

doi:10.13582/j.cnki.1674-5884.2017.07.028

# $\theta$ —指数：一个科研绩效评价的新指标

艾颖

(广东第二师范学院 图书馆,广东 广州 510310)

**摘要:**针对在科研绩效评价中广泛应用的指标,如发表论文总量、论文总的被引次数、篇均引用次数、 $h$ 指数等存在的问题,引入一个新的指标: $\theta$ —指数,并且用此指标分析了广东第二师范学院科研人员的科研绩效,从而说明此指标的优越性。

**关键词:** $\theta$ —指数;科学评价;高被引用的论文; $h$ —指数;评价指数

**中图分类号:**G253      **文献标志码:**A      **文章编号:**1674-5884(2017)07-0123-03

对科研人员的绩效评价一直是个热点问题,根据特殊的原则选择一定的项目是评价的重要部分,本文的目的是提出一个新的科学评价的方法,主要是通过计算科研人员最有影响力的出版物的影响来评价。科研人员最有影响力的出版物的选择不外乎要考虑该人员发表论文的数量及质量。

一般地,许多评价指标考虑的是研究对象的出版物的全体,因此最简单的评价指标就是已发表论文的总量。这个指标代表科研人员的全部的学术成果,而论文的总被引用数则代表其学术成果的总的影响力。当然论文总被引用数依赖于发表论文的数量,一个大的论文引用数可以来自大量的已发表论文,其实,每篇论文的引用分布(即引用率)是不同的,许多论文只有少量的引用次数(包括自引),但是增加了总引用次数。

2005年8月,美国加州大学圣地亚哥分校统计物理学家 J. E. Hirsch 提出的  $h$ —指数<sup>①</sup>,就考虑了科研人员论文的论文数量和被引频次两个指标:如果一个科研人员发表的  $N$  篇论文中有  $h$  篇论文的被引次数至少为  $h$ ,那么他的  $h$  指数为  $h$ 。如 E. Witten 的  $h$  指数 = 110,即 he 已发表了 110 篇论文,其中每篇论文被引的次数至少为 110 次。可以计算一个人的  $h$  指数<sup>②</sup>,也可计算一个团队、一个研究所、一个杂志,甚至一个国家的  $h$  指数<sup>[1]</sup>。也有学者给出了  $h$  指数的一个数学解释<sup>③</sup>,一些改进的  $h$  指数也被提出<sup>[2]</sup>, $h$  指数应用中的问题也可以在他处得见<sup>[3,4]</sup>。

一个科研人员的科研成就的大小应该根据其出版物对科学的影响来评价,而不是根据其出版物的总量来评价。然而,从  $h$  指数的定义可知,一个科研人员的  $h$  指数是小于等于其出版的论文数的( $h \leq P$ )。例如,一个科学家若发表了 25 篇论文,则它的  $h$  指数不可以超过 25,尽管有可能他的一些文章是优质且总引用次数就达 1 000 以上。

论文的被引次数直接反映了其他科研工作者对该论文的肯定程度,而论文被参考和引用是在得到了其它科研工作者对该论文的质量认可的基础上进行的,高的被引论文是科研工作者高质量成果中最

收稿日期:20170425

作者简介:艾颖(1964-),女,江西新余人,图书馆馆员,主要从事图书情报学研究。

① J. E. Hirsch. An index to quantify an individual's scientific research output. (2005,102:16569—16572). Proceedings of the National Academy of Science of the United States of America.

② M. Schreiber. An empirical investigation of the  $g$ -index for 26 physicists in comparison with the  $h$ -index, the  $A$ -index, and the  $R$ -index. (2008,59:1513-1522). Journal of the American Society for Information Science and Technology.

③ W. Glanzel. On the  $h$ -index—A mathematical approach to a new measure of publication activity and citation impact. (2006, 67: 315-321) Scientometrics.

