

用无参数指标 Bcp 识别睡美人文献及其作者动态 h 指数变化规律

李 贺 解梦凡 袁翠敏 余 璐

摘 要 识别睡美人文献,发现其作者的动态 h 指数变化规律,可多维度观察睡美人文献,并从作者影响力角度探索睡美人文献的唤醒要素。本文结合识别睡美人文献的参数标准与无参数指标 Bcp 的计算方法,建立了基于 Bcp 无参数指标的睡美人文献识别标准,依据 Bcp 指数的计算和 Bcp 无参数指标识别标准,识别出 *Science*、*Nature* 和 *PNAS* 三大期刊高被引文献中的睡美人文献作为研究样本,并从中筛选高 Bcp 指数的代表性睡美人文献,探讨其作者影响力指标 h 指数的动态变化规律。研究结果表明:根据 Bcp 识别标准的睡美人文献比例为 2.38%,与前人研究的比例比较吻合,该标准被验证具有可行性;睡美人文献的出现与唤醒时间分布规律具有阶段性特征,且各阶段相互对应,但先后顺序不同;睡美人文献作者历年动态 h 指数曲线整体呈线性趋势;存在部分具有特殊上升规律的动态 h 指数曲线的典型“睡美人作者”。图 6。表 1。参考文献 35。

关键词 睡美人文献 时间分布 动态 h 指数 作者影响力 Bcp 指数

分类号 G250.2

Using the Non-parameter Index Bcp to Identify Sleeping Beauty Literature and Exploring the Changing Law of its Authors' Dynamic h Index

LI He, XIE Mengfan, YUAN Cuimin & YU Lu

ABSTRACT

The study of the identification of Sleeping Beauty literature and the dynamic law of the author's h index can observe the Sleeping Beauty literature from multiple dimensions and provide important reference value for the exploration of the awakening elements of Sleeping Beauty literature from the perspective of author's influence. In this paper, firstly, combining Van Raan's parameter standard and non parameter index Bcp as the theoretical basis, we constructed a standard of Sleeping Beauty literature recognition based on non-parameter index Bcp. Secondly, according to the Bcp index and the standard established above, the Sleeping Beauty literature in the three journals of *Science*, *Nature* and *PNAS* was identified as a research sample, from which the representative Sleeping Beauty literature of high Bcp index was screened to explore the dynamic growth law of the author's h index. The results show: 1) according to the Bcp identification standard, 558 papers of Sleeping Beauty literature are identified, which are about 2.383% of the total sample. Considering various factors such as different literature sources, expansion of time range and the curve's overall trend, this result slightly increases by 1% compared with the ratio 1.29% identified by Li Jiang, which can prove the standard created before is feasible. 2) The publishing and awakening of Sleeping

通信作者:解梦凡, Email: xmf1226@163.com, ORCID: 0000-0002-0670-8160 (Correspondence should be addressed to XIE Mengfan, Email: xmf1226@163.com, ORCID: 0000-0002-0670-8160)

Beauty literature have a centralized distribution rule in phases. The emergence and wake-up period of Sleeping Beauty literature between 1960 and 2007 can be respectively divided into three stages, and each has a corresponding mapping relationship, but the chronological order is different. The positive sequence of the wake-up period corresponds to the reverse order of the publication period. 3) The h-index curves of most Sleeping Beauty literature authors show linear growth trends. And there are three most prominent growth slopes, which are 30, 45 and 60 degrees. During the period from publication to wake-up, the gains of h indexes are mostly at 8-12, which shows the author's influence is moderate. It can be inferred the improvement of the author's influence is one of the awakening elements of the Sleeping Beauty literature, but not an important or the only factor, which is consistent with previous studies. 4) Some typical "Sleeping Beauty authors" are found in this study with a special rising law of dynamic h-index curves. Their growth processes are mainly divided into two stages, starting from the publication date, and leaping at the year of Sleeping Beauty literature awakened. The early h index is basically at a continuous low level with the value of 0-2, then it shows a rapid leap after their corresponding Sleeping Beauty literature awakened. It can be seen that the trends of these authors' h index are not simple linear growth, but coincides with that of the Sleeping Beauty literature citation curves. In such a trend, the rise of h index after the awakening of Sleeping Beauty literature can deduce that the arousal elements of this type of Sleeping Beauty literature do not include the author's influence.

The data collection in this paper is limited to 1960-2007. For the earlier Sleeping Beauty literature, the rule of authors' h-index cannot be speculated. Secondly, the identification criteria for non-parameter indicators needs to be further improved by finding other theoretical basis, and further research on different awakening elements of Sleeping Beauty literature can be carried out. 6 figs. 1 tab. 35 refs.

KEY WORDS

Sleeping Beauty literature. Time distribution. Dynamic h index. Author's influence. Bcp index.

0 引言

“睡美人”文献是指论文发表之初未受到学术界广泛关注或被抵制,一直处于低被引或零被引的状况,之后由于各种原因被“王子”文献唤醒,继而被引频次出现激增的文献^[1]。这些文献之所以低被引或零被引,通常是因为其内容具有变革性或超前性,遭到学术界和社会的忽略或抵制,形成了延迟承认的现象^[2]。通常这些睡美人文献大都代表着领域重大科学发现或转折,因此研究睡美人文献对促进学科发展具有重要意义。睡美人文献由文献内容和作者两部分构成,本文通过代表作者影响力的h指数这一指标,研究睡美人文献作者,发现睡美人

文献作者的动态h指数的变化规律及h指数特征,为睡美人作者的特征识别提供线索。

1 文献综述

1.1 睡美人文献识别方法

20世纪六七十年代,学者开始通过引文曲线关注低被引或零被引文献以及延迟承认现象^[3]。早期对睡美人文献的识别方法集中于曲线拟合^[4],将论文引文曲线与数学公式或具体曲线类型进行拟合,人为观察曲线走势^[5]。中期大多数学者利用人为参数指标界定睡美人文献。Garfield通过对前期平均每年被引次数接近1且低被引时间大于5年的方法识别了5篇典型睡美人文献^[6]。2004年,荷兰科学计量学者

Raan 提出了更加被普遍使用的识别标准,分别从睡眠深度、睡眠时长、唤醒强度三方面进行设定,睡眠时长一般大于等于 5 年,在文献睡眠期间,睡眠深度即平均每年被引频次介于 0—2 次之间,唤醒强度即被唤醒的前 4 年累计被引频次一般应大于 20 次^[1]。后来学者们又利用平均值标准和四分位数标准限定睡美人文献^[7],也有将以上多种标准结合使用的方法^[8]。基于这些评判标准,李江和叶鹰两位学者还提出了四条评判原则,即综合考虑全部引文历史,相同引文总量不同积累过程的文章应弱化前期累积引文量的影响,睡美人的唤醒时间不应随时间而变化,以及避免睡眠期间或觉醒强度的任意阈值^[9]。

