

基于作者主要贡献的 h 指数时间趋势分析^{*}

吕娜¹ 刘扬² 全少颖³

(1. 北京理工大学图书馆信息资源管理研究所 北京 100081;
2. 中国人民大学图书馆 北京 100872; 3. 北京理工大学图书馆 北京 100081)

摘要 基于作者贡献的权重 h 指数研究成果越来越多, 计算复杂度也大幅度提升, 仅用于评价目的会局限 h 指数的应用。结合时间趋势和作者贡献度来分析 h 指数可扩大 h 指数的应用目的, 更多用于分析研究人员的科研成长规律和预测其未来成果。案例研究部分以 4 个科研成果中等的研究人员为例, 构建基于第一作者的权重 h 指数时间发展趋势的计算和分析流程。

关键词 h 指数 科研团队 作者排名权重 时间趋势分析

中图分类号 G350

文献标识码 A

文章编号 1002-1965(2015)04-0054-05

DOI 10.3969/j.issn.1002-1965.2015.04.011

h-index Time Trends Analysis Based on Author Contribution

Lyu Na¹ Liu Yang² Quan Shaoying³

(1. Institute of Information Resource Management, Library of Beijing Institute of Technology, Beijing 100081;
2. Library, Renmin University of China, Beijing 100872;
3. Library of Beijing Institute of Technology, Beijing 100081)

Abstract More and more h-index research literatures are produced based on author contribution. The computing complexity is increased apparently, which limits the application of h-index. h-index based on author contribution in combination with time trend can be used to analyze research growth and predict the future achievement. In the case study, h-index data of four moderately contributive researchers were collected to construct the first author-based procedures for computing and analyzing the time trend of weighted h-index.

Key words h-index R&A Team weighted ranking of authorship time trend analysis

0 引言

h 指数在 2005 年被提出之初, 其原创者 Hirsch 就已经考虑到该指数没有区分作者贡献和作者数量的问题^[1]。2006 年开始很多研究人员也都陆续指出 h 指数最突出的缺点就是没有考虑到科研人员论文成果中合作者的数量^[2]。独立作者对论文的投入或者贡献虽然很大, 但他们的 h 指数与合作发表论文较多的作者相比没有优势。不排除有一部分研究人员的论文数量和 h 指数都较高, 但非第一作者或通信作者的论文数量占较大的比例, 这种现象在规模较大的项目团队中表现会比较突出。而且, 当前的科技评价工作往往采用论文数量和被引次数作为指标, 这在一定程度上推

动了这一现象, 一些项目团队的研究人员在论文发表环节往往会把那些贡献较少甚至没有贡献的其它项目团队成员进行署名, 甚至有些团队只要发表论文就把所有研究人员署名在作者项。这种情况目前还没有被纳入科研诚信标准, 属于灰色地带。

针对以上问题, 2006 年 Batista 提出根据 h 核内论文的作者数量的均值平分 h 指数, 但由于每个作者对于论文的投入和贡献都是不均衡的, 平分的做法有欠公平, 也会在某种程度上妨碍科研人员之间的合作^[3]。Egghe 在 2008 年提出根据作者排名来计算论文的引文数量再提取 h 指数, 2009 年 Schreiber 提出修正 h 指数的研究也都是基于类似的想法, 根据合作作者数量重新计算论文的被引次数再提取 h 指数^[2,4]。

收稿日期: 2014-12-09

修回日期: 2015-01-29

基金项目: 国家社会科学基金项目“科研领域合作关系的识别与关联强度分析”成果之一(编号: 13CTQ024)。

作者简介: 吕娜(1978-), 女, 博士, 硕士生导师, 副研究馆员, 研究方向: 科技情报管理与分析; 刘扬(1976-), 女, 博士研究生, 馆员, 研究方向: 关联数据; 全少颖(1989-), 女, 硕士研究生, 研究方向: 科技情报管理与分析。

