

# 7 信息的测度2-指标设计

7 信息的测度2-指标设计

指标设计

信息的度量

度量信息量的原则

度量信息量的方法

指标

什么是指标

如何认识指标

好的指标

什么是好的指标

如何判断好的指标

判断睡美人指标

4.5 延迟认可

设计好的指标——把多个维度的信息巧妙糅合

1. 累加

2. 天际线

3. H指数

## 指标设计 Word count: 1316

### 信息的度量

观测对象在尚不能量化的时候设计一套指标来进行比较

#### 度量信息量的原则

能度量任何消息，并且与消息的种类无关

度量方法与消息的重要程度无关

消息中所含信息量与信息内容的不确定性有关

消息所表达的时间越不可能发生，信息量越大

某客机坠毁的信息量比今天下雨的信息量大

#### 度量信息量的方法

时间的不确定程度可以用其出现的概率来描述：消息出现的概率越小，则消息中包含的信息量就越大

设  $P(x)$  - 消息发生的概率

$I$  - 消息中所含的信息量

则  $I$  与  $P(x)$  成反比， $P(x)=1$  时， $I=0$ ； $P(x)=0$  时， $I \rightarrow \infty$

### 指标

#### 什么是指标

指标的意思是衡量目标的参数

预期中打算达到的指数、规格、标准，一般用数据表示

指标是对结果、目标的一般性描述：

年会时，今年新增了几个指标：员工流失率、门店客流量

高考成绩，从科学角度出发，语文+英语+数学成绩不能进行累加，累加的结果没有意义，这不见得是最优的方案，但是是最简单的方案，是大多数人朴素可以接受的方案

数据本身的价值有限，无法直接使用，有目的、有方法地对散乱的数据进行拆分、排序、关联后产生的数据指标才有更多的价值

- 首先要对数据指标有所理解
- 数据指标在一定程度上可以反应产品的销售市场、用户体验、发展方向等，从而改进产品设计

- 如同支出成本、用户活跃度、产品转化率等，这些指标往往能够体现出业务发展的关键要素，因此对产品规划十分重要

数据驱动是有道理的，只是数据很重要，**但思维优于数据**

如何认识指标 h2 ^

古德哈特定律：**当一个政策变成指标，它将不再是一个好的政策**

一项社会指标或经济指标，一旦成为一个用以指引宏观政策的既定目标，那么该指标就会丧失其原本具有的信息价值

政策制定者会**牺牲其他方面来强化这个指标**，从而使这个指标不再具有指示整体情况的作用

**如果一项指标一旦变成了政策制定的依据，它将不再是一个好指标**

好的指标 h1 ^

什么是好的指标 h2 ^

**好的指标是可比较的**，数据指标在不同的时间段、用户群体、竞争产品之间有不同的表现，可以更好地洞察产品的走向

简洁、无武断参数、可延展（可考虑未来）、可操作性强

**简单易懂的**

**往往是一个比率**：比率可操作性强，是天生的比较性指标

**会改变行为**：这是最重要的评判标准

- 指标的变化体现目标与现实的实际距离差
- 根据数据确定一条做与不做的准绳，对于规范企业行为大有裨益

**及时、准确、简明、直观**

如何判断好的指标 h2 ^

力求指标是及时的、准确的、明确的、简洁的、直观地将主力的异动汇报给我们

及时、准确是指标的生命线，不准确、不及时信号都会贻误战机，这是指标的内在要求

判断睡美人指标 h2 ^

**引文曲线**：一般而言，一篇学术论文在发表几年时间内会被其它论文引用，并逐渐达到被引次数高峰，之后被引次数慢慢减少，直到被人遗忘，**引文曲线**是被引次数反映在时间上的曲线

4.5 延迟认可 h2 ^

一篇重要的论文在发表很长时间后才被人们认识到它的价值

**基于引文均值的方法：**

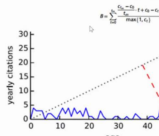
- 1. 设置标准的“沉睡期”「3-5」与“苏醒期”「至少4」
- 2. 当一篇论文在“沉睡期”内被引次数低于某个阈值且在“苏醒期”内被引次数高于某个阈值，就被视为“**延迟认可**”的论文
- 3. 但是如何划分时期和如何设定阈值是个大问题，还可能遗漏本应该入选的论文

**基于百分位值的方法**

- 1. Year 50%：论文达到一半引文数量所用的时间
- 2. P25：样本中最先的25%达到一半引文数量的论文所用的时间
- 3. P75，样本中最后的75%达到一半引文数量的论文所用的时间
- 4. Year 50% > P75, 延迟认可，< P25, 昙花一现
- 5. 不足：计算量很大，且标准过于宽松

**美丽系数：**

- 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



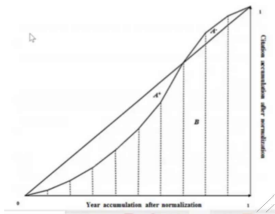
Ke et al. (2015)的实验结果与Redner (2005)的研究十分吻合。用来识别“苏醒”年份的方法也具有重要意义。

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略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。

- 这个指标最大的优势在于没有实参，但是可延展性有一定的缺陷，最大值的点可能会发生变化

**GS指数**

- 1. Gs指数与引用延迟（D1）这两个指标并专门非针对量化“延迟认可”现象而被提出，**它们考虑到整条引文曲线，适用于任何论文**，两个指标具有相关性
- 2. Gs指数，**基尼系数的改版指标**，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
- 3. 如果Gs指数在（0.2，0.6）之间，则有更大可能被“唤醒”并获得认可



美丽系数之前的指标都设定了阈值，这是一个相对主观的标准，参数设定太多，太复杂

设计好的指标——把多个维度的信息巧妙糅合 h2 ^

1. 累加 h3 ^

2. 天际线 h3 ^

考察科学家的学术能力：**有人/很多人说你行，说你行的人很行**

在散点图中有两个维度，如果科学家所在的点没有任何一个点在两个维度上**同时**超过他，那么就画一条线，最后连成的线上的科学家就是该领域中最顶尖的

在层内部没有区分度，不知道线上的人谁更好

3. H指数 h3 ^

将一个科学家的每篇论文的被引次数按降序排列，若H篇论文的被引量 $\leq H$ ，则指数为H