近年来识别睡美人文献多使用无参数指标,即综合考虑文献的相关引文数据来判定睡美人文献。2013 年 Wang 提出被引速率,即文献被引频次的累计增加速率^[10];2015 年 Ke 等人提出“美丽指数(B 指数)”以衡量睡美人文献的“美丽程度”,即描述睡美人文献三个相关参数的表现程度^[11];2017 年杜建提出 Bcp 指数,以“年度被引次数累计百分比”重新计算睡眠深度、唤醒时间和唤醒强度^[12];2018 年叶鹰和 Bornmann 以及国内学者张家榕等引入“引文角”概念测度睡美人文献^[13-14]。目前较为前沿的研究是将机器学习方法引入到睡美人文献的识别中。2017 年 Dey 等人利用机器学习的分类器方法,使用由文献作者、出版地点、关键词和每篇论文的引用情况等元数据所组成的特征集,识别计算机科学领域的睡美人文献并总结其沉睡因素^[15]。同年 Ho 等人利用心理学计算机检索新方法从 30 多万篇心理学文章中识别出三篇典型睡美人文献及其王子文献,分析心理学中睡美人文献的细节特征^[16-17]。

上述方法中,曲线拟合法由人为观察决定,结果偏主观,参数标准对识别睡美人文献的沉睡期要求苛刻,相比于无参数标准公式计算的识别结果,沉睡期年均被引次数大于 2 次但也是低被引的文献可能会被忽略。综合考虑所有

识别方法的优劣特征,本文采用 Bcp 指数法计算 *Science*、*Nature* 和 *PNAS* 期刊中高被引文献的“美丽程度”,并结合 Bcp 指数识别标准,依据计算结果识别出睡美人文献,构建睡美人文献样本,作为文献特征研究的基础。

本文旨在通过研究睡美人文献的特征来完成后续的预测工作。但现有的对睡美人文献的预测研究仍处在探索阶段,集中于开发预测模型。潜在的睡美人文献,现阶段处于低被引状态,除了同行评议外,无法与普通低被引文献区分开;但同行评议仍面临多重问题,如工作量较大,或存在变革性、创造性的新思想和新发现继续被抵制的可能性。因此,本文提出从文献的创造者——即作者的角度出发,研究发现睡美人文献作者具有的共同特征,利用特征指标,识别现有低被引文献的作者,进而实现对睡美人文献的预测和挖掘。

1.2 动态 h 指数

h 指数是 2005 年被美国加州大学圣地亚哥分校物理系 Hirsch 教授提出^[18],作为评估科研人员成果与作者影响力的主要评价指标之一,其综合考量学术产出数量与学术产出水平,客观反映作者在某领域的被认可程度。h 指数是指某一学者在一定期间内所发表的论文至少有 h 篇的被引频次不低于 h 次。关于 h 指数的改进也存在大量研究,但其计算相对复杂,且并未得到广泛运用,因此本文仍以 h 指数作为作者影响力的评价指标。多数学者对 h 指数的研究更多地关注当时 h 指数的意义,忽略了 h 指数的时间动态变化趋势。只有少数学者对动态指数有进一步研究。Rousseau 和 Ye 利用 R 指数与有理数 h 指数基于连续时间变量的导函数的乘积定义了动态 h 指数^[19];Egghe 基于引文老化速率和 Lotka 指数提出了 h 指数的时间函数模型^[20];张晓阳以分子生物学与遗传学的 14 位高被引科学家为例,发现高被引科学家的 h 指数的线性成长规律,以及学术生涯结束后 h 指数的对数成长特征^[21]。吕娜通过 ESI 平台选定物

理学领域篇均被引次数排名前 5 位的作者为研究对象,总结出 3 种 h 指数成长曲线类型^[22],并提出作者贡献程度与作者影响力即 h 指数的关系^[23]。张琳等分别对国内外 8 位高被引作者进行 h 指数趋势监测,指出团队合作者之间 h 指数的相对变化规律^[24]。

在睡美人文献领域研究动态 h 指数还鲜有人涉及,本文从该角度探讨睡美人文献在作者影响力方面的特征及变化趋势,深入了解睡美人文献及其作者的内在性质,有利于后期对睡美人文献的甄别。

2 数据来源与研究方法

2.1 数据来源

睡美人文献研究对象的前提是高被引文献。本文挑选了 *Nature*, *Science*, *PNAS* (*Proceedings of the National Academy of Sciences of the United States of America*) 三大名刊,囊括了自然科学和社会科学两大领域,研究主题覆盖全面,同时杜建的博士论文对相关研究的统计结果也称上述三本期刊是产出睡美人文献最多的期刊^[2]。本文通过 Web of Science (WoS) 平台数据库,以 *Science*, *Nature* 和 *PNAS* 杂志自 1960 年以来发表的论文的年度被引频次为基础数据,文献被引频次下载日期为 2018 年 3 月 5 日,发表文献数据统计截至 2007 年 12 月 31 日。为观察文献的年被引频次变化趋势,本文以 10 年为例,保证每篇文献至少有 10 年的引文窗,由于 2018 年被引频次数据不完整,因此将检索时间条件设定为 1960—2007 年。检索式为 SO = (science OR nature OR PNAS),共得到有效统计数据 270 639 万条。上述三本期刊的 5 年影响因子分别为 38.062、43.769 和 10.414,取平均值 30.748,即平均每年每篇文献被引 30.748 次。以 10 年引文窗为例,取总被引频次大于 307 次的文献作为高被引文献,共计 23 414 条,构成本文的研究文献列表。该文献列表可使用 WoS 平台提供的“创建引文报告”功能及下载功能,获得每篇文

献自发表年起至检索日期的历年被引频次,构成本文的文献引用数据库。

计算 h 指数所需要的作者历年的引用数据仍通过 Web of Science 平台收集。在收集某一作者的文献成果时,为了排除重名作者的文章,特添加“所属机构”限定,但由于该作者在学术生涯期间有可能更换工作单位,因此还要参考作者的研究方向,同时也可以通过文献的其他合作者协同确定,因为他们的合作次数可能不止一次。综上,本文采用“姓名+所属机构+研究方向+期刊领域”的检索策略限定作者的查准率,为确保囊括作者所有成果,检索时间设定为“所有年份”。利用 WoS 平台提供的“创建引文报告”功能获得作者的每篇文献的历年引文数据,数据收集自该作者首篇文献的发表年起。相较于前人的动态 h 指数研究,本文收集了更大规模的数据,选取睡美人文献数据库中 Bcp 指数大于 10 的 77 篇文献的 129 位作者,进行数据统计和分析,结果更具普适性。

2.2 研究方法

2.2.1 睡美人文献识别及唤醒计算

本文通过对相关识别方法优劣的对比分析,选取较为完善的无参数指标 Bcp 指数对数据进行睡美人文献识别,用“年度被引次数累计百分比”重新计算睡眠深度、唤醒时间和唤醒强度。计算公式如下:

$$Bcp = \sum_{t=0}^{t_m} \left(\frac{1-c_0}{t_m} \cdot t + c_0 - c_t \right)$$

论文发表当年被引次数为 c_0 , 年度被引频次累计百分比 $p \in [0, 1]$, t_m 为文献累计百分比达到最高点时的年龄,即该文献的总生命周期, l 为 t_m 点纵坐标和 c_0 点纵坐标的连线;任意 $t < t_m$, c_t 为年龄 t 的累积被引频次百分比, l_t 为年龄 t 对应连线上的点,对 $(l_t - c_t)$ 求和,即为 Bcp 指数值。

