h指数与h型指数研究

叶 鹰 唐健辉 赵 星 等 著

科学出版社

h指数与h型指数研究

Studies on the h-index and h-type indices

叶 鹰 唐健辉 赵 星等 著

科学出版社
北京

内 容 简 介

本书是研究 h 指数和 h 型指数的一部专著。在阐明 h 指数和 h 型指数来龙去脉的基础上，对 h 指数和 h 型指数的理论机理、实证研究和应用研究进行了系统探索。这是在国家自然科学基金项目“h-指数与类 h-指数的机理分析与实证研究”（批准号：70773101）资助下独立研究并结合国内外进展创作的国内有关 h 指数和 h 型指数研究的第一部专著。

本书可作为科研管理与评价、科技政策、科学计量学、信息计量学、文献计量学、图书情报与档案管理等相关领域研究者和工作者的业务参考用书，也可作为信息资源管理、信息管理与信息系统、科学学与科技管理、图书馆学、情报学与文献学等相关专业的本科生和研究生的专题教材。

图书在版编目(CIP)数据

h 指数与 h 型指数研究 / 叶鹰，唐健辉，赵星等著．—北京：科学出版社，2011

ISBN 978-7-03-030263-2

Ⅰ．h… Ⅱ．①叶… ②唐… ③赵… Ⅲ．计量-管理-研究 Ⅳ．TB9

中国版本图书馆 CIP 数据核字（2011）第 021783 号

责任编辑：李 敏 赵 鹏 / 责任校对：何艳萍

责任印制：钱玉芬 / 封面设计：耕者设计

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2011 年 3 月第 一 版 开本：B5（720×1000）

2011 年 3 月第一次印刷 印张：13 1/2

印数：1—2 000 字数：263 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

目　　录

第二篇 理论机理

第三篇 实证研究

第四篇　相关研究

第五篇　延伸研究

引　言

2005年美国物理学家J. E. Hirsch引进的h指数，以简单稳健的特点引起学术界的广泛关注，很快成为学术评价新指标和信息计量学研究热点。

学术界之所以对h指数如此感兴趣，除了h指数是平衡产出（output）与影响（impact）的简洁测度外，在很大程度上还因为该简单参数蕴涵了丰富的信息计量学研究信息，可供信息计量学深入研讨和借鉴发挥，激发了信息计量学的研究活力，进而使研究向更多方向延展。

本专著是在第一作者作为负责人的国家自然科学基金项目“h-指数与类h-指数的机理分析与实证研究”（批准号：70773101，2008.1～2010.12）资助下产生的研究成果，项目执行期间课题组成员发表国际同行评议论文多篇、发表国内核心期刊论文两组，主要有：

Ye F Y. A unification of three models for the h-index. Journal of the American Society for Information Science and Technology, 2011, 62 (1): 205-207

Ye F Y, Rousseau R. Probing the h-core: an investigation of the tail-core ratio for rank distributions. Scientometrics, 2010, 84(2): 431-439

Ye F Y. An investigation on mathematical models of the h-index. Scientometrics, 2009, 81(2): 493-498

Ye F Y, Rousseau R. The power law model and total career h-index sequences. Journal of Informetrics, 2008, 2(4): 288-297

Rousseau R, Ye F Y. A proposal for a dynamic h-type index. Journal of the American Society for Information Science and Technology, 2008, 59(11): 1853-1855

