中国图书馆学报(双月刊)2000 年第 5 期
ZHONGGUO TUSHUGUANXUEBAO

●马费成 陈 锐

科学信息离散分布的机理分析

摘 要 科学信息单元按位次/频率排序法处理后符合布氏分布,也表现为半数轴上的 Logistic 函数,与科学信息的增长老化模型具有内在联系,证明情报学基本定律之间具有天然的一致性。参考文献 11。

关键词 科学信息 布氏分布 Logistic 模型 离散分布 机理研究
分类号 G350

ABSTRACT After sorting, scientific information units can be represented by Bradford Distribution and Logistic Function. This is intrinsically related to the model for the ageing of scientific information. 11 refs.

KEY WORDS Scientific information. Bradford Distribution. Logistic Model. Discrete distribution. Mechanism analysis.
CLASS NUMBER G350

1 两种不同的分布模型及形成机理

"信息离散分布"可以被认为是一条公理。但信息离散分布的规律是什么,机理又是什么,却是一个十分复杂的课题。揭示这一规律在理论上对情报学和信息管理学具有奠基性意义,在实践上对信息服务和信息管理工作具有指导作用。因此,对信息离散分布的研究吸引了众多的学者和图书情报专业实际工作者,但迄今尚未获得满意的结果。

一方面,信息离散分布这一课题难度较大;另一方面,不同类的信息具有不同的特征,有可能具有不同的分布规律,很难用统一的模型来加以描述。于是人们把注意力转向一些具有固定载体的典型特征的信息,例如科学信息。即令是科学信息,其范围也很广,分布也很复杂,而且还没有确切的表示方式,人们只得借助其载体——科学文献为研究对象来分析科学信息离散分布的规律,最具有代表性的便是布拉德福定律。

布氏定律用位次/频率排序法揭示科学期刊中文献的分散规律,并以区域分析和图像模拟描述这一分布规律。虽然它是通过经验统计得到的,而且还显得十分粗糙,但由于它简明而直观地描述了科学文献的离散分布状态,因而受到学者的关注。

学术界对于布氏定律的认识和研究经历了一个逐步深化的过程。

最初,人们这样简单地解释文献分散的布氏定律:设想在某一新学科中写第一批论文时,人们首先把它寄给最合适的杂志发表。这些杂志伴随着该学科的发展,刊载越来越多的这类文章,于是许多著者都希望把他的文章发表在本专业的、以前发表了大量高质量论文的这类杂志上,使投稿数量大大增加,杂志对文章选择性增大,因而对文章的质量要求提高,杂志威信也日益提高,便产生了一些"核心"性的杂志。这种现象就是科学文献分布中的"堆加"效应。与此同时,有关这一学科的文章也在其他杂志上发表,这就产生了科学文献的集中与分散现象。

在以后的研究中发现,除了科学文献呈布氏分布外,社会科学的许多领域中也呈现出类似的分布,如城市按人口多少的分布,居民按收入多少的群体分布,书籍按页数的分布,作者按其论著的分布等等……这些现象乍看起来毫无关系,但仔细观察就会发现它们并不是纯粹偶然因素的堆砌,不象许多自然现象那样受众多相互独立的、细微的、偶然的随机因素影响,而是受人的意志作用的一种有目的活动,具有十分明显的倾向性。人们写作时总爱选择常用的、传递功能强而消耗能量少的词汇。大城市总是人口集中的目标;杂志编辑部总希望选择质量较高的论文;科学工作者总是有目的地撰写论文,并把自己的论文

寄给声誉较高、影响较大的杂志。对于这些受人的意志作用,倾向性很强的现象,只要我们用频率/位次排序法对其观测值进行处理,都会呈现出同样的分布,它所揭示的是这些观测值(具体元素)在其主体来源中的集中与分散规律。费尔桑提出了一个统一的表达式来描述这类分布[1]:

$$P(X) = C/X^{p} \qquad (1 < p \leq 2) \qquad (1)$$

由人的控制因素支配的社会科学诸现象,包括情报现象,对观测值的概率密度分布常常服从上述分布式,我们称其为负幂分布。服从这种分布的现象,尽管其最初的表现形式不一定与(1)式完全相同,但总可以通过变换后得到与(1)等价的表达式。