借鉴 B 指数和 Bcp 指数对唤醒期的计算原理,累计被引百分比曲线各点向曲线起点与终点的连线即 l 作垂线,取使得垂线距离最大时的

年龄 t 对应的年份作为被唤醒年,对公式稍作修正,得到公式如下: $t \in [0, t_m]$,

$$\frac{1 - c_0}{t_m} = \frac{\sqrt{\left(\frac{1 - c_0}{t_m} \cdot t + c_0 - c_t\right)^2 - d_t^2}}{d_t}$$

$$\max d_t = \frac{(1 - c_0) \cdot t + (c_0 - c_t) \cdot t_m}{\sqrt{(1 - c_0)^2 + t_m^2}}$$

2.2.2 无参数指标 Bcp 识别标准

现有的无参数指标均以计算指数大小为主要目的,比较指数之间的相对有效性,而指数的数值只能衡量睡美人的“美丽程度”,并不能给出确切的标准界定是否为睡美人文献。因此若想使用无参数指标识别出睡美人文献,对睡美人文献进行总体分析,需要具体的阈值来划定睡美人文献的 Bcp 指数范围。

本文结合参数指标和无参数指标设定关于 Bcp 指数的睡美人文献识别标准,即按照普遍认可的 Raan 提出的参数标准,利用 Bcp 指数计算公式,计算得到满足睡美人文献定义的文献 Bcp 指数阈值,即睡美人文献的识别标准。假设有两篇文献恰好符合 Raan 提出的参数标准的上限和下限,设定沉睡期取最小值 5 年;沉睡期每年被引频次为 1—2 次,分别取两个数组 $\{1, 1, 1, 1, 1\}$ 和 $\{2, 2, 2, 2, 2\}$ 进行计算并求平均值;前 4 年累积被引频次大于等于 20 次,则取前 4 年共 20 次,平均以每年 5 次累加,构成年被引频次矩阵 $\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 5 & 5 & 5 & 5 \\ 2 & 2 & 2 & 2 & 2 & 5 & 5 & 5 & 5 \end{bmatrix}$,两篇文献

9 年内的总被引频次为 $\begin{pmatrix} 25 \\ 30 \end{pmatrix}$,年被引频次累加

比例矩阵 c_t 为 $\begin{bmatrix} 0.04 & 0.08 & 0.12 \\ 0.0667 & 0.1334 & 0.2001 \\ 0.16 & 0.20 & 0.40 & 0.60 & 0.80 \\ 0.2668 & 0.3335 & 0.5002 & 0.0669 & 0.8336 \\ 1.00 \\ 1.0003 \end{bmatrix}$,文献年龄取整数 $t \in [0, 8]$, t_m 为 8。

根据 Bcp 计算公式,得到 Bcp 指数向量 $\begin{pmatrix} 3.68 \\ 3.79985 \end{pmatrix}$,选定其平均值 $Bcp = 3.739925$ 为

最低“睡美人”文献标准。

该标准兼顾无参数指标的高查全率和客观性优势,避免参数指标对于沉睡期的严苛性和主观性偏差,同时给出了使用无参数指标的判定标准,由于现有相关研究成果匮乏,此标准的理论依据较为薄弱,仍需进一步实践验证。

2.2.3 动态 h 指数

h 指数会随文章发表量和被引频次的增长时刻发生变化,因此本文将动态 h 指数应用于睡美人文献的作者特征研究。而 h 指数及相关改进指数均以当前时间点定义计算方法,若想获取历年 h 指数需尊重 h 指数的原始定义,按照初始计算方法,以文献篇数为单位,取 h 指数为整数,而非连续有理数,人工收集文献历年的总被引频次,并逐年累加,得到截至每年年底的累计被引频次,再对截至每年年底已发文献按照被引频次排序,进而可得到睡美人文献作者的历年 h 指数。

动态 h 指数计算方法如下:①检索某一作者发表的所有文献集合 D_m ;②对于 D 集中文献 $D_i (i=1 \cdots n)$,逐篇提取出版年并检索形成其引证文献集合 C_{im} ;③对于 C 集中的文献 $C_j (j=1 \cdots n)$,逐篇统计其出版年;④对于某一时间节点 T ,可以计算出该时间节点作者发表的文献集合 D_t ,从 C 集中找出 D_t 中每篇文献的引证文献并提取时间节点 T 之前的引证文献集合 C_t 及引证文献数量;⑤将步骤④中 D_t 集中的文献按照引证文献数量由高到低进行排序,根据传统 h 指数计算方式即可获得该时间节点作者的 h 指数^[22]。

由此,本文通过观察睡美人文献发表年的作者 h 指数值的共同特征、发表年到唤醒年间 h 指数的走势、唤醒年 h 指数的变化,以及合作者之间的 h 指数比较与相互影响,归纳总结睡美人文献作者的内在特征与成长规律。

3 结果与讨论

3.1 基于 Bcp 标准的睡美人文献识别

以所收集的期刊高被引文献为数据基础,

利用 Bcp 指数的计算方法,将 Bcp 指数结果由大到小排序,结合无参数指标 Bcp 识别标准,进行最终的睡美人文献识别。

通过计算,共有 558 篇满足睡美人文献标准,占总篇数的 2.38%,而李江等人于 2014 年发表的 *A study of the "heartbeat spectra" for "sleeping beauties"* 一文中利用 Van Raan 的参数标准识别得到的比例为 1.29%^[25],由于文献来源的不同、时间范围的扩大以及 Bcp 指数对曲线整体趋势的兼顾等多个因素,相比之下本文增多 1%,也在一定程度上验证了本文构建的识别标准具有一定的科学性和可行性。本文计算得到的睡美人文献比例与定义高被引文献所采用的前 1%

的标准相比,是其 2 倍之多,可表明睡美人文献不在少数。本文只对 1960 至 2007 年被认可的睡美人文献进行统计分析,仍有大量正在沉睡的睡美人文献等待“王子”的出现,发挥其价值,推动社会发展和科技进步。因此,加强对睡美人文献沉睡原因和唤醒机制的研究,对预测睡美人文献、提前人为寻找“王子”来唤醒的“睡美人”,具有重要意义。

本文摘取 Bcp 指数大于 10 的 77 篇睡美人文献在此展示,并构成最终研究睡美人文献作者的样本数据库,以备后续 h 指数研究使用。文献结果数据见表 1。

表 1 Bcp 指数计算结果列表

序号	Bcp	序号	Bcp	序号	Bcp	序号	Bcp	序号	Bcp
1 112	18.237	9 241	14.837	9 437	12.787	9 186	11.619	9 554	10.782
22 222	17.434	10 887	14.803	9 647	12.782	10 017	11.611	9 225	10.685
10 258	16.908	9 163	14.713	9 343	12.702	9 167	11.588	9 422	10.613
9 138	16.686	10 040	14.557	9 272	12.581	9 212	11.539	10 516	10.535
10 385	16.215	9 366	14.461	9 850	12.537	10 225	11.424	10 132	10.41
9 698	16.185	9 686	14.419	10 041	12.435	9 255	11.414	4 359	10.388
9 128	15.562	9 191	14.195	13 976	12.42	9 306	11.373	10 098	10.265
9 181	15.542	9 585	14.132	10 638	12.385	9 963	11.37	17 410	10.264
9 330	15.535	9 215	13.87	10 430	12.357	11 286	11.326	9 250	10.216
10 305	15.451	9 213	13.59	9 800	12.235	9 199	11.192	10 014	10.214
9 271	15.396	10 123	13.495	10 536	12.168	9 535	11.012	9 903	10.115
9 260	15.186	17 996	13.325	9 354	12.076	10 675	10.948	9 963	10.065
21 803	15.036	9 130	13.311	24 337	11.968	9 774	10.919	22 520	10.033
9 428	14.975	10 190	13.12	9 335	11.872	21 316	10.871		
9 281	14.944	10 708	13.022	9 137	11.74	9 224	10.854		
10 489	14.926	11 250	12.796	9 198	11.682	160	10.788		

