测度论文的“延迟认可”

闵超
南京大学

案例：强连接与弱连接

- Granoveter interviewed people who recently changed jobs and asked them how they discovered their new job
- Findings:
  1) The job was found through personal contacts
  2) The contacts were “acquaintances” instead of “friends”
- Question: how come?
- Guess: acquaintances may have access to different sources of information than friends

讨论：你觉得什么是信息扩散、知识扩散

- 多样的学科视角
- 信息
- 扩散

科学知识的扩散与采纳

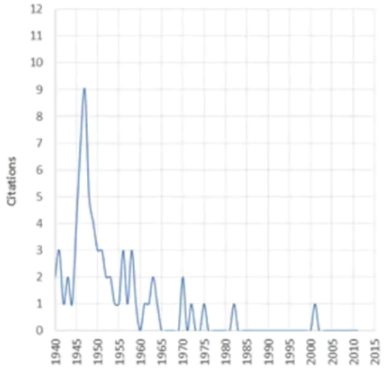
- 根据Kuhn(1962)的理论，科学的发展时常会遭遇偶发的“异常”现象，这些“异常”现象受到社会的、常理无法解释的因素影响，反过来也影响着新的科学发现被人们接受认可的过程。
- 因此，“常规的”科学进程常被视为对现有理论渐次积累的发展过程，而当新理论与常规科学或者其标准做法相悖时，旧数据中可能会涌现出新问题。这最终可能导致激烈的***范式转移***甚至***科学革命***。

什么是“延迟认可”

- “被抵制的发现” (Barber, 1961),“科学上的早熟” (Stent, 1972),“孟德尔效应”(Garfield, 1979; Costas et al., 2011)以及“睡美人” (van Raan, 2004)
- 一篇重要的论文在发表很长时间之后才被人们认识到它的价值

Sleeping Beauties in Science

- Some important discoveries went unnoticed for a long time, and then almost suddenly began to receive much attention.
  - Mendel discovered in 1866 the principle of segregation of hereditary units. But his 1866 paper remained unappreciated for **34 years** until it was “rediscovered” in 1900.

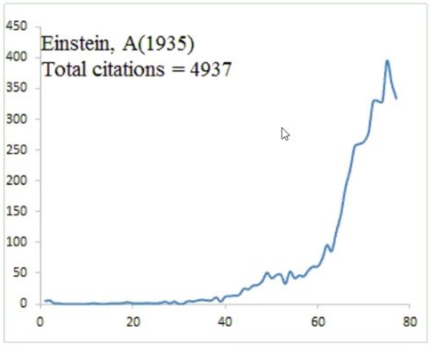


The citation curve of an ordinary paper

Sleeping Beauties in Science

- These publications are also referred to as "premature discoveries" (Wyatt, 1961), "resisted discoveries" (Barber, 1961), "***delayed recognition***" (Cole, 1970), and "***sleeping beauties***" (van Raan, 2004).
- The citation trace exhibits a dormancy period followed by a sudden or gradual lifting of citation counts, just like a princess sleeps for a long time and then is awakened by a prince.

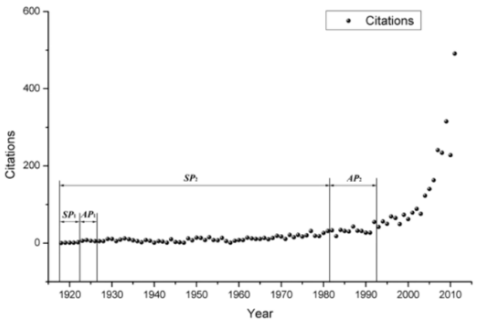
The fairy of Sleeping Beauty



Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete?. *Physical review*, *47*(10), 777.