Hirsch 本人也在 2010 年提出了文献计量指标 h -bar, 能够给独立署名或较少合作发表论文的作者加分, 不鼓励名义上的署名, 同时又不能妨碍真正的合作。他提出在 h 指数基础上, 将以下情况的论文计入 h -bar 核: 独立作者; h 指数高于其合作者的论文; h 指数低于其合作者, 但论文被引频次高于合作者的 h 指数。再考虑将 h 核外符合条件的少数论文纳入到 h -bar 核。 h -bar 指数相比较 h 指数具有一定的优势, 他考虑了作者的贡献, 同时在一定程度上能避免虚假或不必要的署名现象等。但是 h -bar 也有缺点, 主要表现在以下几点:

a. 计算相比较 h 指数来说略为复杂。利用文献数据库计算 h -bar 指数时, 需要获得每篇论文被引次数以及每篇论文合作者的 h 指数数据, 包括 h 核外的论文也要计算在内, 会花费一定的时间。

b. h -bar 指数与 h 指数一样, 对于某些研究人员同样不适用, 比如成果平平的研究人员以及少而精的人员。

c. 获得认可和推广可能会有障碍^[5]。

1 基于作者贡献的 h 指数相关研究现状分析

1.1 现有 h 指数改进研究视角 基于作者贡献的 h 指数改进研究主要有两种视角: 一种是根据作者署名顺序和署名规则仅提取作为主要贡献人的署名成果, 并纳入 h 指数计算范畴。通常有两种情况: 一是将第一作者和通讯作者纳入研究范畴; 二是将署名排在前 3 位的成果纳入研究范畴, 根据作者署名次序, 仅计算排名在前 1、2 或 3 位的论文被引次数后计算 h 指数。

基于作者贡献的另一种研究视角就是根据作者署名顺序计算权重, 将成果的被引次数加权后再重新计算序列提取 h 指数。也就是说, 根据每篇论文的作者数量平分引用次数或根据作者排序加权后计算引文次数再提取 h 指数。

1.2 现有研究成果存在的问题 自 Hirsch 2005 年提出 h 指数以来, 3 年内就产生了几百篇改进 h 指数的文献研究成果, 其中既有 g -index、 r -index 等新指数的提出, 也有 h_i 等 h -index 改进指数^[6], 其目的主要是通过改进 h 相关指数, 提高评价的科学性, 评价的对象包括学科、机构、国家和个体科研人员。在这些成果研究过程中, 主要是通过 Web of Science 等平台获取研究对象的引文数据, 通过识别作为主要贡献人的成果或对主要贡献人的引文次数进行加权后重新计算和提取评价其影响力和水平的 h 指数信息。

h 指数被提出以来, 以其计算简单这一优势被快速用于评价科研人员成果水平, 如果想获取某位研究

人员的 h 指数只需登陆 Web of Science 等数据库平台进行检索, 点击检索结果界面中的引文分析功能, 该功能直接提供 h 指数, 也可通过降序排列被引次数通过浏览获得 h 指数数据。科研管理部门、图书情报部门甚至科研人员自身都能够快速获取 h 指数信息。后续针对 h 指数的改进研究无论是公正性还是客观性都有了很大的提升, 但是随之带来的却是数据检索和处理问题以及计算方法的复杂性大大提升等问题。本来只需动动鼠标简单的点击几次即可完成的工作, 现在则需要计算每篇论文的作者数量及其排序信息, 根据数量和权重计算被引次数后重新对论文进行排序, 进而提取 h 指数, 该工作还需要 Excel 等其他数据分析和工具的帮助才能完成。计算一个人的改进 h 指数所投入的时间从原来的几分钟可能需要花费几个小时。对于科技评价工作往往需要计算大量科研人员的 h 指数则需要的时间就更难以估计了, 这也是为什么后续改进指标没能得以推广的原因。

2 基于主要贡献逐年提取 h 指数的主要思路和方法

h 指数和相关改进指数研究均以当前引文数据为基础, 对研究人员当前成果的产出物进行评价, 主要原因是 Web of Science 平台的引文数据统计是当前的, 也就是说, 现在登陆 Web of Science 平台提取研究人员的 h 指数与较早时间, 甚至几天前的结果都是不一致的。而大多研究成果都是以检索时间点获取的静态数据进行研究, 忽视了 h 指数的时间累积。Web of Science 平台的引文数据是动态增长的, 科研人员当前的 h 指数数值都会经历逐渐累积的过程, 需对科研人员不同时期的 h 指数进行监测来研究 h 指数的动态增长规律。