本文所识别的 558 篇睡美人文献样本涉及学科领域主要有医学、生物科学、化学、物理学、心理学和海洋科学等^[26],分布领域广泛,其所蕴含的学术价值也不可估量。文献作者的所属机构大多为美国、英国和加拿大等发达国家,其中

美国以绝对优势领先于其他国家。其中 Bcp 指数最高的文章是来自 *Science* 杂志的美国学者 Schoenfeld 发表的 *Lysergic acid diethylamide- and mescaline-induced attenuation of the effect of punishment in the rat*, Bcp 指数为 18.237。在 1976

年发表后的 4 至 6 年间,有两次年被引频次分别达到 3 和 5(不排除自引的情况),从此之后沉睡深度一直保持在 0—2 之间,说明当时有极少数人关注过此研究,并未被广泛传播和认可,随后进入深度睡眠时期,沉睡期长达 40 年之久,直到 2016 年才被唤醒,年度被引频次达到 118 次,2017 年被引频次更是高达 753 次,说明在近两年内其成果才开始在学术界扩散且势头迅猛,得到广泛认可,为生物学领域带来革命性的转折。

为了奠定所识别的睡美人文献时间分布规律的研究基础,本文先通过图 1 对睡美人文献的总体时间分布进行描述。从期刊方面观察, *Science* 在 1995 年前为三本期刊中睡美人文献出现概率最高的期刊,但其分化明显,在 1995 年后鲜少再有睡美人文献出现的高峰; *Nature* 前 25 年紧随 *Science* 之后,一直处于较高水平,仅在 1986—1996 的 10 年间表现平平,但在 1996 年后

反增趋势明显,领先其他期刊 2—3 倍,整体上 *Nature* 与年度综合的趋势相吻合,也侧面反映了 *Nature* 在其中的带头作用; *PNAS* 的睡美人文献数量相比前两者一直处于起伏不定的低水平状态,对整体趋势影响甚小。三者比较,不同期刊睡美人文献出现频率的高低可能与期刊影响因子存在正相关性^[27]。从折线总体趋势观察,睡美人文献的出现总体分布在 1964—1984 的 20 年间,其中有三个明显的高峰期,第一个是 1965 年前后,增长和消减的速度都很快,均为一年内变化,三本期刊分别出现了各自的波峰,与年度综合曲线走势相同;1970 年为第二高峰,该波峰增长幅度略大于第一个高峰期,且持续时间较长,可达 3—5 年;第三个高峰为 1977 年,也是增长最强势的一个,达到所有检索时间的最高点,但也很快下降,紧接着 1980 年和 1982 年又反弹出现小波峰,随后睡美人文献的出现一直处于较低水平,相对前期而言仅有些许起伏。

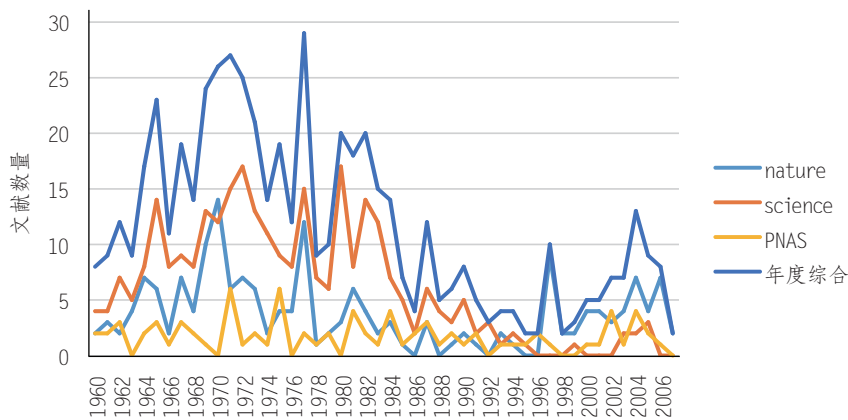


图 1 期刊睡美人文献时间分布图

睡美人文献出现的原因与社会发展的进程有关。第二次世界大战结束后,发达国家经过数年的休整,社会环境安定平稳,而各国为了恢复国力,开始迫切追求科技和经济的发展,加大对科技和经济的投入,在 18 世纪以来工业革命和电力革命的基础之上,信息需求成了人类征服自然最迫切需要解决的问题^[28],因此世界各国纷纷开始了第三次科技革命,这是一场信息

控制技术的革命。主要以原子能、电子计算机、空间技术和生物工程的发明和应用为标志,涉及信息技术、新能源技术、新材料技术、生物技术、空间技术和海洋技术等诸多领域,与睡美人文献涉及的领域相吻合。20 世纪中叶,尤其是 1965—1975 年,以重组 DNA 实验的成功为标志,进入第三次科技革命的第三个 10 年,生物技术快速发展,人类开始控制遗传和生命过程^[28],

思想先进的学者发表了一些比较前沿但不被认可的研究,也就是后来学术界所定义的睡美人文献。之后随着第三次科技革命成果的逐渐显现,各领域科技水平的提高和信息的爆炸式增长,人们对信息的使用需求和储存技术的提升,使得文献的曝光量增加^[29],睡美人文献逐渐被唤醒。

1990年后睡美人文献相对前三个高峰期有所减少,对此 Raan 和 Winnink 给出了解释:因科技专利的引用,大量的可能成为睡美人的文献被提前唤醒^[30]，“专利王子”逐渐代替了“科学

王子”的作用。此外,现代网络与科技的发达,文献的可获取性逐渐提高^[10],使得因文献传播问题而产生的睡美人文献也相应减少。

3.2 睡美人唤醒期识别

本文首先需要识别睡美人文献的唤醒期,才能总结睡美人文献的出版年和唤醒年两大特殊节点的时间分布规律,并观察文献作者动态 h 指数的特殊节点特征及节点之间的成长规律。通过计算满足标准的 558 篇睡美人文献所得到的唤醒年份,如图 2 所示。