如何定量测度科学发现的“延迟认可”？

基于引文均值的方法[1]



An example from Li & Shi (2015)

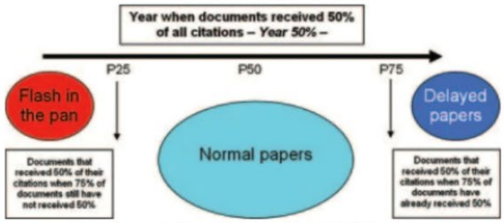
- 设置标准的“沉睡期”与“苏醒期”：
- “沉睡期”通常是3-5年，“苏醒期”则是至少4年或者更长时间
- 当一篇论文在“沉睡期”内被引次数低于某个阈值（例如1或2次）且在“苏醒期”内被引次数高于某个阈值（例如5次），它就被视为一篇“延迟认可”的论文。

- 这种方法确实能够识别出一些“睡美人”，但是如何划分两段时期以及设置引文阈值无疑将对结果产生重要影响。
- 尤其是，这有可能遗漏一些本来应该入选的论文，原因在于不同论文的引文曲线各异，而对所有论文来说“沉睡期”与“苏醒期”也并不一定出现在发表之后的固定时间节点并且持续相同的时间长度。

1 Garfield, 1989; Glänzel, Schlemmer & Thijs, 2003; van Raan, 2004; Li, Shi & Zhao et al., 2014

基于百分位值的方法[1]

- 三个参数：
  - Year 50%，即论文达到一半引文数量所用的时间；
  - P25，即样本中最先达到一半引文数量的论文所用的时间；
  - P75，即样本中最后达到一半引文数量的论文所用的时间。
- 如果一篇论文的Year 50% > P75，它就被认为是延迟认可的；反之如果Year 50% < P25，则它被认为是“昙花一现”的。



This figure is from Costas et al. (2010)

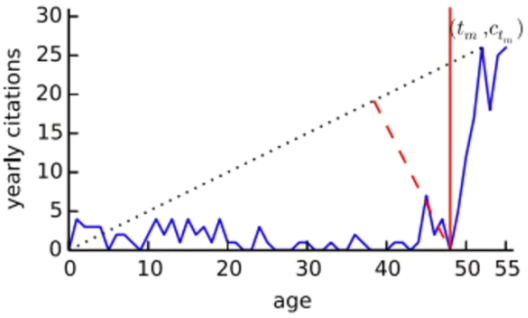
- 这种方法需要很大的计算量，并且标准过于宽松，以致于识别出很大比重的“延迟认可”论文（在其样本数据中约占20%）

1 Costas et al., 2010

“美丽系数”法[1]

- 在引文曲线上的原点与最高点之间连起一条参考线，两点分别对应论文发表年份（记为t_0）与达到最高引文的年份（记为t_m）。
- “美丽系数”=通过加和从t_0到t_m时段内参考线与引文曲线上每年数值的差值

$$B=\sum_{t=0}^{t_m}\frac{\frac{c_{t_m}-c_0}{t_m}\cdot t+c_0-c_t}{\max\{1,c_t\}}.$$



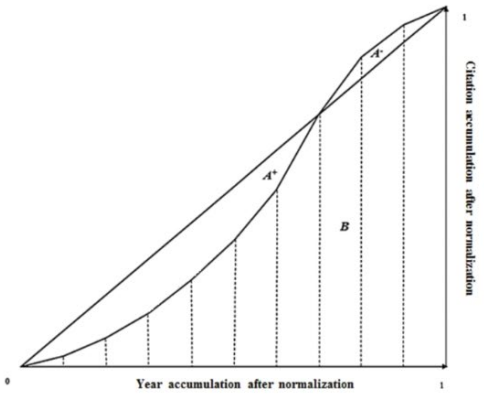
- Ke et al. (2015)的实验结果与Redner (2005)的研究十分吻合，用来识别“苏醒”年份的方法也具有启发意义。
- 另一方面，这一方法只考虑了引文顶峰之前的曲线部分而忽视了顶峰之后的那部分曲线。这可能导致对论文整体引文历史与最终认可程度的不完整的估计。

1 Li et al., 2014; Ke et al., 2015

This figure is from Ke et al. (2015)

“*Gs*指数”法[1]

