中国图书馆学报(双月刊)2000 年第 3 期
ZHONGGUO TUSHUGUANXUEBAO

●李洪喜　于　光　荣毅虹

科技文献老化过程的数学辨识模型*

摘　要　分析研究了几种科技文献老化数学模型,讨论了引文年代分布数据统计中引文频次的采样误差。基于过程辨识理论,提出了科技文献老化的数学辨识模型,即传递函数模型。参考文献 12。图 5。

关键词　科技文献　老化过程　数学辨识模型
分类号　G250

ABSTRACT　In this paper, the authors analyze several mathematical models of the aging of scientific documents, discuss errors in the sampling of citation frequencies in the statistics of date distribution data, and propose a mathematical recognition model based on the process recognition model. 12 refs. 5 figs.

KEY WORDS　Scientific documents.　Aging.　Mathematical recognition model.
CLASS NUMBER　G250

1　科技文献老化研究现状评述

科技文献的老化过程是指文献利用率(用被引用频次表征)随时间变化的过程,是文献计量学研究的重要内容之一,其研究始于 1943 年戈斯纳尔(Gosnell)的博士论文《大学图书馆中文献老化问题》。50 多年来,文献计量学家及文献工作者在科技文献老化问题上做了大量的研究工作,如戈斯纳尔、贝尔纳(J. D. Bernal)、巴尔顿(R. E. Berton)和凯普勒(R. W. Keblen)、布鲁克(Brookes)、普勒斯(D. Price)、莫蒂列夫、刘文及朱西传等人都提出了不同的科技文献老化模型,推动了科技文献老化规律的研究。并出现了几种主要的老化模型。

1.1　引文频次的负指数模型及其改进模型

1958 年,英国物理学家贝尔纳借用放射性物质衰变的过程来类比文献老化现象,提出了描述引文频次随引文年龄增长的衰减过程的负指数模型[1],即:

$$r(t) = Ke^{-at} \qquad ①$$

式中 $r(t)$ 为 t 年所发表的文献的引用频次,k 为常数,a 为老化率。

引文频次的负指数模型形式简洁,基本上反映了引文频次随时间的变化规律,但由于引文频次的统计误差会造成建模误差,后又改用累积引文频次增长模型。1970 年,布鲁克在引文频次的负指数模型基础上提出了文献老化的累积指数模型:

$$R_B(t) = K_0 b^t \qquad ②$$

式中 $R_B(t)$ 为引文中 t 年以前(包括 t 年)发表的文章数(被引文献年龄 $\geq t$),k 为常数,b 为老化系数,$b < 1$。

1991 年刘文提出了一种科技文献老化规律的数学模型[2],即:

$$R(t) = K_0(1 - e^{-at}) \qquad ③$$

式中 t 为年龄,a 为老化常数,$R(t)$ 为文献在 t 年内累积被引频次(注:引文年龄 $\leq t$ 的引文总数)。

实质上,以上 3 种老化模型在数学上是等价的,首先由模型①能推导出布鲁克模型,由 $r(t)$ 和 $R_B(t)$ 的定义得:

* 本文系黑龙江省自然科学基金资助项目(G9703)研究成果。

$$R_B(t) = \int_t^{\infty} (t)\,dt = \int_t^{\infty} Ke^{-at}dt = \frac{k}{a}e^{-at}$$

即得:

$$R_B(t) = \frac{K}{a}b^t = K_0 b^t \qquad ④$$

其中 $b = e^{-a} < 1$,显然式①和式②是等价的。

根据刘文公式的定义,有:

$$R(t) = \int_0^t (t)\,dt$$

$$= \int_0^t Ke^{-at}dt = \frac{K}{a}(1 - e^{-at})$$

即:

$$R(t) = K_0(1 - e^{-at}) \qquad ⑤$$

可见式①与式③也是等价的。

在实际应用中布鲁克模型总体上与实际统计数据吻合较好,但当 t 较小时,即在 0～5 年范围内其拟合效果不佳。基于此原因,文献[3]在布鲁克模型基础上引入了文献传播阻碍因素的影响,即考虑了文献从发表到被引用需要一个传递和选择过程,建立了引文年代分布数学模型:

$$R(t) = R_0\left[1 - \frac{\beta}{\beta - \lambda}e^{\lambda t} + \frac{\lambda}{\beta - \lambda}e^{\beta t}\right] \qquad ⑥$$