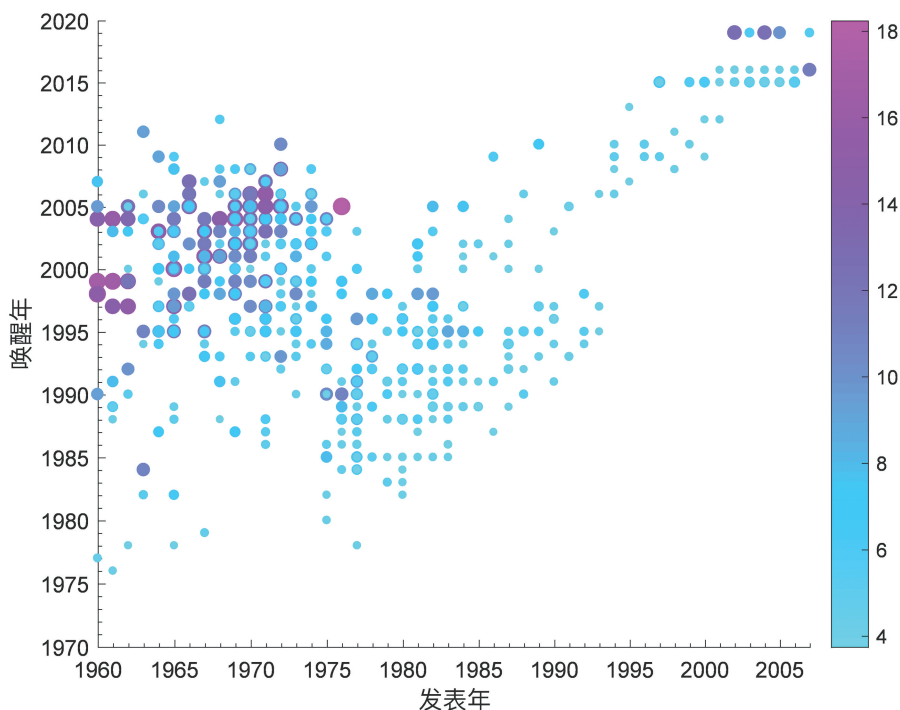


图 2 睡美人文献发表年和唤醒年的分布及文献 Bcp 指数

右侧色标图例的颜色由浅到深代表 Bcp 指数的增加,图中各圆点的颜色和大小均可表示 Bcp 指数的大小。由图 2 观察可见,睡美人文献被唤醒主要分三个时期,1982—1995 年为第一时期,唤醒了大量处于上文分析得到的睡美人发表年第三高峰的文章,文献的 Bcp 指数偏小,仅介于 4—8 之间,原因推测为文献的沉睡期都

在 10—15 年,沉睡时间较短;1998—2006 年为第二时期,集中唤醒了数量更大、沉睡期更久即发表年更早的文章,这批文献主要集中在 1965—1970 年发表,即睡美人文献发表的第一、二高峰,沉睡期大概在 30 年左右,图中颜色偏红色且直径大的圆点大多分布在此区域,说明高质量的睡美人文献(即 Bcp 指数较大)基本集中

在该时期;2000 年至今为第三时期,相对前两个时期唤醒数量大幅下降,文献的发表年份也更晚,且 Bcp 指数大小不等。

从整体上看,满足 Bcp 指数标准的睡美人文献,其沉睡期基本都大于 10 年,1982 年之后文章被大量唤醒,且先唤醒的是时间距离较近的文献。又过了 10 多年,学者们的注意力转移到更早的文献,这也符合知识的涉猎规律,根据解释水平理论,人们对时间距离、空间距离和社会距离较近的事物有具体形象的表征,表示为低解释水平^[31]。对于低解释水平的事物人们更容易产生低解释水平的行为,表现为关于自身的、近期的、简单易行的行为,将该理论应用于文献的发现与挖掘,人们近期更易对低解释水平的事物即时间距离较近的文献产生搜索、浏览等行为,远期才会去发现高解释水平的事物即时间距离较远的文献。而且,知识的发展具有持续性^[32],人们对当前知识的掌握与前不久的先进知识更易产生联系,因此,时间距离较近的文献首先被检索发现。同时随着科学技术和思想的进步,社会和学术界也逐渐认知到被忽略的变革性思想和技术的可行性,被发现的文献随即被认可,即被唤醒后成为睡美人文献。需要说明的是,本文针对的是出现睡美人频率较高的三本期刊的文献,该规律对于其他期刊文献的适用性还有待考证。

3.3 睡美人文献作者动态 h 指数

为了研究睡美人文献作者在特殊节点的影响力特征及其成长规律,本文通过 h 指数代表作者的影响力水平,观察作者的动态 h 指数变化情况。

根据 Web of Science 平台的作者数据计算得到每位作者的历年 h 指数,进行数据整理和筛选。为了使研究特征明显突出,本文选取 Bcp 指数大于等于 10 的 77 篇睡美人文献,按照每篇文献作者贡献度的排名顺序,排名靠后的作者贡献度较小,不具备研究价值,因此对于一篇文献而言,最多抽取前 4 名作者,不足或等于 4 名

的则全部纳入分析范围,共得到 129 名作者作为研究对象。对于 h 指数小于等于 2 的作者来说,其 h 指数浮动不超过 1,自始至终几乎无变化,这对观察其 h 指数的动态变化和走势基本毫无意义,因此本文决定筛选 h 指数小于等于 2 的作者,对其他 116 条数据进行分析,得到 h 指数变化趋势图(见图 3)。图中每条曲线表示该作者从学术生涯开始至今的 h 指数的变化趋势,“x”号标记出该作者在所发表的睡美人文献发表年时的 h 指数值,“x”号集中分布的年份较为靠前,约 1960—1980 年之间,“*”标记出该作者在所发表的睡美人文献被唤醒时的 h 指数值,“*”号集中分布在 1990 年之后。

从图 3 中可以清晰地观察到,睡美人文献的作者们在文献发表时期的影响力基本都不超过 5,而且从图中 h 指数的起始点观察,多数睡美人文献的出现都发生在这些作者的学术生涯初期,基本不超过 10 年,因此可推测,这些作者当时的年龄都在 20—30 岁的青年阶段。以上两点结论也可以充分说明睡美人文献在发表初期被忽略甚至被抵制的原因,这些作者进入研究领域时间较短,成果和被认可程度均处于较低水平,在学术领域不具有权威的发言权。由于“马太效应”的存在,使得他们的新发现并不能在第一时间吸引领域内领军人物的关注,或者支持他们观点的学者人数较少,在他人选择参考文献时就被剔除^[33],没有这些权威的“伯乐”,学术“新秀”们自然会被埋没。

从 h 指数值来看,根据美国科睿唯安(Clari-vate)公布的“高被引科学家”名单及白皮书,查询所有“高被引科学家”的 h 指数,发现 h 指数大于 40 可认定为高被引作者^[34]。截至 2017 年底,在 116 名睡美人文献作者中有 20% 的人可被认为“高被引作者”,大部分作者的 h 指数均处于 30 左右及以下,属正常影响力水平。出乎意料的是,个位数的 h 指数作者(包含 h 指数小于等于 2 的作者)也具有一定比例,常作为第三作者或第四作者出现。本文推断,因作者年龄年轻化特征,多数作者在年轻时期攻读学士、硕