- Li et al. (2014) 提出一个基尼系数的改版指标，称之为*Gs*指数，用来分析“睡美人”论文在“沉睡期”年度引文分布的不均衡程度。
- 一篇论文，如果*Gs*指数在(0.2, 0.6]之间，则有更大的可能性被“唤醒”与获得认可



Wang (2013)的“引用延迟”

- Wang (2013)基于单篇论文被引次数的累积百分比构建了一个称为“引用速度”的指标，并进一步定义了一个称为“引用延迟”的指标，以此衡量论文接收引文的快慢程度(Wang, Thijs & Glänzel, 2015)。
- Gs指数与引用延迟这两个指标并专门非针对量化“延迟认可”现象而被提出，但是它们考虑到整条引文曲线，适用于任何论文。

两个指标之间的相关性

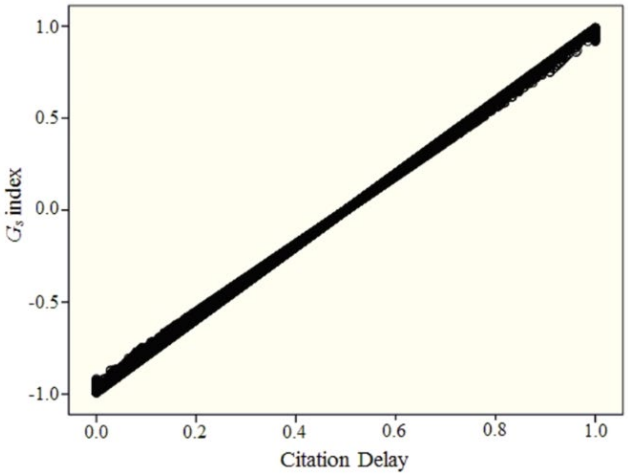


Figure The G_s index vs. the *Citation Delay*

两个指标之间的相关性

- Li et al. (2014) 通过加和引文累积曲线中的n个直角梯形面积来计算*Gs* 指数。其原始公式为：
- $G_S = \frac{A}{A+B}$, (1)
- 其中A与B分别是年度引文累积图中被引文累积曲线划分开的两个部分。*Gs* 指数的最终表达式是一个分段函数：
- $G_S = \begin{cases} 1 - \frac{2\times[n\times c_1+(n-1)\times c_2+\cdots+c_n]-C}{C\times n}, & C > 0 \\ 1, & C = 0 \end{cases}$ (2)
- 其中*n*是论文发表后的时间（年份），*C* 是*n*年内获得的引文总数， $c_i(i \in \{1,2,\cdots,n\})$是在第*i*年被引用的次数。

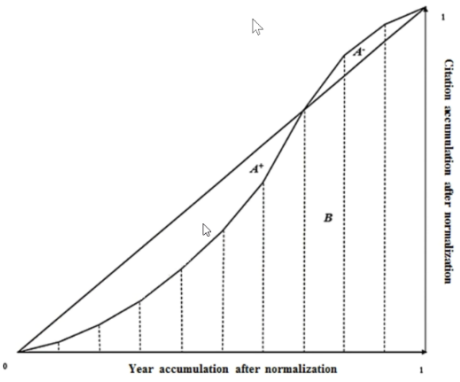


图 单篇论文标准化之后的引文累积曲线（改编自文献**Li et al., 2014**中的图**2**）

两个指标之间的相关性

- 在构造“引用延迟”方面， Wang (2013) 先提出了一个指标“引用速度”：
- 引用速度 $= \frac{\sum_1^{n-1} c_i/c_n}{n-1}$ (3)
- 其中n与方程2中相同，C_i 是i年之内的累积引文数量。Wang et al. (2015) 随后将“引用延迟”定义为 1 – “引用速度”。如果将“引用延迟”的公式展开，它可以表示为：
- 引用延迟 $= 1 -$ 引用速度 $= 1 - \frac{\sum_1^{n-1} c_i/c_n}{n-1} = 1 - \frac{C_1+C_2+\cdots+C_{n-1}}{(n-1)\times C_n}$
- $= 1 - \frac{(n-1)\times c_1+(n-2)\times c_2+\cdots+c_{n-1}}{(n-1)\times C_n}$ (4)