不同研究领域科学家发表论文数量、 h 指数增长都会呈现不同的趋势, 不同排名署名的论文对 h 值的贡献也会反映不同类型科学家, 能描述其科研主动性和主导性程度, 也是对其个人科研能力的衡量, 对科研人员 h 指数历年增长趋势进行统计, 进而分析其科研成长规律是 h 指数时间趋势研究的重点所在。

由于 h 指数改进研究计算方法的复杂性, 那么研究 h 指数的目的性就必须要从评价角度转移到分析和预测。基于作者贡献的 h 指数可用于分析研究人员科研经历特征, 如果能够对科研人员过去一段时间 h 指数逐年或根据时间进行分别计算的话可用于提取科研人员成长模式, 预测其未来成就。Hirsch 也在 2010 年提出应该研究科研人员 h 指数随时间的演变规律。金碧辉在 2007 年以 10 年期为观测窗口提取了某一科学家 A 的 h 指数, 考核的是不同时期其 h 指数变化率。同年她与张晓阳对科学家在研究生涯内 h 指数线性成

长规律进行了研究,并对 h 核内论文对 h 指数贡献情况进行分析,发现当高被引科学家不再发表论文后, h 指数仍在 5~8 年期呈现对数型成长^[7-8]。

论文后面实证分析部分主要基于以下两点思路:首先,考虑作者排名因素,现阶段仅对排名第一的论文引用情况分别进行计算,提取 h 指数,暂不考虑权重。其次,逐年提取研究人员的 h 指数数据,即获取某一科研人员过去某一时间节点 T 的 h 指数数据(也就是该作者当时的 h 指数),这就要求既要成果文献(来源文献)限定在过去某一时间节点之前又要筛选出该时间节点之前的引证文献并逐年统计篇数,假设计算科研人员 A 在 2010 年的 h 指数,所有的引证文献及频次数据也应该统计截止到 2010 年及之前。主要步骤如下:

第一步,收集来源数据。可利用数据库平台检索某科研人员的论文成果数据以及对应的引证文献数据。

第二步,提取统计字段。根据来源数据,提取计算 h 指数所需要的数据,包括作者项信息,用于提取第一作者的论文成果,以及每篇论文引证文献的数量和出版年信息。

第三步,生成提取 h 指数的数据序列。逐年统计论文成果信息和引证文献逐年累积数量,并根据被引

次数降序排列,根据该序列提取 h 指数。

3 逐年提取第一作者成果的 h 指数实例研究

3.1 研究对象的选择 根据 ESI 平台物理学领域篇均引文数据从高到低排列,同时考虑科研人员供职机构的稳定以便提高来源数据的准确性,避免同名不同人所造成的数据冗余问题,同时可减少数据清洗工作量,选取了 4 位 h 指数在 60~70 之间的研究人员 GEIM AK、NOVOSELOV KS、FERRARI AC、KA-TSNELSON,MI 作为研究对象。通过 Web of Science 平台利用作者检索功能,即“姓名+研究领域+机构”的检索策略获取来源数据,包括论文成果信息和引证文献。

根据上节提出的方法和步骤进行计算,通过 Web of Science 平台提供的来源数据下载功能,将来源文献及对应的引证文献的元数据下载并导入 Excel 进行处理和统计。由于 Web of Science 平台每次支持最多 500 条记录的下载,所以对于动辄上万篇引证文献的数据量来说需分批多次下载,这个过程是非常耗时的。利用 Excel 处理作者和引文数据,提取作者排序和引文时间数据。4 位研究对象作者数量排序基本数据如表 1 所示。

表 1 研究人员作者数量和排序信息

作者	h 指数	总论文数量	平均作者数量 (署名序列平均数)	最大作者数量 (署名序列最大数)	最小作者数量 (署名序列最小数)	署名第一论文数量 (独立署名论文数量)
AK	71	144	7.84(6.47)	21(20)	1(1)	10(4)
KS	68	174	7.86(5.71)	21(21)	1(1)	24(6)
AC	67	246	7.62(5.69)	112(48)	1(1)	29(5)
MI	59	298	5.30(3.86)	21(19)	1(1)	32(13)