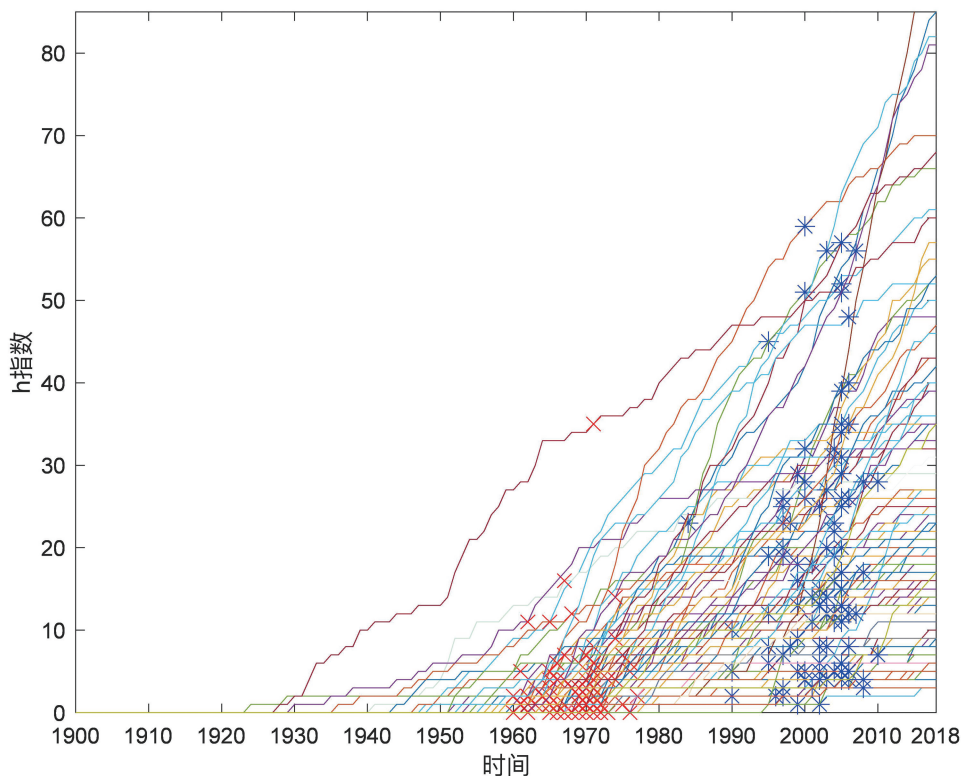


图3 作者h指数走势与分布图

士或博士期间,参与指导教师相关研究,并在过程中作出一定贡献,而学业顺利完成后,可能选择就职于其他行业,没有继续进行更加深入的学术研究,使得一部分作者h指数较低。

3.3.1 动态h指数总体趋势

从h指数曲线走势看,总体睡美人文献作者的h指数均呈线性增长趋势,Burrell在2007年也根据大量数据证实了科学家h指数的线性增长特点^[35],同时国内学者张晓阳也提出分子生物学和遗传学领域的高被引作者h指数呈线性成长规律^[21]。本文在线性增长规律基础上,发现了三个表现突出的成长速率,以倾斜角度30°、45°和60°最为明显,h指数以每十年增加5、10、20的速度上升,分别在其h指数区间[10,20]、[25,35]、[50,60]时被唤醒。因此对于不同h指数成长速度的作者来说,均在其学术生

涯的中期阶段被唤醒,究其原因可能是彼时睡美人文献作者找到了自己感兴趣的领域,并已在该领域发展了一段时间,文献的被认可度逐渐提升,同时也具备了一定学术影响力,综合其他因素,文章即被唤醒。

图4列出了作者在文献发表年和唤醒年之间h指数的成长幅度。依图可见,唤醒年与发表年h指数之差为2—4的人数较多,推测原因主要是部分作者在求学之后并未从事科学研究,h指数受限于文献发表数量,整体水平较低,数量上也具有一定比例,因此使得[2,4]区间人数偏多。正常情况下h指数的增幅分布较多的区间为[8,12],中位数为10,即睡美人文献被唤醒时,作者h指数提升10的概率最大,该数值在所有增幅大小排列中属偏小值,以其h指数判断也鲜有高被引作者(h指数大于40可被认为

是高被引作者),因此可知睡美人文献的作者影响力水平并不需要达到该领域的最高标准,只需具有中等左右的水平即可被关注。作者学术影响力的提升,使其文章受到关注,进而被广泛引用,可见作者影响力也是文章被发现和唤醒

的原因之一,但并非重要因素或唯一因素,因为较低影响力水平的作者发表的睡美人文献也将被唤醒,对于睡美人文献被唤醒的要素还需进一步研究。

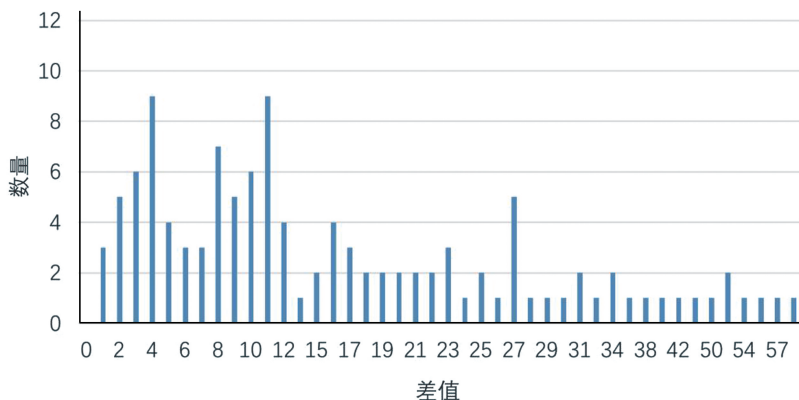


图4 唤醒年与发表年 h 指数之差

3.3.2 典型“睡美人作者”

在针对睡美人文献作者进行研究时,发现了不同于线性增长的部分作者,本文称之为典型“睡美人作者”,这些作者的 h 指数成长趋势类似于睡美人文献的被引趋势,即前期 h 指数基本为持续低水平状态,在其发表的睡美人文献被唤醒后 h 指数呈现迅速飞跃的现象,本文从中摘取了 6 个典型案例进行展示,见图 5。

图 5 和图 6 中蓝色且变化较为平缓的曲线代表该作者的 h 指数的变化趋势,黄色且起伏较大、变动频繁的曲线代表该作者所发表的睡美人文献的引文曲线,在两条曲线中,“x”分别标记出该作者在睡美人文献发表年时的 h 指数值及总被引频次,“*”分别标记出该作者在睡美人文献被唤醒时的 h 指数值及总被引频次。

在每个坐标内,特意标明睡美人文献唤醒年份,并给出该文献引文曲线作对照。观察两条曲线的变化趋势可知,其在时间分布上具有一致性,即在睡美人文献被引频次发生突变时,该作者的 h 指数也在该点之后发生激增现象。

该现象并非是简单的线性成长,其成长过

程主要分为两阶段,通常以睡美人文献发表年为起点,睡美人文献唤醒年为两阶段转折点。第一阶段是“被埋没的作者”,h 指数一直处于 0—2 区间,作者影响力水平很低;第二阶段则是从转折点至今的“飞跃期”,在转折点后的 1—5 年期间,h 指数的涨幅可达到前期的 10 倍,甚至更高。其主要原因还是在于作者文献的被引情况。在作者发表睡美人文献后,该文献被引情况不佳,即使作者继续发表更多的文章,但是受制于每篇文献的被引频次,该作者的 h 指数并没有得到提升;在该文献被唤醒之后,与该文献相关的一系列文献也将被关注,因而该文献作者的 h 指数曲线则会瞬间上升,这种情况符合睡美人文献的被引规律,与睡美人文献的引文曲线基本吻合。其中,图 5 中(3)和(6)作者的 h 指数激增趋势不明显,因该作者的 h 指数相对引文的被引频次数值较小,在同一坐标轴中显示时,其趋势被掩盖,因此在图 6 中分别画出了作者 h 指数曲线和睡美人文献的引文曲线,以此观察可以确定(3)和(6)作者属于典型的“睡美人作者”。(4)作者在唤醒年时的 h 指数

