中国科技资源导刊 ISSN 1674－1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2009 年 1 月 第 41 卷第 1 期 46－51 ISSN1674－1544 Vol. 41 No. 1 46－51,Jan. 2009

基于 h 指数的核心作者遴选方法的比较研究

周春雷
(武汉大学科学评价研究中心,湖北武汉 430072)

摘 要:“核心期刊高发文量和 h 指数相结合”是一种具有优势的评选高影响力作者的有效方法,但传统的 h 指数手工统计方法制约了该方法的应用。为此,本文提出一种适合程序处理的、基于领域专业期刊被引信息的 h 指数统计新方法,并以图书情报领域为例进行了实证研究。经与手工统计结果的对比证明了该方法的优越性。

关键词:h 指数;图书情报领域;核心作者;实证研究;论文引用;文献计量
中图分类号: G350 文献标识码: A DOI: 10.3772/j.issn.1674－1544.2009.01.007

1 引 言

“核心作者”虽然是文献计量学的专业术语,但其测量方法并没有统一的标准。我国图情界学者在核心作者评价方面进行过一系列研究,但对领域核心作者遴选方法并未达成共识。刘东维(1986)选择我国情报学界较有影响的 7 种学术期刊,根据各期刊自创刊之日到 1985 年底的发文篇数、被引证篇次数、平均被引率和基础文献发文数的综合评价指标,用定量化的方法确定了我国情报学研究领域的 33 位核心著者[1]。郦金花和苏新宁(2004)通过对中国社会科学引文索引(CSSCI)1998－2002 年间的图书馆学情报学论文的统计分析,分别得出该领域发文最多以及被引最多的前 32 位作者[2]。方太强、周蓉、胡英(2005)根据发文篇数、被引证篇(次)数和重要文献发文数等因素为指标的综合评价体系,确定了我国图书情报学研究的 78 名核心作者[3]。马费成、宋恩梅(2006)综合其他学者的核心著者评价研究成果,确定了 37 位情报学核心作者[4]。李彩云(2007)统计了《情报科学》1998－2005 年间的核心作者,采用普赖斯公式 N = 0.749 * sqrt(最高产作者的发文量)规定了核心作者候选人的最低发文数,并以第一作者发文量和被引量为指标进行加权计算,得出各候选人的综合指数,最后据此综合指数得出《情报科学》的核心作者[5]。

2 已有的核心作者遴选方法评价

不难看出,虽然这些学者所采用的基本思路都是试图通过综合利用发文和被引信息来确定图情领域的核心作者,但由于所取样本、时间段、评价指标等差异,所得到的核心作者名单各不相同。由于缺乏统一的遴选标准,核心作者遴选的可操作性及结果的客观性、可信性均难以保证。美国统计物理学家赫希(Jorge E. Hirsch)教授 2005 年提出了 h 指数,即一个人的 h 指数是指其至多有 h

作者简介:周春雷(1977－),男,系统分析师,讲师,武汉大学信息管理学院情报学博士研究生,研究方向是知识管理与科学评价。

收稿日期:2008 年 11 月 22 日。

篇论文分别被引用至少 h 次[6]。由于 h 指数能够综合反映作者的发文和被引信息，一经提出即在国际上引起很大反响。普遍认为 h 指数能够较好地评价优秀学者的终身成就。邱均平、缪雯婷(2007)[7]、张学梅[8]等人对国内部分图书情报学者 h 指数的统计表明，那些在图情界有较高影响力的作者的 h 指数要明显高于普通作者。因此，笔者认为，基于 h 指数的评价可以较好地解决核心作者遴选问题，我们可以将某个 h 指数值作为核心作者的标准，不同领域、不同时期的评选标准可以不同。

笔者曾提出“核心期刊高发文量和 h 指数相结合”是一种具有优势的评选高影响力作者的有效方法，进行了规模较大的实证研究，并建议用 h≥5 作为国内图情界高影响力作者的参考尺度。与传统的基于高发文量的核心作者评选方法相比，该方法能剔除那些虽发文多却不为同行所看重的作者；与传统的基于高被引的核心作者评选方法相比，该方法能剔除那些偶有被引很高佳作的低产作者；与单纯的 h 指数方法相比，该方法能筛选出主要研究领域非研究者所关心领域的跨学科高影响力作者；与专家评审方法相比，本方法具有操作简单、客观、准确等优势，在评价效率和所花费的代价方面也具有较大的优势[9]。该方法作为一种能综合发文和被引信息的新的核心作者评选方法虽具有种种优势，但有依赖手工统计 h 指数这一劣势。在进行该项研究时，笔者尽管耗时月余也不过精确统计了 1241 人的 h 指数。