对于社会科学和情报学中的许多现象,这种呈规律性的集中与分散是普遍存在的。于是有的研究者认为,在这方面存在着所谓"马太效应",即"……谁若有,就给他,并不断增加;而谁没有,则连已有的都要被夺走"。有的研究者则认为,导致这种集中与分散规律的是"成功产生成功"的机理。这里"成功"有较广的含义,诸如:论文的写作与发表,收入增加,杂志声誉的提高,词汇被选用等等。已取得的成功次数越多,就越容易在此基础上获得新的成功。例如高产作者撰写一篇论文十分容易,百万富翁增加一点收入毫无困难,声望高的杂志更容易获得高质量的稿件。这是个体自身能力和特性的显示[2]。当一系列同类对象被选择时,这种个体性的差异就常常成为选择的依据,有的经常被选择,有的不常被选择,这种频度不均的选择又可以反过来作为再次选择的依据。如果我们把对象受到一次选择视为一次成功,那么,这种成功的累积必然导致新的成功。而位次/频率法正是将这些个性突出、经常被选择的元素排在高位,而把那些不常被选择的元素排在末位,从而表达了这类特殊的分布。

与上面讨论的相反,某些现象受众多独立的、细微的因素影响,每一种因素都不起主导作用。例如在任意一段固定长的时间间隔内,由某块放射性物质放射出的 α一质点,到达某个计数器的质点数;从一个真空管的阴极发射出的电子到达阳极的电子数;来到某公共设施要求给予服务的顾客数(这里的公共设施诸如百货商店的售货员,工厂仓库的保管员,图书馆出纳员,机场的跑道,港口装卸货物的设备,电话交换台的干线等等);事故、错误、故障及其它灾害性事件数。

这些现象可以说是纯粹的随机现象,而且这些随机变量大致上都有如下特点:它们都取正整数为值,并且与时间间隔长度有关,当时间间隔极短,取值为2以上几乎是不可能的(例如,在极短的时间间隔内,可以认为不能有两个或两个以上的电话呼叫同时来到)。另外,他们取值的概率与时间间隔的长度有关,而与从哪个时刻算起没有什么关系,并且在不相重叠的时间间隔内,彼此没有什么影响。我们可以证明,在满足上述相应的条件下,这种与时间有关的随机现象服从泊松分布,我们可以称其为一个泊松过程或泊松流。

$$p(k\lambda) = \frac{\lambda^{k}}{k!}e^{-\lambda}$$

$$(k = 0, 1, 2 \cdots\cdots \lambda \text{为常数}) \qquad (2)$$

大量自然科学中的随机现象,或者只受偶然因素支配的一系列现象,都可以用泊松分布研究处理。与前面对应,我们称其为泊松分布系。

与泊松分布系相比,布—齐分布系无论从机理分析、适应范围还是数学表达都显得很不成熟。许多学者还从不同角度用定量化方法模拟这种成功累积效应,导出了负幂型函数,即广义布氏分布模型。最有代表性的便是西蒙引入的球—罐子模型。这里,主体来源被看作了罐子,某一具体元素被看作投入罐子中的球。对布氏分布、齐夫分布、洛特卡分布、帕累托分布来说,罐子好比科学期刊、词汇集合、科学工作者、居民集合等,球则好比相关文章、单词、所著论文、单位货币等。他设计有充分多的不同的罐子,在这些罐子中输入 R 个质量均匀的球,假定新进入罐子的球到达某一个罐子的概率与罐子内已有的球数成正比,求进入了 r 个球的罐子的个数 n(r,R),结果得到与(1)同型的负幂分布函数。经过变换,就可以得到布氏分布模型。

上述实验也可换一种方式进行。设想在一只袋中装着同样数量的红球和白球,每次从袋中随机取出一个球,如果是红球视为成功,将其放回袋中,再加入一个红球;如果是白球,视为失败,不再放回袋中。随着时间推移,袋中的红球将会越来越多,白球则越来越少。求某一时刻从袋中取出红球的概率,结果与球罐子模型的结果一致。

2　科学信息离散分布与 Logistic 模型的一致性

1988年,我们获得国家自然科学基金资助,进一步研究科学信息离散分布的规律,试图不仅在文献层次(宏观层次),而且在内容层次(微观层次)上揭示科

学信息离散分布的机理与模型。我们以布氏定律为参照系,选择电子学、物理学、生物学、工程技术等具有代表性的学科领域,利用 BIOSIS、INSPEC、COMPENDEX 光盘数据库输出记录,用计算机分析、统计记录中的信息单元的集中、聚类和分散状态,并分别用布氏区域分析和图像模拟对文献单元和内容单元的集中分散进行研究,发现科学信息在文献层次和内容层次上都表现出同样的分散态势,具有相同的分散规律[3~5]。计算机绘制出的曲线是一条生长曲线,通过 K- S 检验,发现与 3 参数 Logistic 函数拟合得很好[6,7],且比莱姆库勒函数精确得多[8]。因此在文献信息流规律中,除了增长、老化具有生长曲线描述的规律、满足 Logistic 函数外,我们又发现科学信息的分布也具有同样的规律。