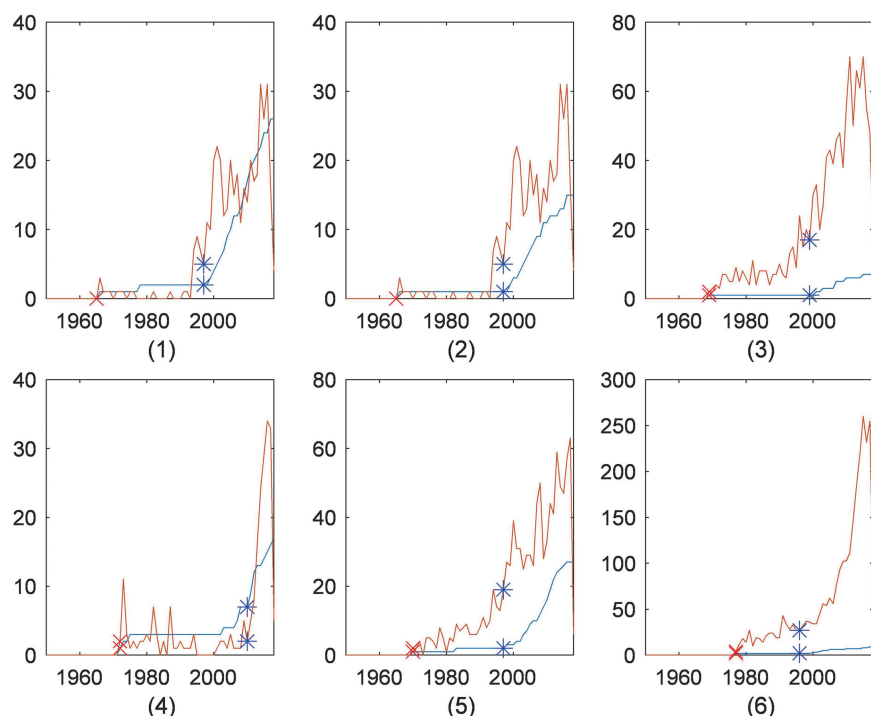


图5 典型睡美人作者 h 指数与睡美人文献引文曲线案例

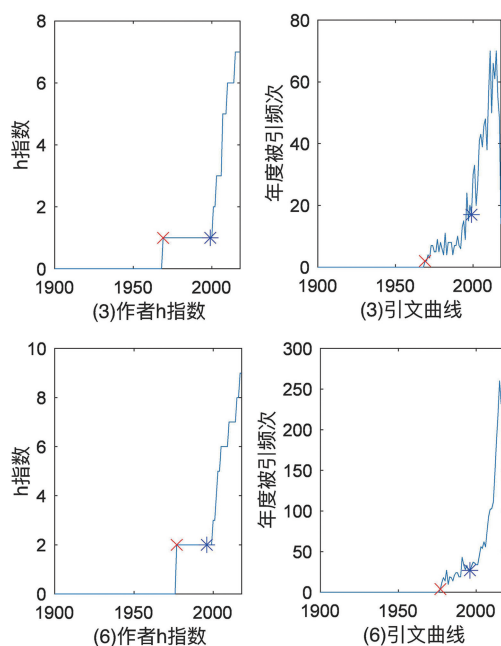


图6 (3)和(6)作者 h 指数与引文曲线

没有像其他文献一样恰好处在转折点上,但也在 h 指数迅猛上升的初期,其原因可能是由于睡美人文献在被唤醒前期,其被引频次也有小幅度的累加,使得该作者的 h 指数也有所增加。由于其 h 指数的飙升是在睡美人文献被唤醒之后,睡美人文献被引频次的增加才带动了作者 h 指数的提升,说明在文献被唤醒时,该作者还不具备一定的学术影响力水平,可推断这一类作者的睡美人文献被唤醒要素并非作者影响力的提升,对于此类睡美人文献的唤醒要素还可进行更深层次的研究。

4 结论

本文针对睡美人文献时间分布和睡美人文献作者动态 h 指数进行研究,分别得出以下结论。

(1)以所收集的 Web of Science 平台 *Science*、

Nature 和 *PNAS* 期刊数据为基础,分析睡美人文献识别及睡美人文献时间分布规律,得到以下结论。①以 Van Raan 的参数标准和 Bcp 无参数指标的计算方法为理论依据,构建了睡美人文献的 Bcp 无参数指标识别标准。②利用 Bcp 指数公式计算睡美人文献的“美丽程度”,并根据 Bcp 无参数指标识别标准,识别出 558 篇睡美人文献,比例约为总篇数的 2.383%,该结果与李江等人识别的睡美人文献比例比较吻合,可验证所构建的标准具有一定的可行性。③睡美人文献的出版与唤醒时间分布规律具有阶段性特征。1960—2007 年之间睡美人文献的出现分为三个阶段,唤醒期分布也主要分为三个阶段,并与出现的三个阶段相互对应,但先后次序不同,“唤醒一期”先唤醒了“出现三期”的文章,后唤醒了“出现一期”和“出现二期”的文章。

(2)通过计算睡美人文献作者的动态 h 指数,研究分析睡美人文献作者影响力特征及成长规律,得到以下结论。①多数睡美人文献作者 h 指数成长规律总体呈线性趋势,与高被引作者的 h 指数成长线性规律相同,且有三个成

长斜率的表现最为突出,分别是倾斜角度为 30°、45°和 60°,在发表至唤醒的时间段内,作者 h 指数的成长幅度以 8—12 最为明显,该增幅属偏小值,作者影响力属中等水平,可推断作者影响力的提高是睡美人文献唤醒的要素之一,但并非是睡美人文献被唤醒的重要因素或唯一因素。②发现典型“睡美人作者”的特殊成长规律。研究发现,部分作者的动态 h 指数成长曲线并非简单的线性成长,而是与其发表的睡美人文献的年度被引曲线的变化趋势相吻合,呈现出前期 h 指数成长速度很低或一直处于 0—2 的水平,后期睡美人文献的引文曲线与 h 指数同时得到飙升的变化规律。由于 h 指数的上升与文献被唤醒同时发生,可推断该类睡美人文献的唤醒要素不包括作者影响力。

本文对数据的收集只限于 1960 至 2007 年,对于更早期的睡美人文献未收录在内,其睡美人作者 h 指数的变化规律也无法推测;另外对于无参数指标的识别标准,有待进一步寻找其他理论依据进行完善,对于不同睡美人文献的唤醒要素还可进行深入研究。