- $G_S = \begin{cases} 1 - \frac{2\times[n\times c_1+(n-1)\times c_2+\cdots+c_n]-C}{C\times n}, & C > 0 \\ 1, & C = 0 \end{cases}$ (2)
- 引用延迟 $= 1 - \frac{(n-1)\times c_1+(n-2)\times c_2+\cdots+c_{n-1}}{(n-1)\times C_n}$ (4)

两个指标之间的相关性

- 方程2与方程4在某种程度上有些相似。注意，Wang et al. (2015) 在计算“引用延迟”时将最后一年的引文数量排除在外了。如果我们将其添加到方程4中，就能发现两个方程中都有一个公共因子：
- 公共因子 $= -\frac{n\times c_1+(n-1)\times c_2+\cdots+c_n}{n\times C_n} = -\frac{\sum_1^n (n+1-i)\times c_i}{n\times C_n} = -\sum_1^n \frac{(n+1-i)}{n\times C_n} \times c_i$ (5)
- 容易推导出-1 < 公共因子< 0，原因在于：
- $0 <$ −公共因子 $= \frac{n\times c_1+(n-1)\times c_2+\cdots+c_n}{n\times C_n} < \frac{n\times c_1+n\times c_2+\cdots+n\times c_n}{n\times C_n} = \frac{n\times\sum_1^n c_i}{n\times C_n} = 1$ (6)
- 正是这个公共因子，既是*Gs* 指数与“引用延迟”的共同内核，同时也决定了这两个指标数值之间的正相关关系。换言之，这个公共因子在测量单篇论文的延迟引用上扮演着关键而同等的作用。如果说两组研究者使用不同的方法分别提出了两个不同的指标，而这两个指标具有相同的内核，那么这个相同内核背后的逻辑是值得我们更深入的研究的。

背后的逻辑：时域加权求和

- 从方程2我们可以得到：

$$G_s = 1 + 2 \times \text{公共因子} + \frac{1}{n} \quad (7)$$

- 在最后一年的引文信息被加入到“引用速度”的公式后，有引用速度 $= \frac{\sum_1^n C_i/C_n}{n}$。因此：

$$\text{引用延迟} = 1 - \text{引用速度} = 1 - \frac{\sum_1^n C_i/C_n}{n} = 1 + \text{公共因子} \quad (8)$$

$2 \times$ 方程(8) $-$ 方程(7)，我们有

$$\text{引用延迟} = \frac{1 + G_s}{2} - \frac{1}{2n} \quad (9)$$

- 因此，Gs 指数与“引用延迟”之间可以通过公共因子相互转化。为了更好了解公共因子发挥的作用，我们继而检查方程5中的分子与分母。

背后的逻辑：时域加权求和

- 一方面，方程5的分子可以被视为一篇论文每年被引次数的加权总和，其中权重为($-n + i$ -1) ($i \in \{1,2,\cdots,n\}$)，n 为论文发表之后经历的时间（年份）。方程的分母则起到规范化因子的作用，使得公共因子的取值在-1至0之间。而公共因子在整体上也是每年引文的加权总和，此时权重变为$-\frac{(n+1-i)}{n \times C_n}(i \in \{1,2,\cdots,n\})$。
- 现在两个指标的公共因子背后的逻辑十分清楚了。在考虑对一篇论文的年度引文数量进行加权求和时，对权重的选择取决于我们所要关注的那部分年度引文。年度引文需要强调得越多，赋予它的相应权重就应越大。
- 例如，方程5中的权重为：$-\frac{(n+1-i)}{n \times C_n} = \frac{i-n-1}{n \times C_n}(i \in \{1,2,\cdots,n\})$，与时间呈单调递增的关系，原因在于($i$-$n$-1) 随着 i的增加而增加。由于需要为靠后的年度引文赋予比靠前的年度引文更大的权重，以此凸显引文在时间上的延迟，这种引文求和的方法与“延迟认可”现象不谋而合。*Gs* 指数与“引用延迟”两个指标的合理性在此得到了证实。