叶鹰．一种学术排序新指数——f指数探析．情报学报，2009，28(1)：142-149

叶鹰．对数f指数及其评价学意义．情报科学，2009，27(7)：965-968

丁楠，潘有能，叶鹰．基于CSSCI的文科学者h指数实证研究．大学图书馆学报，2009，27(2)：55-60

潘有能，丁楠，朱佳惠等．基于Web of Science的理科学者h指数实证研究．大学图书馆学报，2009，27(2)：61-65，84

周英博，马景娣，叶鹰．国际基础科学核心期刊h指数实证研究．大学图书馆学报，2009，27(2)：66-70

程丽等．国际大学h指数与综合指标排名的比较研究．大学图书馆学报，2009，27(2)：71-75

次仁拉珍，乐思诗，叶鹰．世界百强企业h指数探析．大学图书馆学报，2009，27(2)：76-79

乐思诗，叶鹰．专利计量学的研究现状与发展态势．图书与情报，2009(6)：63-66，73

次仁拉珍，叶鹰．专利权人h指数研究．图书与情报，2009(6)：67-69，107

唐健辉，叶鹰．3G通信技术之专利计量分析．图书与情报，2009(6)：70-73

乐思诗，唐健辉．汽车节能技术之专利计量分析．图书与情报，2009(6)：74-77

这些成果为本专著的写作奠定了坚实基础。

专著内容分为发展概况、理论机理、实证研究、相关研究、延伸研究共五篇15章，各章大多有研究论文基础，因而相对独立且参考文献自成体系，符合专著构造特点。

尽管h指数原产于作者论文按引用次数排序后的发文数与引文数的平衡点且原意作为评价和排序学者的测度，但在应用推广和理论研究中发现这一简单有趣的指数不仅可以推广应用于学术期刊、研究机构、专利权人，甚至国家，而且在理论上还能启发不少优雅的研究构思，因此获得学术界的青睐。不过，其片面单一的特性也遭到不少批评，有必要认识到高h指数不必然等于高学术水平和高知识创造。有鉴于此，兼顾h指数的优点和缺点，客观公正地研究并运用h指数和h型指数成为本专著的学术立场，论著内容由此展开。

第一篇 发展概况

2005年11月15日，一篇名为“An index to quantify an individual's scientific research output”的论文发表在美国科学院院报（PNAS）102卷46期上（Hirsch，2005），署名J. E. Hirsch，虽然该文2005年8月已经在arXiv上公开，但PNAS的正式发表强化了学术界的关注。*Nature*评论员P. Ball在当年*Nature*第436卷900页上的一页评论（Ball，2005），正面肯定了h指数的效果，后来尽管也有不同意见（Lehmann et al.，2006），h指数仍以其简单新颖引人关注，很快，h指数研究的学术帷幕开启并成为学术热点。

第 1 章　h 指数及其变体

1.1　Hirsch h 指数

按照 Hirsch 的原始定义，一名科学家的 h 指数是指其发表的 N_p 篇论文中有 h 篇每篇至少被引 h 次、而其余 N_p-h 篇论文每篇被引均小于或等于 h 次（A scientist has index h if h of his or her N_p papers have at least h citations each and the other (N_p-h) papers have $\leqslant h$ citations each），也就是说：一位学者的 h 指数等于其至多发表了 h 篇每篇至少被引 h 次的论文，亦即一个学者的 h 指数表明其至多有 h 篇论文被引用了至少 h 次。Braun 等将原来针对学者的 h 指数概念用于期刊（Braun et al.，2005），提出一种期刊的 h 指数等于该期刊发表了至多 h 篇每篇至少被引 h 次的论文，或者说一种期刊的 h 指数表明该期刊所发表的全部论文中最多有 h 篇论文至少被引用了 h 次。一般地，可将一个学术信息源的 h 指数定义为该信息源至多有 h 篇每篇至少被引用了 h 次的学术发文数，这一概念可普遍适用于学者、期刊、机构（包括大学）、专利权人乃至国家。

h 指数的计算首先需要把论文按照被引次数从高到低排序，表 1.1 是对其形成机制的数据示意。

表 1.1　h 指数形成机制的数据示意

发表及被引数据			排序数据	
PY	P	C	TC	r
1996	1	2	32	1
1997	2	3+5	25	2
1998	3	4+6+8	20	3
1999	2	10+9	18	4
2000	3	32+16+25	17	5
2001	2	20+18	16	6
2002	1	15	15	7
2003	5	1+2+3+17+11	12	8

续表

发表及被引数据			排序数据	
PY	P	C	TC	r
2004	4	12+8+6+3	11	9
2005	3	9+7+5	10	10→h
2006	2	2+1	9	11
			9	12
			8	13
			8	14
			7	15
			6	16
			6	17
			5	18
			5	19
			4	20
			3	21
			3	22
			3	23
			2	24
			2	25
			2	26
			1	27
			1	28

注：PY=publishing year；P=papers；C=citations from publishing to present；r=order of paper；TC=total citations in decreasing order

设r是按被引次数降序排列的论文的序次，TC_r是论文r的被引总数，则有以下序列：

$$\boldsymbol{r} = (1,2,\cdots,r,\cdots,z) \tag{1.1}$$

$$TC = (TC_1, TC_2, \cdots, TC_r, \cdots, TC_z); TC_1 \geqslant TC_2 \geqslant \cdots \geqslant TC_r \geqslant \cdots \geqslant TC_z \tag{1.2}$$

h指数就是

$$h = \max\{r : r \leqslant \mathrm{TC}\} \tag{1.3}$$

即把一位学者发表的论文按其被引次数（TC）从高到低排序（r）后，h 指数等于按被引次数从多到少排列的单篇论文总计被引次数（TC）大于等于 r 时对应的最大序数 r。