具代表性的最高水平成果,理应对科研评价的贡献大,但是在  $h$  指数的应用中没有充分体现出来。另外, $h$  指数与其它指标相比,存在区分度不够明显的缺陷,尽管许多学者的  $h$  指数一样,但是有可能其  $h$  排序及不同论文不同被引次数叠加而产生的最终的学术影响不同。

## 1 $\theta$ —指数的定义

引进  $\theta$ —指数这一新的文献计量指标的目的就是给具有高影响的论文(高被引用论文,即在所分析的论文中获得相对高的引用的文章)更多的考虑。具有高影响的论文集合的论文数规定为全部论文数的 20%(小数按四舍五入), $\theta$ —指数的具体定义是:设一个科学家发表了  $P$  篇论文,则他的高引用论文集(即优秀论文集)中的论文篇数为  $P_\theta = 0.2P$ 。将该科学家所有发表的论文按引用次数大小排序,记位于前  $P_\theta$  篇的论文的被引用次数之和为  $C(P_\theta)$ ,则定义该科学家的  $\theta$ —指数  $= 0.01C(P_\theta)$ 。

$\theta$  指数可根据许多数据库(如美国 ISI 数据检索平台、中国学术文献网络出版总库等)来计算。由计算公式,根据某个科研人员的论文,考虑每篇论文的引用次数,我们可以计算出其对科学进步的影响。

之所以将研究对象的优秀论文集中文章的篇数确定为全部文献的 20%,这主要是考虑到“二八定律”,即巴莱多定律,它是 19 世纪末 20 世纪初意大利经济学家巴莱多发明的。他认为,在任何一组东西中,最重要的只占其中一小部分,约 20%,其余 80%,多数都是次要的。习惯上,二八定律讨论的是顶端的 20%,而非底部的 20%。

为说明  $\theta$ —指数的应用,我们对广东第二师范学院(原广东教育学院)几位科研工作人员在 ISI 数据检索平台被检索到的数据进行分析,计算出相应的  $\theta$ —指数,再将之与传统的文献计算指标相对比。

## 2 应用分析

### 2.1 数据来源

本文利用美国 ISI 网络检查平台(科学引文索引),它包括 SCIE,SSCI,A&HCI,CPCI-S,CPCI-SSH,CCR,IC 等 7 个数据子库,在其中检索出至 2010 年 12 月,广东第二师范学院科研人员发表的论文及被引频次的数据,统计分析得到部分科研人员的每一篇论文的被引用次数,最后计算出他们各自的  $\theta$ —指数。

### 2.2 数据获取原则

至 2010 年 12 月底,广东第二师范学院科研人员发表的论文在美国 ISI 数据平台中共检出 186 篇,其中被引论文 103 篇,总被引数 668 次,我们选择被收录 8 篇以上的学者,再以姓名检查出每篇论文的被引次数,再将被引文献按被引频次降序排列,计算出作者的  $\theta$ —指数。

1) 也许是 ISI 检查平台的数据库录入存在一些错误,也有可能是科研人员发表论文的单位署名不规范,故我们在收集数据对一些单位名称作了更正。

2) 作者姓名的不同拼法尽量查全,如杨必成有:Yang BC 和 Yang B,慧萍有 Hui P 和 Ping H 等。

3) 选取所有学术论文和综述,除去 book, meeting, abstract letter, editorial material 和 correction 等类型文献。

4) 特别注意的是,论文的单位必须是署名广东第二师范学院(或广东教育学院),不区分排名。另外,我们分析的数据,引用频次中没有删除自引。

5) 符号: $P$  表示文章总数; $h = P_h$  表示  $h$  指数及计算  $h$  指数时的优秀论文数量; $P_\theta$  表示计算  $\theta$ —指数的论文数,即高影响的文集(优秀论文集)中文章的篇数, $P_\theta = 0.2P$ ;  $P$  篇论文总的引用次数  $C$ ;篇均引用次数  $(C/P)_i$ ;高影响的论文集中论文篇均引用次数  $(C/P)_\theta$ ;  $\theta$ —指数  $\theta$ ;  $h$  指数  $h$ ;  $C_\theta:P_\theta$  篇论文被引次数。

### 2.3 结果

表 1 给出了根据美国 ISI 数据平台获得的广东第二师范学院(原广东教育学院)科研人员发表论文

的数据分析的结果。

表 1 数据分析结果

| 姓名  | $P$ | $C$ | $P_{\theta}$ | $C_{\theta}$ | $(C/P)_i$ | $(C/P)_{\theta}$ | $h = P_h$ | $\theta$ |
|-----|-----|-----|--------------|--------------|-----------|------------------|-----------|----------|
| 杨必成 | 36  | 207 | 7            | 116          | 6.08      | 16.57            | 8         | 1.16     |
| 李样明 | 31  | 129 | 6            | 101          | 4.161     | 16.83            | 6         | 1.01     |
| 江俊勤 | 27  | 87  | 5            | 46           | 3.44      | 9.2              | 6         | 0.46     |
| 慧萍  | 20  | 54  | 4            | 35           | 2.7       | 8.75             | 4         | 0.35     |
| 张秀莲 | 11  | 110 | 2            | 89           | 13.286    | 44.5             | 4         | 0.93     |
| 尹伟  | 11  | 21  | 2            | 10           | 1.625     | 5                | 3         | 0.10     |
| 余超凡 | 8   | 3   | 2            | 2            | 0.286     | 1                | 1         | 0.01     |