用 h 指数来评选核心作者首先需要了解整个研究领域所有作者的 h 指数分布情况，然后才能根据分布情况确定合适的阈值，其中准确获取领域作者的 h 指数是其关键。但对传统 h 指数统计方法来说，快速、准确地测量大范围领域作者的 h 指数被普遍认为是枯燥、繁重且易错的工作，手工检索是其难以逾越的瓶颈。笔者曾指出“核心期刊高发文量和 h 指数相结合”的方法虽保证了高 h 指数作者统计的准确性，但尚无法揭示大量低 h 指数作者的分布情况，对低 h 指数作者分布情况的准确调查有待其他研究方法的出现[9]。总之，落后的研究手段已成为制约 h 指数研究发展的重大障碍。为此，本文提出一种新的 h 指数统计方法，统计出国内图情领域各作者的 h 指数，并根据 h 指数分布情况确定了该领域的核心作者。

3　基于领域期刊引文信息的 h 指数统计方法

3.1　h 指数传统统计方法的不足

利用赫希教授对 h 指数的定义只能进行单个作者 h 指数的统计，不能快速、准确地测量某个研究或学科领域全部作者 h 指数。现有的 h 指数统计流程一般为选择某个引文数据库，如 Web of Science、Scopus、Google Scholar 以及国内的 CSSCI 等，按照某个名单，手工逐一检索作者被引信息并按被引次数降序排列，统计出各作者的 h 指数。这种方法的弊端如下：

(1) 无法区分不同领域同名作者的引文信息。各引文数据库均包含了不同领域研究者的信息，采用标准 h 指数统计方法得到的 h 指数往往因混杂了同名作者的引文信息而被夸大。

(2)抽样的代表性无法保证。传统的 h 指数统计一般依据某个名单由人工逐一进行，这个名单可能来自学者们统计出的领域核心作者，也可能来自期刊的编委列表等。这种统计方法仅是对整个研究领域的抽样，其代表性难以保证，普通研究者由于人数众多，难以出现在这些名单中。

(3)不适合大范围的快速统计。传统方法仅适合小范围的手工统计，是枯燥、繁重、易错的工作，不适合涉及成千上万作者的整个研究领域大范围、快速、精确统计。

3.2　一种基于引文信息的 h 指数统计方法

为解决上述弊端，本文提出了一种基于领域期刊引文信息的 h 指数统计方法，设想将统计范围限制在领域专业期刊。具体方法是首先选择某个引文数据库，接着用被引期刊名称进行检索并汇集全部领域期刊的引文信息，然后按照被引作者和被引次数排序，最后使用程序从中自动获得各作者的 h 指数。

表 1 示例了按作者姓名和被引次数降序排列的引文数据汇总表。通过使用自编程序分析“被引作者”和“被引次数”两列数据，自动得出各作者相应的 h 指数(表 2)。

这种方法的关键在于领域专业期刊的选取。

表 1　将全部专业期刊引文信息汇总按作者和被引次数排序的引文数据示例

被引作者	被引文献篇名	被引次数	被引期刊	被引文献发表时间
艾静	关于公共图书馆跨世纪发展的思考	4	图书馆理论与实践	2000(3)
艾露	超文本在情报检索中的应用	2	图书馆学刊	1998(6)
艾露	梁启超目录学思想与实践研究综述	1	国家图书馆学刊	1999(1)
艾冰	图书馆自动化建设中的机读目录	1	晋图学刊	1998(3)

表 2　作者 h 指数计算

被引作者	被引次数降序列表	h 指数
艾静	4	1
艾露	2	1
	1	
艾冰	1	1

由于各领域都有公认的专业期刊，而且这些期刊名单可以从重要学术数据库的分学科期刊列表获得，所以这个问题不是本方法应用的障碍。根据文献计量学常识可知，任何研究领域都有该领域公认的“专业期刊”，绝大多数与该领域有关的文献都发表在这些专业期刊上。因此，笔者认为，通过将某领域全部专业期刊的被引信息汇总，可以得到涵盖该领域绝大多数研究者成果的数据，通过对这些数据的分析可以得到比较准确的该领域全部作者的 h 指数。这一假设将在下文的实证研究中得到证实。此外，由于引文数据库主要是基于核心期刊所载论文所附的参考文献建立起来的，而发表于非核心期刊的论文也可能为核心期刊上的论文所引用，仅采用来自核心期刊的引文信息是否能较好地反映作者的 h 指数也是本文关心的问题。因篇幅所限，有关该方法的详细论述笔者将另文介绍。