4 位研究人员科研经历较为丰富,成果较为突出。他们的 h 指数都在 60~70 之间,差距不大,但他们的论文数量从 144~298,差距较大,署名第一作者的论文成果数量为 10~32,作者项署名排序平均值为 3.86~6.47,4 人随着发表论文数量的增多,其平均作者数量以及署名排序的平均值反而下降。论文数量大且作者项平均值较低,说明其所属的科研团队规模较小,并在其科研团队中处于负责人的位置,例如 4 人中表现较为突出的 MI。

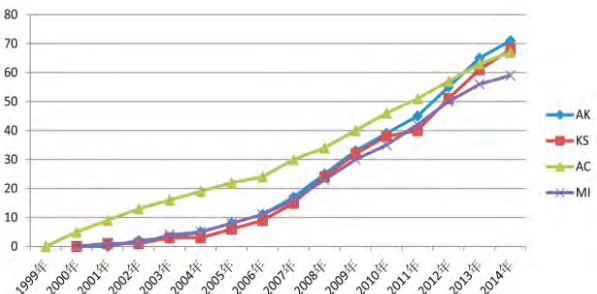
3.2 基于作者排名的逐年 h 指数结果分析 基于来源文献,提取第一作者的论文成果和引证文献的时间累积数据,逐年依次提取 h 指数、历年发文数量、历年第一作者署名发文数量、历年 h 核内第一作者署名论文数量,具体数据见表 2。其中 h_i 代表该作者作为第一作者发表的论文的 h 指数; P_n 代表该作者发表论文总数; P_1 代表作为第一作者发表论文的数量; $P_1 \ln$

(h) 代表 h 核内论文中作为第一作者发表的论文数量,该变量与 h_i 不同,前者基于全部论文序列而且逐年累积可能会下降,后者基于第一作者论文序列,与 h 指数一样,逐年累积只升不降。 $P_1 \ln(h)$ 用于侧重衡量作为论文主要贡献者对 h 值的贡献大小,也是对基于个人学术研究和领导能力对 h 指数的贡献大小的评价,对于团队合作较多的情况下,可用于识别核心科研人员,评价其学术能力。

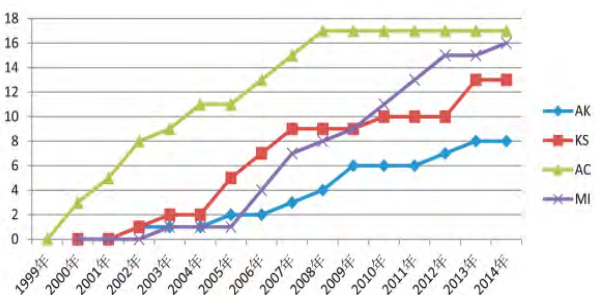
3.2.1 h 指数与 h_i 趋势比较 图 1 显示了 4 位研究人员 h 指数逐年变化趋势,AK、KS 和 MI 呈现出较高的线性拟合。这是因为 AK、KS、MI 虽然分别隶属两个机构,但他们之间有较多的合作。他们所发表的论文会全部贡献三人的 h 核,所以通过 h 指数很难看出三人科研实力的区别以及他们在团队中的角色和能力。

如果仅把作为主要贡献的成果,即排名第一的论

文以及引证文献数量排出序列,逐年提取 h 指数,即 h_t 指数,如图2所示,可以明显看出 h 指数拟合度比较高的三人中,MI的 h_t 曲线处于快速上升期,结合表1,其总论文数量、第一作者论文数量都较高,平均排序、平均合作者数量却较低,说明MI作为第一贡献者的论文成果影响力或者说质量较高,被引用频次超出AK和KS。相比较,MI的 h 指数趋势在三人中处于微弱的劣势,说明合作成果相对拉低了他的 h 指数,而其作为主要贡献者的成果表现较为突出,进而提升了其 h_t 曲线的强度。