根据表 1 的数据,我们可以分析有关指标的相关性。通过计算,我们得出如表 2 的相关系数:

表 2 相关系数

|                  | $P$      | $C$      | $P_{\theta}$ | $C_{\theta}$ | $(C/P)_i$ | $(C/P)_{\theta}$ | $h = P_h$ | $\theta$ |
|------------------|----------|----------|--------------|--------------|-----------|------------------|-----------|----------|
| $P$              | 1        |          |              |              |           |                  |           |          |
| $C$              | 0.81 808 | 1        |              |              |           |                  |           |          |
| $P_{\theta}$     | 0.99 462 | 0.80 117 | 1            |              |           |                  |           |          |
| $C_{\theta}$     | 0.71 689 | 0.95 658 | 0.69 728     | 1            |           |                  |           |          |
| $(C/P)_i$        | 0.04 039 | 0.54 670 | -0.0 073     | 0.66 848     | 1         |                  |           |          |
| $(C/P)_{\theta}$ | -0.0 011 | 0.49 577 | -0.0 468     | 0.65 475     | 0.98 941  | 1                |           |          |
| $h = P_h$        | 0.93 813 | 0.91 545 | 0.90 543     | 0.82 363     | 0.32 131  | 0.26257          | 1         |          |
| $\theta$         | 0.69 754 | 0.95 116 | 0.676 237    | 0.999 487    | 0.691 621 | 0.678422         | 0.81315   | 1        |

3 结论

根据  $h$ —指数的计算方法,将被研究对象的论文,按被引用次数,按降序排序,序号小于等于  $h$ —指数的论文被考虑,此集合中的论文可认为是  $h$ —指数的优秀论文集。比如, $h$ —指数=8,则优秀论文集中论文的篇数为 8。一般地,用 Hirsch 的方法计算  $h$ —指数时,要考虑的论文占全部论文的比例要大于 20%,即  $P_h > P_{\theta}$  (参考表 1 相应栏)。我们希望引入一个能给高引用次数的文章(即优秀论文集)更多关注的指标,但是要考虑的论文的数量又不是太多,因此我们提出的新的  $\theta$ —指数是比较好的。

我们注意到李样明与江俊勤、慧萍与张秀莲的  $h$ —指数都是一样,但是  $\theta$ —指数揭示了他们的出版物的影响力的差距,因此  $\theta$ —指数可有效克服  $h$ —指数区分度不够明显的缺陷。

从表 1 中注意到,张秀莲的出版物的篇均引用率与优秀论文集中篇均引用率都是最高的(见表 1 中  $(C/P)_h$ ,  $(C/P)_{\theta}$  栏),她有几篇具有较高引用次数的论文,但她发表的论文总量不多( $P=11$ ),故  $\theta$ —指数为 0.93,在统计对象中排第 3。由此可看出  $\theta$ —指数是一个综合评价指标,既考虑研究对象出版物的数量,又考虑质量,是一个综合评价指标。

显然,当根据两个数量的样本数据计算出来的相关系数  $r > 0.8$  时,则这两个数量有高度线性相关。从表 2 可以看出, $\theta$ —指数与  $h$ —指数高度线性相关(当然,我们的样本数不多。)

参考文献:

[1] 姜春林.  $h$  指数和  $g$  指数期刊学术影响力评价的新指标[J]. 图书情报工作, 2006(12): 632-651.

[2] 金碧辉. Rousseau Ronald.  $R$  指数、AR 指数:  $h$  指数功能扩展的补充指标[J]. 科学观察, 2007(3): 128-131.

[3] Vinkler P. Eminence of Scientists in the light of the  $h$ -index and other scientometric indicators[J]. Journal of information Science, 2007(33): 481-491.

[4] Costas R, Bordons M. The  $h$ -index: Advantages, limitations and its relation with other bibliometric indicators at the micro level[J]. Journal of Informetrics, 2007(1): 193-203.

( 责任校对 谢宜辰)