参照学者们的总结（Costas and Bordons，2007；Bornmann et al.，2007；叶鹰，2007；赵基明等，2008；Rousseau，2008；Alonso et al.，2009），h 指数的主要优点有以下几个方面：

（1）具有数学简单性（It is a mathematically simple index）；

（2）具有数值稳健性（It is a robust indicator）；

（3）结合了产出与影响（It incorporates both output and impact）；

（4）适用于各种层次（It can be applied to individual and aggregative levels）；

（5）数据容易获取（Data are easily obtained）。

缺点相应如下：

（1）具有数据源依赖性（It is dependent on database source）；

（2）缺乏敏感性（It lacks sensitivity to changes in performance）；

（3）只升不降会导致吃老本（It allows scientists to rest on their laurels）；

（4）会受自引影响（Self-citations etc. can positively influence its value）；

（5）很难收全决定 h 指数的完整数据（It is difficult to collect complete data for the determination of the h-index）。

尽管褒贬俱存，各种 h 指数变体、h 型指数仍层出不穷（Alonso et al.，2009；Egghe，2010；Cabrerizo，2010），择要综述如下。

1.2 h(2)指数与 h_f 指数

1.2.1 h(2)指数

由于在利用 ISI Web of Science 计算学者 h 指数过程中，存在英文姓名缩写相同或英文姓氏改变等原因导致学者文献数量与引文次数变化，造成 h 指数计算误差。Kosmulski（2006）通过增加高被引文献的权重，提出了 h(2) 指数，其定义是：一个科学家的 h(2) 指数是其发表论文中有前 $h(2)$ 篇高被引文献至少被引用 $[h(2)]^2$ 次的最大自然数（“A scientist's h(2) -index is defined as the highest natural number such that his $h(2)$ most cited papers received each at least $[h(2)]^2$ citations.”）。

例如，h(2) 指数为 10，代表这位科学家至多发表了 10 篇至少被引用 100 次的论文。

h(2) 指数的计算过程与 h 指数相似，以表 1.1 数据为例转换为表 1.2。

表 1.2 学者 h(2) 指数计算过程示意

发表及被引数据			排序数据		
PY	P	C	TC	r	r^2
1996	1	2	32	1	1
1997	2	3+5	25	2	4
1998	3	4+6+8	20	3	9
1999	2	10+9	18	4→$h(2)$	16→ $[h(2)]^2$
2000	3	32+16+25	17	5	25
2001	2	20+18	16	6	
2002	1	15	15	7	
2003	5	1+2+3+17+11	12	8	
2004	4	12+8+6+3	11	9	
2005	3	9+7+5	10	10→h	
2006	2	2+1	9	11	
			9	12	

由此可见，对于任何一位科学家而言，h(2) 指数将不高于其 h 指数数值 (Jin et al.，2007)。h(2) 指数的最大优点在于减少了误差问题，为利用 Web of Science 数据库计算 h(2) 指数数值减少了论文数据核对等耗时，尤其是计算 h 指数过程中辨析学者姓名的大量工作 (Bornmann et al.，2007)，为实证研究带来便利。

但是 h(2) 指数并未能解决 h 指数因不同学科与不同年龄、自引与互引等因素导致的不足和缺陷。

1.2.2 h_f 指数

用 h_f 指数泛指经过归一化处理的 h 指数 (Sidiropoulos et al.，2007；Iglesias and Percharroman，2007)，其目的是实现对不同学科主体进行直接比较。Sidiropoulos 等把归一化 h 指数记为 h^n 并定义为

$$h^n = \frac{h}{N_p} \tag{1.4}$$

其中 N_p 为 h 指数主体发文总数，这实际上是一种篇均 h 指数，而且标记符号容易混淆指数 n 次方。我们更倾向更一般的 h_f 指数定义 (周英博等，2009)：

$$h_f = \frac{h}{f_m} \tag{1.5}$$

其中 f_m 表示不同学科的篇均被引次数，相当于归一化因子。最简便 f_m 取值可用 ESI 十年累积篇均被引作为实际数据。表 1.3 是 1998～2008 年 ESI 中 22 个学科的具体数值。