3.3　基于引文信息的 h 指数统计方法的优点

(1)减少同名作者引文信息混杂现象。由于现有引文数据库均未妥善解决作者唯一标识问题，来自不同领域的同名者其被引信息往往混杂在一起，导致作者 h 指数被夸大。本文所提出的方法将被引文献限制在领域专业期刊，虽然依然无法区分同一研究领域中的同名者，但能够剔除其他领域的同名者降低同名者出现的概率，从而有利于提高作者 h 指数统计的准确性。

(2)大大提高统计效率。与某领域的研究者数量相比，该领域的期刊数量相对要少得多。因此，与以作者为单位进行统计的方法相比，采用本文提出的统计方法将大大减少所需的查询次数。以图情领域为例，根据笔者的统计，近 30 年来核心期刊第一作者人数为 30274，仅发文量在 5 篇及以上的第一作者高达 3911 人；而图情领域共有期刊 73 种，南京大学版核心期刊 20 种。相对于这 3 万多作者逐一进行检索显然不如以期刊为单位进行检索经济而方便。使用笔者自编程序对期刊被引信息进行分析，可以大大提高统计效率。

(3) 减少遗漏，提高统计准确性。本文所介绍的方法不仅可以准确统计出大量在手工统计中被忽略的普通作者的 h 指数，而且能较好地涵盖本领域的重要研究者。

3.4　基于引文信息的 h 指数统计方法的不足

本文所介绍方法的准确性在很大程度上依赖于领域数据的完备程度。由于统计范围限于领域专业期刊，这可能会遗漏那些发表在非本领域专业期刊上的成果和以专著等形式发表的成果，从而导致作者 h 指数的降低。

4　国内图情领域核心作者实证研究

4.1　研究方法

从 CSSCI 分别获取 73 种国内图情专业期刊的被引信息，汇总后按被引篇名和被引作者进行排序并将被引次数合并，从而得到该领域所有研究者所发表专业文章的被引信息，然后分别以被引作者和被引次数为第一、第二排序依据进行降序排列，最后采用笔者自编软件统计出该领域所有作者的 h 指数。为验证前文提及的核心期刊引文对作者 h 指数研究的代表性，笔者还抽取了南京大学版 20 种图情核心期刊的信息进行了对照研究。笔者曾通过多种途径广泛搜集了人数多达数千人的图情领域知名学者名单，利用 CSSCI 逐一统计了其 h 指数，其中精确统计了 1241 人[9]，其结果被用于检验本文所提方法统计 h 指数的准确性。本文的检索时间是 2008 年 3 月，受 CSSCI 引文数据库的限制，检索时间跨度为 1998－2006 年。

4.2　研究结果

国内图情领域作者 h 指数与相应人数分布情况如图 1 所示。从图 1 可以看出，随着 h 指数的升高，相应的作者人数锐减。换言之，高 h 指数作者

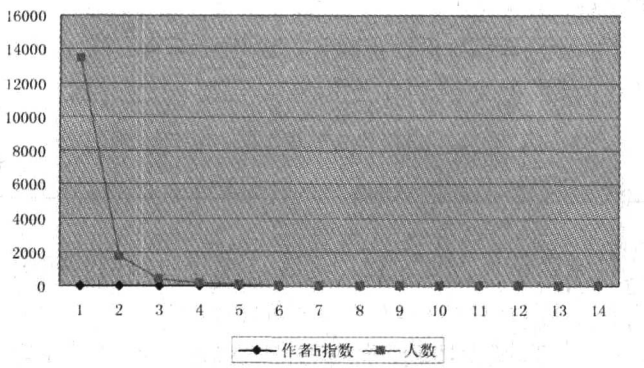

图 1　图情领域作者 h 指数与作者人数分布

人数占总作者比例很小。h≥5 的作者共有 102 人，在全部作者 15790 中所占比例仅为 0.646%。h≥5 的绝大多数作者都具有高级职称，在业内的知名度较高。考虑到有的作者虽然发表了论文，但并没有被引用的情况，全部作者实际人数应该大于 15790，按照对 30 年来图情领域 20 种核心期刊所发文章第一作者的统计，作者人数至少有 30274 人 [9]。则 h≥5 的作者所占的比例将进一步降低到 0.337%。因此，笔者认为，根据 CSSCI 在 1998－2006 年间的引文数据，可以把图情领域核心作者的 h 指数门槛设为 5。当然，这一门槛不是绝对的，研究者可以根据不同研究领域、不同时期的实际情况进行调整。

