的，并不适用于中文环境。可以看出，国内对 Altmetrics 研究取得了一定的成果，但基本上还是处于理论研究阶段，在实证研究方面，尤其是对 Altmetrics 指标研究方面还有很大的不足，而深入研究中国 Altmetrics 指标内涵的更少，尤其对中文环境下 Altmetrics 指标模式的研究、构建国内 Altmetrics 指标评价体系和开发国内可用的 Altmetrics 工具等方面都还处于探索阶段。所以笔者尝试挖掘中文环境下的社交网络平台，以揭示 Altmetrics 指标的特征和规律。

2 数据及数据处理

2.1 数据来源

探索中文环境下 Altmetrics 指标评价体系，需要选择中文数据库中的文献进行分析研究。目前中国最主要的中文数据库有中国知网（CNKI）、万方数据知识

服务平台和维普资讯中文期刊服务平台，3 个数据库均存在同质化现象，但总体来讲 CNKI 收录的数据相对完整，且对核心期刊的收录较全，收录质量就相对较高[16]。所以笔者选取 CNKI 作为研究数据的来源。

按照 ESI 数据库的界定，将近 10 年来的被引频次降序排序，排名在前百分之一的论文就是高被引论文[17]。由于学科的差异，每个学科内文献的被引频次均不相同，有些学科文献的被引频次普遍较高，而另一些学科文献的被引频次就相对较低，如在 CNKI 中将文献分类选择为“计算机软件与计算机应用”，其中被引频次最高的文献共计被引 3 174 次，而文献分类为“图书情报与数字图书馆”的文献中，最高被引频次仅为 1 418 次。所以高被引论文并不能一概而论，首先要确定检索对象和检索范围。

笔者选取 CNKI 中，文献分类为“图书情报与数字图书馆”这一分类，得到近 10 年（2009 年 1 月 1 日—2018 年 12 月 31 日）发表的文献记录共计 301 381 条，其中被引频次≥ 1 次的文献记录为 2 371 条（检索日期为 2019 年 1 月 1 日）。选取近 10 年来被引频次在前 1%的论文约为 237 篇，其被引频次均＞80 次。经过去重操作，并未发现重复记录。

2.2 数据处理及信息获取

在 Internet 网络中包含着海量的信息，为了能从海量的信息中快速准确地获取到笔者研究数据中的 237 篇文献在网络中的被提及情况，必须要利用计算机网络爬虫的手段。网络爬虫是指通过编写程序，可以实现根据用户提供的检索条件，自动地抓取网络中符合条件的网页 URL 地址，并对抓取结果进行统计分析。笔者的研究就是通过编写网络爬虫程序实现对 237 篇文献数据的处理。

笔者数据处理中使用的爬虫程序是基于 Python 语言[18]开发的，爬虫程序的功能完全按照笔者研究的需求进行设计开发，包含了数据接口模块、爬虫调度模块、爬虫管理模块、网页下载模块和网页分析模块 5 个功能模块。其中，数据接口模块可以实现对 237 篇文献数据的自动遍历，将每一条研究数据处理为一次爬虫任务，并将任务分配给爬虫调度模块；爬虫调度模块根据爬虫管理模块中配置的爬虫对象和爬虫层级等信息，调用网页下载模块、网页分析模块进行网络爬虫，从而实现获取研究数据中的每一篇文献在网络中的被提及情况，并将提及该文献的网页下载，通过分析网页内容，得到相关 URL 的网站出处。

通过利用计算机网络爬虫的方法，对研究数据进行处理和分析，获得了 237 篇文献信息在网络中被提及的详细情况，为 Altmetrics 指标评价体系的研究和构建提供了研究依据。

3 Altmetrics 指标评价体系的构建

3.1 高被引论文数据来源分析

通过将文献数据中的题目和作者作为网络爬虫抓取的条件，以百度（www.baidu.com）为目标网站，获取其搜索结果页面的 URL 并进行解析分析。在百度搜索的结果中，排名靠前的搜索结果的相关性要高于排名靠后的，为提高爬虫效率和有效性，笔者仅对百度搜索结果的前 30 条进行抓取。将爬虫层级设定为两层，一层是自动抓取百度搜索结果列表页的 URL，另一层是依次抓取列表页内各个搜索结果的 URL，经过上述的两层挖掘，可以抓取到相关搜索结果页面的 URL，通过对 URL 的域名字段进行解析，可以得出相应的网站出处。在完成所有文献数据的网络抓取操作后，以网站出处作为统计分析的分组标签，分别统计在其中追溯到的文献篇数，并计算该网站追溯到的文献数在本次研究文献总数中的占比情况。最终得出高被引论文在网络平台的被提及情况统计表，如表 1 所示。