Logistic 模型最早起源于生态学领域对于种群生物学的研究,即研究植物、动物与它们所处环境之间的相互关系。所谓种群(population)是指在特定时间内占据一定空间的同一物种的集合,一个最基本的定量单位就是所研究种群的个体数,而种群生物学主要研究种群的时间动态及调节机理。

一定空间内单个生物种群随时间变化的模型主要有 Malthus 模型、McKendric 模型、Logistic 模型和离散模型。通常假设种群数是时间 t 的函数 N(t),并认为它关于时间 t 是连续的并且充分光滑,它的导数 dN/dT 给出了这个种群增长的速率,Logistic 模型是其中最著名的一种[9,10]。

运用 Logistic 模型研究一定空间内单个生物种群随时间变化的规律,一般有以下几个假设和限制:①种群仅仅是时间 t 的函数 N(t),忽略了个体间的差异,如年龄、性别、大小等对种群增长的影响;②N(t)是连续且充分光滑的;③生育和死亡对任何生物个体来说都是随机发生的;④种群个体的平均增长率是种群大小的一个减函数 r(N),并且存在一个饱和水平 K > 0,使得 r(K) = 0;⑤生物体处于一种不随时间变化的定常的环境中,即环境变化不会对种群增长行为产生影响;⑥种群是在一定的空间内封闭的,即不存在迁移现象。

只有在完全满足上述几个条件时,单个生物种群随时间变化的规律才能体现出 Logistic 模型所揭示的生长曲线规律。

Logistic 模型的基本结论是,在一个有限资源环境内种群是不可能无限增长的,它总会存在一个饱和水平,当种群增长到接近于这个饱和水平时,其增长速度应该逐渐减慢而渐近于零。

Logistic 模型最初是用来研究生物种群增长规律的,在情报学领域,我们曾利用这一模型来描述科学文献的增长和老化,得到了符合实际的结论。但这一模型为什么能够较为精确地拟合科学信息离散分布规律呢?我们给出如下分析:

(1)科学信息的离散分布主要研究某一学科主题范围内科学信息单元的分布规律。如在取自 BIOSIS 数据库的第一组数据中,我们仅仅选择了分类号为 CC33508、分类名为 VIROLOGY - PLANT - HOST - VIRUSES 范围内的信息单元,学科分支或主题范围就相当于某一生物种群。

(2)科学信息在随时间推移的演化进程中有两个相互联系、相互影响、不可分割的趋势:即总量增长和离散分布。前者表现为科学信息在纵向上量的累积,后者表现为科学信息在横向上量的扩散,这两个趋势实质上都是科学信息的增长。如同科学文献的增长,也如同生物种群的增长一样,因而可以用 Logistic 模型来描述。不同的是科学信息离散分布的 Logistic 模型是在半对数坐标轴上取得的,时间变量隐含于按载文量递减排列的期刊序号中,与科学文献增长的变量之间刚好存在一个对数歪曲。

(3)我们在研究科学信息离散分布时,将科学信息单元的累积仅仅定义为期刊累积数 r 的函数 R(r),即科学信息累积量只随统计的期刊累积数量的变化而变化,而期刊累积数本身就是时间 t 的累积值且是函数中唯一的自变量,忽略了其他可能影响科学信息单元累积量变化的因素。

(4)R(r)是连续且充分光滑的,即假设 R(r)在任何一点均可微。

(5)假定科学文献的增长、发展是按正常规律进行,不考虑非正常情况下的离散分布如战争、重大自然灾害等,如在本研究中取自 BIOSIS 的第一组数据取值范围为 1995 年到 1997 年,其间全球范围内并未发生对生物学有重要影响的非科学事件。

(6)基于情报学中的文献增长与老化规律,我们确认科学信息的增长是分阶段的,在增长到一定程度时必然会进入一个相对平稳的状态。

科学信息总量增长遵循生长曲线规律早已被学术界所认识,突出体现在人们运用 Logistic 模型来描述科学文献的增长和老化规律,而科学信息的离散分布也在相当程度上遵循生长曲线的规律却未曾被认识。事实上,经典布拉德福分布曲线和莱姆库勒函数