表 3 为图情领域高 h 指数部分作者名单，为节约篇幅，仅列出 h≥7 的作者。其中“手工统计”是指笔者在文献[9]中统计出的作者 h 指数；“全部”是指基于全部 73 种图情期刊统计出的作者 h 指数；“核心”是指基于 20 种南京大学版图情核心期刊统计出的作者 h 指数；“手工－全部”是指手工统计结果与基于全部期刊结果的差值，其他类推。从表 3 可以发现，基于核心期刊的 h 指数与基于全部期刊统计的相差不大，但与手工统计的数值有较大偏差。

4.3　研究结果分析

基于全部领域专业期刊的程序自动提取方法与手工统计两种方法所得的 h 指数与人数分布对比如表 4 所示。笔者使用自编程序对两种方法获得的名单进行了对比，其 h 指数差异情况如表 5 所示。

通过仔细对比发现，绝大多数采用本文方法所得的 h≥5 的核心作者均出现在手工统计的 h≥5 作者名单中，但有 2 名 h＝5 的作者没有出现在手工统计的名单中。这说明尽管笔者尽可能全地搜集了图情领域知名学者名单，但依然难以避免有遗漏。这

表 3　部分高影响力图情领域专家的 h 指数及其不同统计方法的差异

作者	手工统计	全部	核心	手工－全部	手工－核心	全部－核心
邱均平	16	14	14	2	2	0
吴慰慈	14	12	12	2	2	0
张晓林	14	13	12	1	2	1
马费成	13	9	9	4	4	0
胡昌平	12	10	10	2	2	0
黄宗忠	12	11	9	1	3	2
蒋永福	12	12	12	0	0	0
张琪玉	11	7	7	4	4	0
吴建中	11	8	8	3	3	0
肖希明	11	9	9	2	2	0
黄俊贵	11	10	10	1	1	0
彭斐章	10	7	6	3	4	1
王知津	10	8	8	2	2	0
包昌火	10	8	7	2	3	1
王子舟	10	9	9	1	1	0
范并思	10	9	9	1	1	0
马海群	10	9	9	1	1	0
盛小平	10	10	10	0	0	0
来新夏	9	1	1	8	8	0
谢　康	9	3	3	6	6	0
董小英	9	4	4	5	5	0
卢泰宏	9	5	5	4	4	0
乌家培	9	6	5	3	4	1
王崇德	9	6	6	3	3	0
陈光祚	9	7	7	2	2	0
程亚男	9	7	7	2	2	0
霍国庆	9	9	9	0	0	0
王世伟	9	9	9	0	0	0
王重民	8	1	1	7	7	0
苏新宁	8	5	5	3	3	0
柯　平	8	5	4	3	4	1
叶继元	8	6	6	2	2	0
岳剑波	8	6	6	2	2	0
汪　冰	8	6	6	2	2	0
谭祥金	8	6	6	2	2	0
黄晓斌	8	7	7	1	1	0
初景利	8	7	7	1	1	0
李国新	8	7	7	1	1	0
程焕文	8	8	8	0	0	0
邹志仁	7	3	3	4	4	0
周文骏	7	4	3	3	4	1
谢新洲	7	4	4	3	3	0
刘　嘉	7	4	4	3	3	0
孟广均	7	5	5	2	2	0
靖继鹏	7	5	5	2	2	0
焦玉英	7	5	5	2	2	0
赖茂生	7	5	5	2	2	0
于良芝	7	5	4	2	3	1
陈传夫	7	6	6	1	1	0
严怡民	7	6	5	1	2	1
莫少强	7	6	6	1	1	0
朱　强	7	6	4	1	3	2
王　波	7	6	5	1	2	1
马恒通	7	7	6	0	1	1
叶　鹰	7	7	7	0	0	0
查先进	7	7	7	0	0	0
刘兹恒	7	7	7	0	0	0
刘植惠	7	7	7	0	0	0
赵继海	7	7	7	0	0	0
杨宗英	7	7	7	0	0	0
肖　珑	7	7	7	0	0	0
索传军	7	7	7	0	0	0
马文峰	7	7	7	0	0	0
杨文祥	7	7	7	0	0	0

表 4　基于全部期刊与手工统计所得领域作者 h 指数分布情况对照

h 指数	作者人数	
	全部期刊	手工统计
1	13438	151
2	1719	393
3	387	389
4	144	183
5	46	84
6	20	39
7	17	25
8	4	11
9	7	10
10	3	7
11	1	4
12	2	3
13	1	1
14	1	2
16	0	1
注释	H≥5 人数为 102	h≥5 人数为 187
		h≥5 人数为 103