图1 逐年 h 指数趋势

与以上三人分属不同机构也不存在合作的科研人员AC的 h 指数曲线一直处于增长状态,这符合大多数 h 指数时间增长趋势,但从图2可以看出,AC作为主要贡献者,其 h_t 指数却在近5年一直处于停滞状态,这说明一直贡献其 h 指数的论文成果主要是合作成果。当然不排除其部分作为主要贡献人的论文成果的价值还没有被充分发现的可能,未来也会有上升的空间。所以,虽然AC的 h 指数曲线是三人中最为突出的,但 h_t 曲线却说明AC的科研发展空间和潜力是三人中最弱的。

图2 h_t 指数趋势

3.2.2 h 核第一作者贡献情况分析 h 指数和 h_t 指数随时间只会增长不会下降, h 核内第一作者论文数量是会下降的(见图3),尤其当与他人合作较多且被引次数增长较快的情况下,部分原本在 h 核内的第一作者论文会随着被引频次增长缓慢而不在 h 核内,引起曲线的下降。这种下降可能是总体趋势的下降,也可能是短暂的下降,如果论文价值足够高的话,终究会回归 h 核。图1中,KS的 h 核内第一作者论文数在2006和2007年连续两年下降。如果能识别

出这些偏离 h 核的论文,可作为判断该研究人员阶段性成果价值的依据。

另外,作者还选择国内两个团队进行了 h 指数和 h_t 指数逐年变化趋势分析。表2和图4展示的是国内某985高校化学领域研究团队的 h 指数累积增长情况;表3和图5描述了国内某985高校团队 h 指数增长情况。

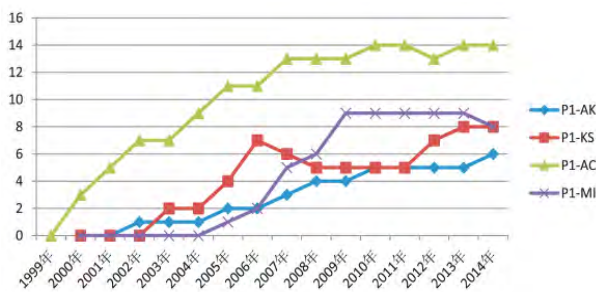
图3 h 核内第一作者论文数量趋势

表2 Qu LT 团队基本信息

	Qu LT	Hu CG	Zhao Y	Cheng HH
论文总数	104	31	46	25
第一作者论文数量	46	9	11	6
平均合作者数量	3.5	6.5	4	7
平均排名	2	1.5	2.5	2.5
当前 h 指数	25	12	15	14

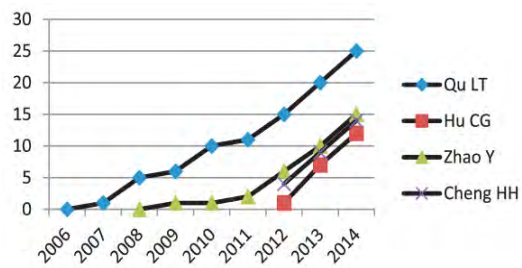
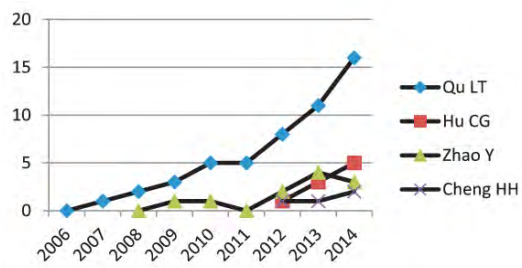
图4a Qu LT团队 h 指数增长情况。图4b Qu LT团队 h_t 指数增长情况

表3 Gao HJ 团队基本信息

	当前 h 指数	第一作者论文数量	平均排名	作者平均数	论文总数
高会军	51	40	2	4	280
刘国平	36	47	2.5	3	245
王常虹	30	0	2	3	小于78
张立宪	21	45	1.5	3	小于119

该团队特点:典型的老、中、青合作团队,整体趋势上升;在团队贡献的基础上注重个人独立研究成果。

结合以上两个图,团队主要成员影响力处于快速上升期,团队尤其是注重青年人才的培养和成长,他们能够独立或作为主要贡献者发表成果,贡献其 h 指数。与表3案例相比,虽然该团队 h 指数较低,但从增长趋势以及合作模式来看更为健康。