表 1.3 ESI 平均引文率：f_m 取值

领域	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	All Years
All Fields	17.74	17.11	16.38	15.08	13.56	11.62	9.74	7.33	4.69	2.35	0.46	9.91
Agricultural Sciences	11.21	11.12	11.10	9.97	8.99	8.05	6.61	4.81	3.12	1.43	0.25	6.20
Biology and Biochemistry	29.13	27.61	26.69	24.11	21.37	18.42	15.09	11.11	7.04	3.52	0.66	16.41
Chemistry	16.47	15.69	15.64	14.04	13.51	11.67	10.01	7.84	5.14	2.65	0.54	9.72
Clinical Medicine	20.62	20.05	19.21	17.92	16.42	14.42	12.14	9.32	5.95	2.91	0.52	11.99
Computer Science	7.20	6.36	5.84	6.02	5.98	3.81	2.66	1.99	1.06	0.80	0.15	3.15
Economics and Business	10.50	9.55	8.84	7.77	7.49	6.17	4.94	3.42	1.94	0.86	0.17	5.19
Engineering	6.75	6.66	6.45	6.15	5.48	4.81	4.18	3.06	1.92	1.00	0.18	3.93
Environment/Ecology	19.20	17.84	17.80	15.33	13.86	11.93	9.80	7.11	4.50	2.15	0.38	9.75
Geosciences	17.95	16.35	14.73	13.59	11.27	9.92	8.08	6.00	4.04	1.76	0.43	8.72
Immunology	36.18	33.00	32.95	30.70	26.96	23.33	20.08	15.00	9.87	5.10	0.93	20.92
Materials Science	9.60	9.41	9.57	8.92	7.93	7.40	6.13	4.61	3.07	1.50	0.27	5.72
Mathematics	5.86	5.79	5.18	4.52	4.27	3.61	2.99	2.27	1.44	0.69	0.15	3.07
Microbiology	28.23	26.44	24.67	22.55	20.00	17.32	14.62	11.59	7.04	3.42	0.65	15.04
Molecular Biology and Genetics	46.63	44.38	41.95	38.51	34.16	28.64	23.71	17.49	11.33	5.64	1.12	25.13
Multidisciplinary	3.54	3.46	3.74	5.45	7.32	6.65	5.76	4.24	4.87	3.42	1.12	4.17
Neuroscience and Behavior	32.67	31.69	29.88	27.99	24.24	20.02	16.67	12.62	8.18	3.88	0.71	18.18
Pharmacology and Toxicology	17.87	18.22	17.89	17.06	16.05	13.21	11.83	8.50	5.97	2.82	0.51	10.96
Physics	13.81	13.35	13.10	11.86	10.51	9.25	8.18	6.36	4.30	2.06	0.55	8.19
Plant and Animal Science	12.81	12.30	11.76	10.71	9.58	8.21	6.87	4.97	3.13	1.50	0.31	7.06

续表

领域	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	All Years
Psychiatry/Psychology	18.39	18.34	16.78	15.66	13.38	11.94	9.68	6.96	4.28	1.86	0.37	9.93
Social Sciences, general	7.64	7.35	7.09	6.34	5.85	5.04	4.32	3.26	1.96	0.88	0.20	4.16
Space Science	21.34	23.14	18.11	19.68	15.40	15.77	12.99	10.73	7.39	4.49	0.90	13.17

数据来源：http：//esi.isiknowledge.com/baselinespage.cgi

1.3 h_I 指数与 h_m 指数

1.3.1 h_I 指数

作为一项科学研究评价的新指标，h指数对不同研究领域的合作方式、引文方式等因素较为敏感，导致h指数存在学科依赖性和差异化。例如，Hirsch列出了物理学领域学者最高h指数为110（Witten E），而生命科学领域学者最高h指数为192（Snyder S H）。同时，h指数是一个整数，当两位学者h指数相等时，为进一步区分需要选择另外的指标进行评价。

为弥补h指数的上述不足，Batista等考虑学者合作这一因素的影响，以物理学、化学、生物学/生物医学和数学四大学科为例，将h指数除以 h 篇论文中学者人数的平均值，对各学科差异进行标准化处理，提出了 h_I 指数（Batista et al.，2006），计算公式为

$$h_I = h^2 / N_a^{(T)} \tag{1.6}$$

其中，$N_a^{(T)}$ 是 h 篇论文中的学者总数（允许学者重复出现）。假如某位学者在其 h 篇论文中均为单独作者，此时 $N_a^{(T)} = h$，h_I 指数与h指数值相等。h_I 指数近似表征了某一位科学家独立写作的 h_I 篇文献均至少被引用了 h_I 次，可相对有效地测量属于作者自己的产出及影响。但是，h_I 指数的算法仅考虑了合作作者数量，忽略了作者署名排序的差异，故高小强等提出了按作者署名顺位对被引次数进行分权的 h_{AW} 等指数作为补充（高小强，赵星，2010）。

1.3.2 h_m 指数

h_m 指数是Schreiber针对 h_I 指数的不足提出的，他认为简单地以 h 篇论文学者数的平均值作为标准化处理方法，会过多地降低大规模合作论文的影响力和过多地增加单个学者论文的影响力。