通过上述得到的高被引论文在网络平台的被提及情况统计表的结果可以发现，百度学术（中英文文献检索的学术资源搜索平台，涵盖了各类学术期刊、会议论文）、百度文库（在线分享文档的平台）、豆丁网（中文社会化阅读平台）、道客巴巴（在线分享文档的平台）这 4 个网站平台中被追溯到的论文篇数最多，覆盖率相对较高，所以这 4 个平台是获取学术成果社

DOI：10.13998/j.cnki.issn1002-1248.2019.05.19-0315

表 1　高被引论文在网络平台的被提及情况统计表

追溯到网站	追溯到的文献篇数/篇	占高被引论文的百分比/%
百度学术	199	83.97
百度文库	165	69.62
豆丁网	100	42.19
道客巴巴	84	35.44
新浪博客	7	2.95
淘豆网	5	2.11
新浪爱问共享资料	5	2.11
中国社会科学网	3	1.27
科学网博客	3	1.27
CSDN 博客	2	0.84
中国图书馆网	2	0.84
中华文本库	2	0.84
优酷视频	1	0.42

会影响力的主要数据来源平台，其相关指标可以作为 Altmetrics 主要参考评价指标。百度学术、百度文库、豆丁网和道客巴巴 4 个网站平台，其网站的主要功能就是学术交流和资料共享，因此在理论上，这 4 个网站平台对学术成果的社会影响力应该要高于其他类型的网站，笔者研究的统计结果与实际也是相符合的。而在这 4 个网站平台中，百度学术的占比最高，也说明了学术性越强的网站平台，学术成果的社会影响力也就越高。

新浪博客（国内主流的社交网站）、淘豆网（文档在线分享和销售网站）、新浪爱问共享资料（在线资料分享网站）、中国社会科学网（社会科学类学术网站）、科学网博客（学术社交平台）、CSDN 博客（中国 IT 社区和服务平台）、中国图书馆网（图书馆专业网站）、中华文本库（科研、教案等相关文档分享网站）以及优酷视频（视频网站），这些平台的高被引论文的覆盖率虽然较低，但其相关指标也可作为辅助参考评价指标。上述平台中的占比排序，在一定程度上受研究分析领域的影响，网站平台本身拥有的用户体量越大、用户对研究领域的关注度越高，其占比排名也就越高。笔者主要以图书情报与数字图书馆研究领域的高被引论文作为研究对象，统计排名情况只能代表各社交网站平台对该领域学术成果的社会影响力情况。

至于新浪微博，笔者在其中并未检索到相关高被引文献，这可能也是源于学科的差异，因为有些学科大众的关注度并不高。但早在 2014 年，Altmetric.com 就与其开展合作，追踪其对学术论文的关注度，并将相关指标其作为分析中文数据的主要来源，虽然也有数据指出其 Altmetrics 指标的覆盖率不足 1%[19]，然而作为参考评价指标还是要予以考虑的。

3.2　数据来源平台指标分析

通过上述方法进行网络爬虫数据获取并对相关数据进行统计分析，可以获得 Altmetrics 指标来源的网站平台，这些平台来源众多且类型各异，因此要构建适用于中国中文期刊的 Altmetrics 指标评价体系，还需要对这些平台上的指标数据进行进一步的分类汇总，不同网站平台其支持和包含的评价指标也有一定的区别。按照高被引论文被提及的数据来源网站平台的类型可进行如下分类（见表 2），主要可分为学术平台、在线文档分享库、社交平台、视频网站和相关学科网站 5 个类别：

（1）学术平台：主要代表就是百度学术，它是由百度提供的一个中英文学术资源检索平台，可以为学术型网站提供大众服务的快捷通道，百度学术并不直接提供学术文献的全文，而是链接到其他专业学术平台（如 CNKI、万方数据等），以获取文献的全文。相对于专业的学术平台，百度学术是一个综合性的检索平台，同时也是一个学术共享平台（可以分享到 QQ、微信、微博等社交媒体上）。

（2）在线文档分享库：包括百度文库、道客巴巴、豆丁网、淘豆网、新浪爱问共享资料等类似的文库，其平台提供各种文档（包括学术期刊论文）资料的全文浏览、阅读和下载等功能，也提供了一些收藏、点赞、分享、转发以及评论的功能。