对“延迟认可”的新认识

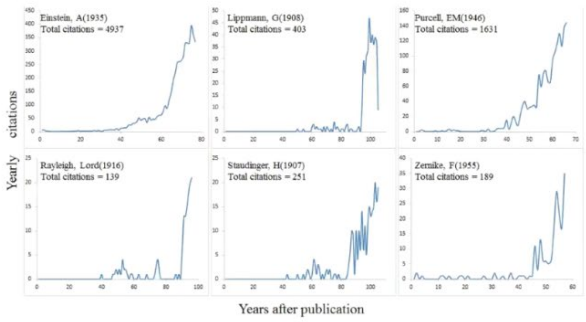
对年度引文的不均等加权

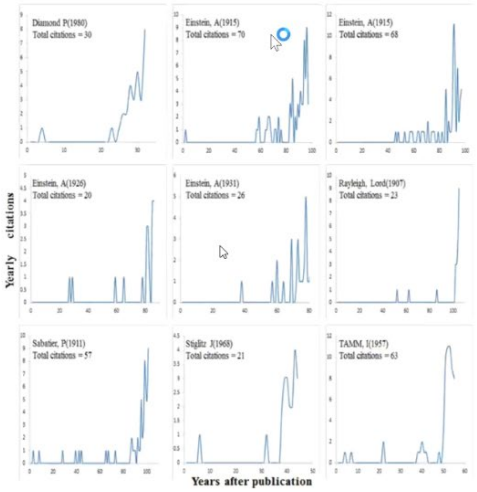
- 由于*Gs*指数与“引用延迟”都有一个公共因子，而这个公共因子在测度延迟引用方面起决定作用，所以这两个指标本质上是同一个指标，并且可以通过线性转换互相转化。其隐含的逻辑在于根据不同的测量目的为年度引文赋予不均等的权重，然后把它们相加。
- 这里的关键因素是设置每年引文数量的权重。
- 我们可以利用加权求和的方法开发新的指标来测度论文的引文在多大程度上被延迟了。如果设置权重为$\frac{i}{n\times C_n}(i\in\{1,2,\cdots,n\})$（随时间而增加），我们可以得到一个统一了*Gs*指数与“引用延迟”的测度延迟引用的指标D_1：

$$D_1=\frac{1}{n\times C_n}\times c_1+\frac{2}{n\times C_n}\times c_2+\cdots+\frac{n}{n\times C_n}\times c_n=\sum_1^n\frac{i}{n\times C_n}c_i$$
$$=\frac{\sum_1^n i\times c_i}{n\times C_n} \qquad (10)$$

- 其中c_i是年份*i*时的引文数量，C_i时截止到年份*i*的累积引文数量。在方程10中，$n\times C_n$同样是使得$0<D_1<1$成立的规范化因子。

利用D_1与引用总数识别“延迟论文”





引用延迟与引文总数衡量的是一篇论文的不同维度；同时，尽管D_1能够很好测度一篇论文的引文在多大程度上被延迟了，但是只有D_1并不能有效地识别出“延迟引用”，更不用说识别“睡美人”的存在了。因此，我们认为，在考虑“延迟认可”现象时，引用延迟与引文总数这两个方面缺一不可。

结论

- 我们建议从引用延迟的程度与整体认可的程度两个方面分开测量。
- 如果同时具有较高的D_a值与引用总数，则常常是“睡美人”论文。
- D_a值较高而引文总数中等或者偏低的论文同样值得关注。尽管它们目前没有获得很多即时认可，但是它们的学术价值有可能在经历长时间的“沉睡期”之后被发现，甚至有成为“睡美人”的潜力。
- 我们建议将D_a作为单篇论文的基本属性，以此来量化引文在多大程度上被延迟了，其优势在于：（1）它充分利用了一篇论文的所有引文数据，没有重要的引文信息被遗漏；（2）它避免将引文曲线武断地划分为不同的时段；（3）它适用于具有任何引文曲线的论文；（4）它简单易算；以及（5）它的取值范围在0与1之间，方便对引用延迟现象的后续研究。