也形似"S"形曲线,只不过人们在模拟这些曲线时,排除了"格鲁斯"下降部分,没有将其当作"S"曲线来处理。最有代表性的便是布鲁克斯方程和莱姆库勒函数,前者分别用两个函数表示核心区曲线和相继各区的直线部分,后者则用2参数函数模拟分布曲线。这两个函数的共同优点是模型简单、参数少、直观性强。布鲁克斯方程的直线部分过于简化粗糙,误差较大,莱姆库勒函数被认为是最精确的模型,但其对曲线中间段的拟合较差。我们用计算机绘制出科学信息(文献单元和内容单元)的布氏分布曲线,直接用Logistic函数模拟整个曲线,将布氏分布曲线的3个部分统一到一个模型中,而且较好地通过了误差检验,对理论研究和实际应用都具有重要意义[11]。西蒙的广义布氏分布模型将布氏定律、齐夫定律、洛特卡定律统一到一个函数中,科学信息离散分布符合生长曲线规律的现象进一步将布一齐分布、科学信息的增长与老化统一到Logistic模型中,说明情报学基本定律之间的天然一致性,同时也证明了科学信息离散分布规律在情报学中的奠基性意义。

科学信息离散分布规律与Logistic模型的一致性似乎并不偶然,因为种群生物学中生物的属按其种的分布本身就符合布拉德福一齐夫分布。

需要指出的是,有可能影响本项目结论的是数据来源。我们用以表征知识单元的主题词或关键词在这三个数据库中均是对应于每篇文献出现的,是一种静态的、表面的联系,并不能表达知识的内在逻辑关系,因而抽取的主题词或关键词数量基本上与文献量对应,其分布显示出相同的规律是必然的。唯一例外的是核心词的分布。如果说一般的主题词或关键词尚不能完全代表一个学科领域的知识单元或内容单元的话,那么核心词则无疑是一个学科领域最基本的概念,最能反映该学科领域的基本内容和实质。核心词的分布也表现出与文献单元同样的规律,不同的仅仅是其离散程度更大。今后的研究需要在此基础上进一步考虑知识单元之间的内在逻辑联系,可以用引文索引、关系索引等手段建立这种联系,从科学信息的生产和利用过程中去考察其离散分布,这可能是研究科学信息离散分布最好的途径和方法。

本研究利用生长曲线来描述科学信息在文献单元和知识单元层次上的离散分布规律尽管取得了一定成功,但却是在半对数轴上取得的,即以Logr为横坐标对曲线模拟的结果,如果将对数坐标转换成一般坐标,则函数的表现形式将会是另一种模型,其形式也并不简化,应用也不一定方便。我们之所以没有将对数形式的自变量转换为一般形式的变量,正是希望保持其Logistic函数形式,从理论上阐明科学信息离散分布的态势,同时保持其与科学信息增长趋势的一致性。

基金项目　本文系国家自然科学基金资助项目"科学信息离散分布的机理与模型研究"(批准号:79770067)的研究成果。

参考文献

1 Fairthorne, R. A. Empirical hyperbolic distributions (Bradford - Zipf - Mandelbrot) for bibliometric description and prediction. Journal of Documentation. 25(4). 1969
2 严怡民主编. 情报学概论. 武汉:武汉大学出版社,1983
3 马费成等. 科学信息离散分布规律的研究 —从文献单元到内容单元的实证分析(Ⅰ):总体研究框架. 情报学报,1999,18(1):79～84
4,11 马费成,陈锐. 科学信息离散分布规律的研究 —从文献单元到内容单元的实证分析(Ⅱ):文献离散分布的布氏区域分析. 情报学报,1999,18(2):171～182
5 马费成,陈锐. 科学信息离散分布规律的研究 —从文献单元到内容单元的实证分析(Ⅲ):文献单元离散分布的莱姆库勒函数拟合. 情报学报,1999,18(3):270～277
6 马费成,陈锐. 科学信息离散分布规律的研究 —从文献单元到内容单元的实证分析(Ⅳ):以布氏区域分布为参照系的知识单元分布. 情报学报,1999,18(4):376～383
7 马费成,陈锐. 科学信息离散分布规律的研究 —从文献单元到内容单元的实证分析(Ⅴ):知识单元离散分布的图形模拟. 情报学报,1999,18(5):463～479
8 马费成,陈锐. 科学信息离散分布规律的研究 —从文献单元到内容单元的实证分析(Ⅵ):比较与总结. 情报学报,2000,19(1):78～86
9 刘来福,曾文艺. 数学模型与数学建模. 北京:北京师范大学出版社,1997
10 安鸿志,陈敏. 非线性时间序列分析. 上海:上海科学技术出版社,1998

马费成　武汉大学传播与信息学院院长、教授、博士生导师。通讯地址:武汉大学。邮编430072。
陈　锐　武汉大学传播与信息学院博士研究生。

(来稿时间:2000-02-01)