（3）社交平台（包括学术社交平台）：主要涉及新浪博客、新浪微博、CSDN 博客、科学网博客（学术社交平台）等网站，这些网站不止包含阅读量等数据，更重要的是存在大量的评论，其最主要的指标应是被提及的次数。

表 2 数据来源及其指标分析表

数据平台类型	数据来源	可直接获取指标	不可直接获取指标
学术平台	百度学术	阅读量、被引量	收藏量、分享量
	百度文库	浏览量、下载量、评分数	点赞量、收藏量、转发量
	道客巴巴	浏览量	点赞量、下载量、收藏量、分享量
在线文档分享库	豆丁网	浏览量、点赞量、评价	分享量、收藏量、下载量
	淘豆网	收藏量	下载量
	爱问共享资料	阅读量、下载量、评论	无
社交平台	新浪博客	阅读量、收藏量、评论	转载量、分享量
	新浪微博	阅读数、评论量、转发量、点赞量	收藏量
	CDSN 博客	阅读数、点赞量	收藏量、分享量、评论
	科学网博客	阅读量、推荐量、评论	分享量、收藏量
视频网站	优酷视频	无	收藏量、下载量、分享量
相关学科（专业）网站	中国社会科学网	分享量、推荐量、浏览量	无
	中国图书馆网		

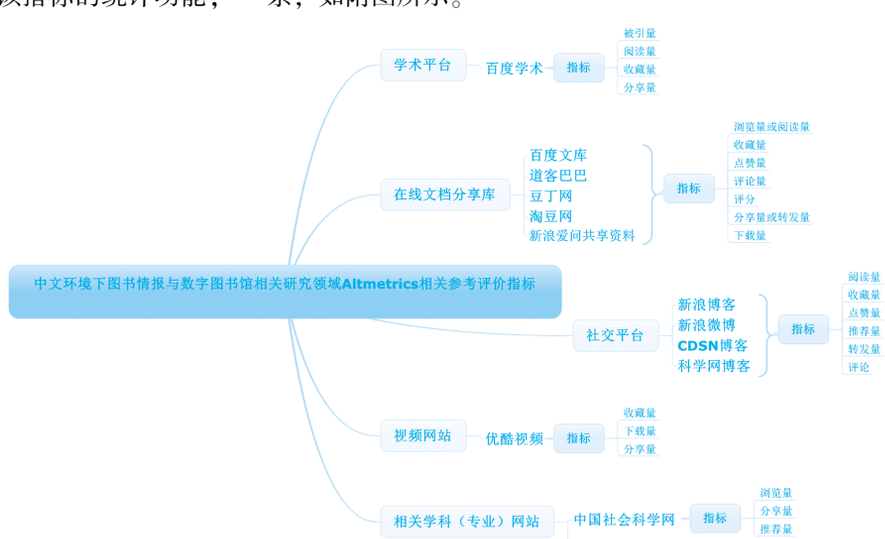
（4）视频网站：如优酷视频等。主要以视频的收藏、下载和分享指标为主。

（5）相关学科（专业）网站：由于笔者采用的高被引论文均源自于图书情报与数字图书馆方向，所以得到如中国社会科学网，中国图书馆网这样的相关专业领域的网站平台。如检索其他学科网站将会得到各种不同的结果，具体研究指标与具体学科领域相关。

按照网站平台的分类，对各个平台上的指标参数展开调查分析，可直接获取的指标就是可以从网站平台直接获得的指标参数，即网站平台本身已经对用户提供了该指标的统计情况，具备对该指标的统计功能；不可直接获取的指标则是虽然网站平台提供了这些功能，但并未进行总数统计或未将统计结果公布在网站上的，也就是说其参数值暂时还不能从网站平台上直接获得，但这些指标仍是计量文献社会影响力的重要参考指标。

以表 2 数据来源及其指标分析表的统计结果来看，浏览量或阅读量这类反应学术成果在网站平台被关注次数的指标几乎各大网站平台都具备，因此，该指标也是高被引论文 Altmetrics 指标评价体系中最重要最基本的一个指标。而例如被引量、收藏量、点赞量、推荐量等可以反馈读者评价的指标，则可以进一步体现学术成果的社会影响力，其指标值越高，代表学术成果的社会影响力、学术成果的被认可率也就越高，是学术成果在高被引论文 Altmetrics 指标评价体系最主要的加分项。