参考文献

- [1] Raan A F J V. Sleeping beauties in science[J]. *Scientometrics*, 2004, 59(3):467-472.
- [2] 杜建. “睡美人”文献的识别方法与唤醒机制研究[D]. 南京: 南京大学, 2017. (Du Jian. A study on systematic identification of “sleeping beauty” publications and on their awaking mechanisms[D]. Nanjing: Nanjing University, 2017.)
- [3] Merton R K. Resistance to the systematic study of multiple discoveries in science[J]. *European Journal of Sociology*, 1963, 4(2):237-282.
- [4] 杜建, 武夷山. 睡美人与王子文献的识别方法研究[J]. *图书情报工作*, 2015, 59(19):84-92. (Du Jian, Wu Yishan. Defining and identifying the sleeping beauty and her princes in science[J]. *Library and Information Service*, 2015, 59(19):84-92.)
- [5] Barber B. Resistance by scientists to scientific discovery[J]. *Science*, 1961, 134(3479):596-602.
- [6] Garfield E. Delayed recognition in scientific discovery: citation frequency analysis aids the search for case histories[J]. *Current Contents*, 1989, 23:3-9.
- [7] Glänzel W, Schlemmer B, Thijs B. Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon[J]. *Scientometrics*, 2003, 58(3):571-586.
- [8] Li J, Shi D. Sleeping beauties in genius work: when were they awakened?[J]. *Journal of the Association for Infor-*

- mation Science and Technology, 2016, 67(2):432-440.
- [9] Li J, Ye F Y. Distinguishing sleeping beauties in science[J]. *Scientometrics*, 2016, 108(2):821-828.
- [10] Wang J. Citation time window choice for research impact evaluation[J]. *Scientometrics*, 2013, 94(3):851-872.
- [11] Ke Q, Ferrara E, Radicchi F, et al. Defining and identifying sleeping beauties in science[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2015, 112(24):7426-7431.
- [12] 杜建, 武夷山. 一个用于识别睡美人文献的新的无参数指标——基于“Science”和“Nature”上睡美人文献的验证[J]. *情报理论与实践*, 2017, 40(2):19-25. (Du Jian, Wu Yishan. A new parameter free index for identifying sleeping beauty literature: verification of sleeping beauty literature based on *Science* and *Nature*[J]. *Information studies: Theory & Application*, 2017, 40(2):19-25.)
- [13] Ye F Y, Bornmann L. “Smart girls” versus “sleeping beauties” in the sciences: the identification of instant and delayed recognition by using the citation angle[J]. *Journal of the Association for Information Science and Technology*, 2018, 69(3):359-367.
- [14] 张家榕, 曾继城, 叶鹰. 3S 引文现象的特征测度及学术意义——“睡美人”、“时髦女”与“天鹅”综论[J]. *情报学报*, 2017, 36(12):1241-1246. (Zhang R J, Zeng C J, Ye F Y. Characteristic measures of “3S” special citation phenomena with academic implications: a synthetic essay on “sleeping beauty”, “smart girl” and “swan” [J]. *Journal of the China Society for Scientific and Technical Information*, 2017, 36(12):1241-1246.)
- [15] Dey R, Roy A, Chakraborty T, et al. Sleeping beauties in Computer Science: characterization and early identification[J]. *Scientometrics*, 2017, 113(5439):1-19.
- [16] Ho Y S, Hartley J. Sleeping beauties in psychology[J]. *Scientometrics*, 2017, 110:1-5.
- [17] Hartley J, Ho Y S. Who woke the sleeping beauties in psychology?[J]. *Scientometrics*, 2017, 112(3):1-4.
- [18] 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.
- [19] Rousseau R, Ye F Y. A proposal for a dynamic h-type index[J]. *Journal of the American Society for Information Science & Technology*, 2008, 59(11):1853-1855.
- [20] Egghe L. Dynamic h-index: the Hirsch index in function of time; brief communication[J]. *Journal of the American Society for Information Science and Technology*, 2007, 58(3):452-454.
- [21] 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[J]. *科学学研究*, 2007(3):407-414. (Zhang Xiaoyang, Jin Bihui. A study on h-index increasing of highly cited scientists in molecular biology & genetics[J]. *Studies in Science of Science*, 2007(3):407-414.)
- [22] 吕娜. 科研人员 h 指数动态变化趋势分析[J]. *情报理论与实践*, 2015, 38(5):112-115. (Lü Na. Analysis of the h-index dynamic changing trend of scientific research personnel[J]. *Information Studies: Theory & Application*, 2015, 38(5):112-115.)
- [23] 吕娜, 刘扬, 全少颖. 基于作者主要贡献的 h 指数时间趋势分析[J]. *情报杂志*, 2015, 34(4):54-58. (Lü Na, Liu Yang, Quan Shaoying. H-index time trends analysis based on author contribution[J]. *Journal of Intelligence*, 2015, 34(4):54-58.)
- [24] 张琳, 董颖, 毕德强, 等. 基于动态数据整合的学者影响力 h 指数趋势监测与分析[J]. *图书情报工作*, 2017, 61(17):116-121. (Zhang Lin, Dong Ying, Bi Deqiang, et al. H-index trend monitoring and analysis of scholars' influence based on dynamic data integration[J]. *Library and Information Service*, 2017, 61(17):116-121.)
- [25] Li J, Shi D, Zhao X, et al. A study of the “heartbeat spectra” for “sleeping beauties”[J]. *Journal of Informet-*

- rics, 2014, 8(3):493-502.
- [26] Williams A. Sleeping beauties of science[J]. Scientific American, 2016, 314(1):80-80.
- [27] Kokol P, Blažun H V, Vermeulen J. Exploring an unknown territory: “sleeping beauties” in the nursing research literature[J]. Nursing Research, 2017, 66(5):359-367.
- [28] 孙寿涛.信息革命:称谓及其历史地位[J].北京邮电大学学报(社会科学版),2007(2):55-60. (Sun Shoutao. On the title and historical status of the information revolution[J]. Journal of Beijing University of Posts and Telecommunications(Social Sciences Edition),2007(2):55-60.)
- [29] Wang J, Ma F, Chen M, et al. Why and how can “sleeping beauties” be awakened?[J]. Electronic Library, 2013, 30(1):5-18.
- [30] Raan A F J V, Winnink J J. Do younger sleeping beauties prefer a technological prince?[J]. Scientometrics, 2018, 114(2):701-717.
- [31] 杜伟宇,蒋运霞.消费者的解释水平对其选择偏好的影响[J].心理研究,2014,7(1):15-22. (Du Weiyu, Jiang Yunxia. Effect of consumer's construal level on the choice preference[J]. Psychological Research, 2014, 7(1):15-22.)
- [32] 王亮.基于SCI引文网络的知识扩散研究[D].哈尔滨:哈尔滨工业大学,2014. (Wang Liang. Research on knowledge diffusion based on SCI citation network[D]. Harbin:Harbin Institute of Technology,2014.)
- [33] 郭斐,鄢小燕.“睡美人”文献研究综述[J].图书馆建设,2016(5):40-45. (Guo Fei, Yan Xiaoyan. Research review of sleeping beauty in science[J]. Library Development,2016(5):40-45.)
- [34] Tehmina A,Ding Y,Xu J, et al. Standing on the shoulders of giants[J].Journal of Informetrics.2017,11(1):307-323.
- [35] Burrell Q L. Some comments on “A Proposal for a Dynamic h-Type Index” by Rousseau and Ye[J].Journal of the American Society for Information Science and Technology,2009,60(2):418-419.

李 贺 吉林大学管理学院信息管理系教授。吉林 长春 130022。

解梦凡 吉林大学管理学院情报学硕士研究生。吉林 长春 130022。

袁翠敏 吉林大学管理学院图书情报与档案管理专业博士研究生。吉林 长春 130022。

余 璐 吉林大学管理学院情报学硕士研究生。吉林 长春 130022。

(收稿日期:2018-04-15;修回日期:2018-06-25)