该团队种,高会军、张立宪、王常虹是联系较为紧密的合作者,王常虹作为非第一作者与高会军和张立宪合作发表论文数量较多,并依靠合作论文获得较高被引次数,其中与高会军合作48篇,与张立宪合作7篇,三人共同合作论文仅为3篇,尤其借助与高会军的合作使得其 h 指数高于张立宪。高会军、王常虹和张立宪合作较多的三人中,明显看出高会军 h 指数的快速增长,其独立和第一作者署名成果数量较多,属于三人中未来科研潜力和成就最大的一个;相比较张立宪独立研究能力较强, h 指数多依靠自己的成果论文做贡献,属于稳定增长。王常虹虽然 h 指数增长,但其没有独立或第一作者署名的论文,不宜作为学术带头人。四人中的刘国平属于科研年龄最长的一个,与其他三人几乎没有合作,其 h 指数呈现出稳定的增长态势,属于健康稳健类型。

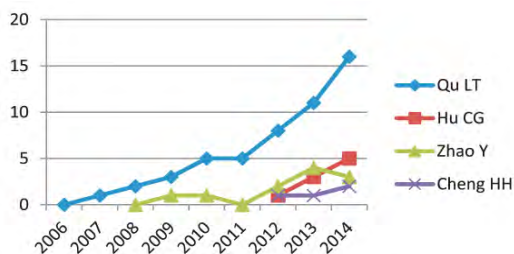


图5 Gao HJ 团队 h 指数增长情况

4 总结与讨论

本文的研究立足点并不在于提出一个易于计算的科研绩效评价标准,重点在于关注科研人员研究历程,分析其成长趋势和规律,尤其是科研团队成员的 h 指数发展规律,可以结合合著网络或科研合作网络的研究,识别团队成员的角色和影响力以及与团队其他成员的关系。通过 h 指数成长曲线来综合评价科研人员学术影响力则更为全面,一些较大规模的科研团队会存在 h 指数协同增长现象,基于作者贡献和作者排序

重新计算 h 指数则可辅助用于对团队成员关系进行分析,作为预测其未来成就的依据。由于时间所限,作为阶段性研究成果,实证研究仅采用4位研究人员的论文数据和引证文献数据作为对第一作者论文成果 h 指数的提取方法流程的验证,并对结果进行初步分析。基于作者贡献的 h 指数时间趋势研究还需要更多的数据支持和深入的分析,将通讯作者、前三排名以及权重信息纳入到研究范畴,最终目的是形成科研人员未来潜力预测模型。

参考文献

- [1] Hirsch J E. An Index to Quantify an Individual's Scientific Research Output[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(46): 16569 - 16572.
- [2] Schreiber M. A Case Study of the Modified Hirsch h_m Accounting for Multiple Coauthors[J]. Journal of the American Society for Information Science and Technology, 2009, 60(6): 1274 - 1282.
- [3] Batista P D, et al. Is it Possible to Compare Researchers with Different Scientific Interests[J]. Scientometrics, 2006, 68(1): 179 - 189.
- [4] Egghe L. Mathematical Theory of the H - and G - index in Case of Fractional Counting of Authorship[J]. Journal of the American Society for Information Science and Technology, 2008, 59(16): 1608 - 1616.
- [5] J E Hirsch. An Index to Quantify an Individual's Scientific Research Output That Takes into Account the Effect of Multiple Coauthorship[J]. Scientometrics, 2010, 85(3): 741 - 754.
- [6] 杜建, 张 玢. 作者合作视角下的 h 指数计量方法: 比较与归纳[J]. 图书情报工作, 2011, 55(24): 52 - 55, 136.
- [7] 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, 2(3): 1 - 8.
- [8] Jin B H, Liang L M, Rousseau R, et al. The R - and AR - indices: Complementing the h - index[J]. Chinese Science Bulletin, 2007(52): 855 - 863.
- [9] 张晓阳, 金碧辉. 高被引科学家 h 指数成长性探讨——以分子生物学与遗传学领域为例高被引[J]. 科学学研究, 2007, 25(3): 407 - 425.

(责编: 刘影梅)