综上，可以初步得到中文环境下图书情报与数字图书馆研究领域高被引论文 Altmetrics 指标评价参考体系，如附图所示。



附图 中文环境下图书情报与数字图书馆领域高被引论文 Altmetrics 指标评价体系

DOI：10.13998/j.cnki.issn1002-1248.2019.05.19-0315

4 结语

笔者选取 CNKI 中，文献分类为“图书情报与数字图书馆”的高被引论文为研究数据，通过计算机网络爬虫的方法获取其在各网站平台的被提及情况，并对其网站出处和评价指标进行分析研究，初步建立了中文环境下高被引论文 Altmetrics 指标评价体系。

笔者的研究只是对高被引论文 Altmetrics 指标评价体系的初步构建，还有很多方面需要进一步展开研究。首先，笔者只是在图书情报与数字图书馆研究领域初步获得了中文环境下 Altmetrics 指标评价体系，想要获得整个中文环境高被引论文 Altmetrics 的指标还需要对更广泛的研究领域开展调查分析研究。其次，笔者并没有对这些指标相关性等问题进行深入分析，如分析指标的用户群体特征和用户行为特征，确定指标的权重，对指标的效度和信度进行分析。所以，笔者未来的研究方向将朝着以上两个方向进行尝试，这样才能使 Altmetrics 在中文环境下发挥其对学术论文评价的真正作用和价值。

参考文献：

[1] 杨思洛.引文分析存在的问题及其原因探究[J].中国图书馆学报,2011,37(193):108-117.

[2] J. Priem, D. Taraborelli, P. Groth, C. Neylon (2010), Altmetrics: A manifesto, 26 October 2010 [EB /OL].[2019-02-15].http://altmetrics.org/manifesto.

[3] 刘春丽.Web 2.0 环境下的科学计量学:选择性计量学[J].图书情报工作,2012,56(14):52-56.

[4] 顾立平. 开放数据计量研究综述: 计算网络用户行为和科学社群影响力的 Altmetrics 计量[J]. 现代图书情报技术,2013,(6):1-8.

[5] 由庆斌,汤珊红. 补充计量学及应用前景[J]. 情报理论与实践,2013,36(12):6-10.

[6] 余厚强,任全娥,张洋,刘春丽. Altmetrics 的译名分歧: 困扰、影响及其辨析[J].中国图书馆学报 ,2019,45(239):47-59.

[7] 王睿,胡文静,郭玮.常用 Altmetrics 工具比较[J].现代图书情报技术,2014,(12):18-26.

[8] 余厚强,邱均平.论替代计量学在图书馆文献服务中的应用[J].情报杂志,2014,33(9):163-172.

[9] 赵雅馨,杨志萍.研究热点探测的替代计量学方法和应用—以信息与计算机科学为例[J].情报杂志,2016,35(11):39-44.

[10] 邱均平,张心源,董克.Altmetrics 指标在机构知识库中的应用研究[J].图书情报工作,2015,59(2):100-105.

[11] 由庆斌, 汤珊红. 不同类型论文层面计量指标间的相关性研究[J].图书情报工作,2014,58(8):79-84.

[12] 刘晓娟,宋婉姿.基于 PLOS ALM 的 altmetrics 指标可用性分析[J].图书情报工作,2016,60(2):93-101.

[13] 余厚强,邱均平. 替代计量指标分层与聚合的理论研究[J].图书馆杂志,2014,33(10):13-19.

[14] 邱均平, 余厚强. 基于影响力产生模型的替代计量指标分层研究[J].情报杂志,2015,33(5):53-58.

[15] 赵蓉英,郭凤娇,谭洁. 基于 Altmetrics 的学术论文影响力评价研究—以汉语言文学学科为例[J].中国图书馆学报,2016,42(221):96-108.

[16] 杨慕莲.数字环境下中文学术信息资源的比较与分析—以 CNKI、万方、维普三大期刊全文数据库为例[J].湖北科技学院学报,2013,33(11):190-191.

[17] 高被引论文[EB /OL].[2019-01-20].https://baike.so.com/doc/8499622-8819923.html

[18] Python [EB /OL].[2019-02-11].https://baike.baidu.com/item/Python/407313?fr=aladdin.

[19] 余厚强,Bradley M.Hemminger,肖婷婷,邱均平.新浪微博替代计量指标特征分析[J].中国图书馆学报,2016,42(224):20-36.