表 5　基于全部期刊与手工统计所得 h 指数差异对照

h 指数差异	人数	备注
0	46	相同
–1	29	基于全部期刊比手工统计少数 1
–2	16	基于全部期刊比手工统计少数 2
–3	7	基于全部期刊比手工统计少数 3
–4	2	基于全部期刊比手工统计少数 4
5	2	手工统计没有发现的

也证实了本文方法具有发现被忽视的核心作者的功能。

通过对比还发现，手工统计的 h≥5 的部分作者并没有出现在采用本文方法统计的 h≥5 作者名单中。其原因是：(1) 个别手工统计的作者主要研究领域并非图情领域，这证明了本方法具有过滤非本领域研究者的优势。(2) 手工统计得到的 h 指数不小于 5，但采用本文方法的 h 指数小于 5，这是因为：其他领域存在同名者；某些作者的部分高被引文献来自专著或网络文献；个别 CSSCI 引文数据著录不一致导致程序无法像人工那样汇总被引次数，从而影响了作者的 h 指数。

如上所述，基于核心期刊的 h 指数虽因引文类型、领域等外部因素而与手工统计的数值有较大偏差，但与基于全部期刊统计的相差不大。因此，基于核心期刊的分析可以较好地代表对全部领域期刊的分析。换言之，本文所介绍的基于领域核心期刊的 h 指数批量统计法可以取代笔者此前提出的“核心期刊高发文量和 h 指数相结合”的方法而应用于领域核心作者评选。

总之，对比实验很好地验证了本文的假设，证明在限制引文类型为期刊的前提下，通过对某领域全部专业期刊被引信息的汇总和分析，可以得到比较准确的该领域全部作者的 h 指数；仅用领域核心期刊来统计 h 指数即可保证较高的准确性，这可以进一步提升本文方法的研究效率。

5　结　语

h 指数被普遍认为可以用于对优秀学者进行学术成就评价，但 h 指数手工统计方法制约了将 h 指数引入核心作者遴选的尝试。本文提出了一种适合程序处理的新的 h 指数统计方法，从而使快速、准确地统计某领域全部作者的 h 指数成为可能。本文对图书情报领域的实证研究表明，使用这种基于领域专业期刊引文数据的 h 指数统计法可以得到准确的全领域作者 h 指数分布数据，从而为确立核心作者标准提供了重要依据。与传统的基于单纯发文量、被引或专家评审的方法相比，该方法具有操作简单、客观、准确等优势，在评价效率和所花费代价方面也具有较大的优势。总之，本文所提出的 h 指数统计新方法是对笔者以前提出的“核心期刊高发文量和 h 指数相结合”的核心作者评选方法[9]的重大改进，对提高人物学术成就评价的效率以及核心作者、领域专家遴选等评价活动的科学性有一定的积极意义，对科研人才库建设也会有一定的启发。

参考文献

[1] 刘东维．我国情报学基础文献和核心著 [J]．情报科学, 1986(4): 9–16.

[2] 郦金花，苏新宁．近 5 年我国图书馆学情报学研究之影响 [J]．情报学报, 2004(5): 515–523.

[3] 方太强，周蓉，胡英．我国图书馆学情报学核心作者分析 [J]．图书情报工作, 2005(1): 69–73.

[4] 马费成, 宋恩梅．我国情报学研究分析: 以 ACA 为方法 [J]．情报学报, 2006(3): 259–268.

[5] 李彩云．《情报科学》1998–2005 核心作者测评 [J]．情报科学, 2007(2): 236–239.

[6] J E Hirsch. 衡量科学家个人成就的一个量化指标 [J]．科

学观察, 2006(1): 2 – 7.
[7] 邱均平, 缪雯婷 . h 指数在人才评价中的应用——以图书情报学领域中国学者为例[J]. 科学观察, 2007(3): 17 – 22.
[8] 张学梅 . 用 h 指数对我国图书情报学界作者进行评价[J]. 图书情报工作, 2007(8): 48 – 50.
[9] 邱均平, 周春雷 . 发文量和 h 指数结合的高影响力作者评选方法研究[J]. 图书馆论坛, 2008(6): 44 – 49.

Comparative Study of Core Authors Selecting Method Based on H – Index in a Field

Zhou Chunlei
(Research Center for Chinese Science Evaluation, Wuhan University, Wuhan 430072)

Abstract: The method using the amount of a author ' s published papers and his h – index to select high influence authors was proved efficient and advantageous before, but the traditional h – index counting method by hand handled its application. A new h – index counting method is introduced which is based on the citation information of all the journals in a field and can be dealt by software to solve it. By applying the new method to library and information field, the advantages is proved by contrasting with the traditional method by hand.

Keywords: H – index, library and information field, core authors, case study, artcle citation, bibliometrics