式中 λ 和 β 分别为文献老化系数和引文阻碍系数,R(t) 为累积引文量(文献年龄 ≤t 年的引文量),R_0 为累积引文总量。从文献[4]的验证过程看,此模型优于布鲁克模型。

1.2　巴尔顿——凯普勒方程及其改进形式

1960 年巴尔顿和凯普勒提出了一个关于累计引文频次的老化方程[5]:

$$Y = 1 - \left(\frac{a}{e^t} - \frac{b}{e^{2t}}\right) \qquad ⑦$$

式中 a + b = 1,Y 为文献在年内累积被引次数与总引文量之比,t 为以 10 年为单位的时间,该模型在初始段和统计数据也吻合不好。

计算表明巴尔顿——凯普勒老化方程与实际统计数据之间存在着显著差别,特别是在初始阶段。1980 年莫蒂列夫对此又提出了修正公式:

$$Y = 1 - \left(\frac{a}{e^{t-0.1}} + \frac{b}{e^{2t-0.2}}\right) \qquad ⑧$$

式中 a,b,t,y 的意义同式⑦。

莫蒂列夫修正式揭示了文献引用过程中存在的延时现象,但是它只把延时统一定为 0.1(即 1 年),文献引用中的延时现象是由于文献发表过程延时引起的,由文献发表过程的规律得知此延时随文献载体的不同而变化[6],所以方程⑧有一定的局限性。

1992 年北京大学的丁学东提出了巴尔顿——凯普勒方程的级数修正式[7]:

$$\begin{cases} Y = 1 - \sum_{i=1}^{n} a_i e^{-it} \\ \sum_{i=1}^{n} ai = 1 \end{cases} \qquad ⑨$$

$(i = 1,2,3,\cdots\cdots,n)$

式中 Y,t 意义同式⑦,当 $i = 2$ 时式⑨与式⑦相同,当 $i > 2$ 时,式⑨拟合效果比巴尔顿凯普勒好一些,然而从数学分析角度而言,任一函数都可以用某一级数形式来展开,但这样就使得模型具有过多的待定参数而复杂化,这不符合建模的"吝惜原则"。

1.3　科技文献老化的延时修正模型

文献[8]分析了文献发表延时对引文统计结果的影响,提出了在各种老化模型中进行延时修正的观点,即在模型中加入延时修正项 τ,并验证了老化模型加入发表延时修正的必要性和有效性,然而该模型是在原有的老化模型基础上实验的,未能将老化模型归于一种形式,应用时较为复杂。

2　引文频次的采样误差分析

文献老化模型到目前为止仍以经验模型为主,因而引文年代分布的数据统计是建模所必不可少的环节,建立文献老化模型的目的是描述引文频次随引文年龄(时间 t)的变化过

程[设引文频率的时间函数为 r(t i)],但某一时刻的引文频次实际上是无法统计得到的,能统计得到的量是用某一时段[t,t + △t]内的引文数表示的平均引文频率$\overline{r(t\,i)}$。

$$\overline{r(ti)} = \frac{1}{\triangle t}\int_{t}^{\triangle t} r(t)\,dt \qquad ⑩$$

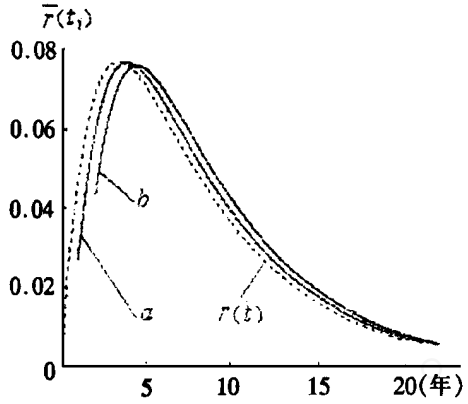

图 1　引文频次数据的采样误差

a. △*t* = 1 年　　*b*. △*t* = 2 年

由于 *r(ti)* 是随时间变化的(如图 1),显然$\overline{r(t\,i)}$将不同于 r(t i)且随统计时段大小的不同而变化;如果把实际统计出的平均引文频率数据当作理论引文频率,则将引入图 1 所示的采样统计误差,将影响建模精度。另外,建模结果必然与统计间隔有关,这是极不合理的,因为统计间隔是主观因素决定的,而模型应反映老化过程的客观规律,决不应受主观因素的影响,这是引文频次模型的不便之处。正是由于这个原因,人们才转向累积引文频次模型的研究。累积引文频次通常定义为区间[0,t]内的引文数(布鲁克模型中特别定义为 t 及 t 年前的引文数,区间应为[t,∞)),依定义有:

$$R_B(t) = \int_{t}^{\infty} r(t)\,dt \qquad ⑪$$

$$R(t) = \int_{0}^{t} r(t)\,dt \qquad ⑫$$

显然,R(t)是可直接统计的,统计间隔的选取仅影响统计数据的样本数,而不影响统计的数值,消除了采样误差。

3　科技文献老化过程的数学辨识模型

3.1　模型选择

本文采用辨识方法建立累积引文频次老化模型。模型类型的选择是辨识建模的首要问题,选用的模型应具有一定的覆盖能力,能比较灵活地描述动态过程,另外,模型的选择要本着吝惜原则,尽量选用简单的模型,以降低算法的复杂性。

任何一物理系统的动态过程,都有一个输入量和输出量,输入量和输出量能用一定的关系式来表示。根据科技文献引用过程这一物理现象,可把输入量看成:总的引文频次/总引文量 = I(为阶跃函数),输出量为累积引文频次随时间的分布与总引文频次之比。

根据老化过程的研究及统计数据,老化过程一般如图 2、图 3 所示。

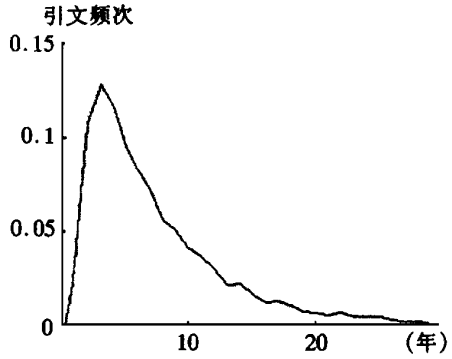

图 2　生物化学文献的引文频次曲线

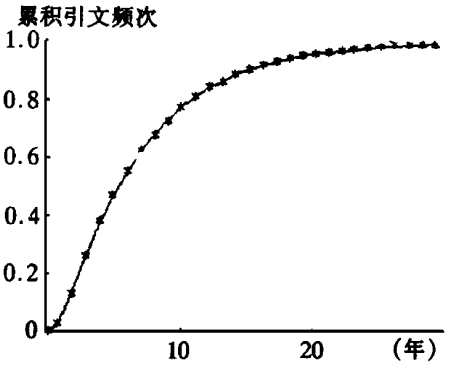

图 3　生物化学文献累积引文频次分布曲线

图2 中引文频次随时间的变化曲线是老化过程的脉冲响应，输入是强度为总引文次数的理想脉冲信号。图 3 中累积引文增长曲线是老化过程的阶跃响应，输入是幅值为总引文次数的阶跃信号。根据阶跃响应和脉冲响应的形状，初步选择如下形式的模型结构：

$$W(s) = \frac{Ke^{-\tau s}}{(T_1 s + 1)(T_2 s + 1)} \qquad ⑬$$

或

$$W(s) = \frac{(T_4 s + 1)Ke^{-\tau s}}{(T_1 s + 1)(T_2 s + 1)(T_3 s + 1)} \qquad ⑬$$

其中 K 为放大系数，T_1、T_2、T_3、T_4 为惯性时间常数，τ 为纯延时项。这里分母阶数比分子阶数高 2 阶是考虑到脉冲响应的 0 初始值特点而选取的。

3.2　模型验证

现引用两组数据来验证模型与实际数据之间的拟合程度。

辨识算法采用了时域响应的最小二乘拟合法[9]，待辨识过程取自文献[10]中 1980 年统计的生物化学文献累积引文频次数据(图 3)及《科学引文索引》(SCI)中 1992 年统计的累积引文频次数据[11](图 4)。

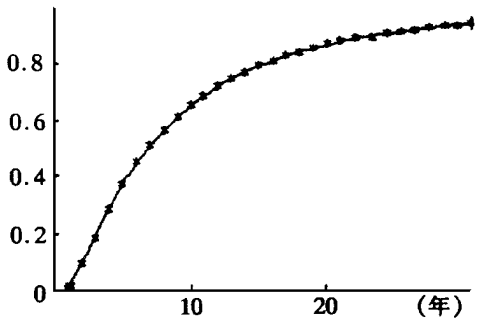

图 4 SCI1964～1992 年累积引文频次分布曲线

对于文献[12]中的 1980 年《生物化学杂志》和《生物化学》发表的 2595 篇论文所附的被引文献的统计数据，辨识得到的模型为式⑫所示的模型，图 3 为统计数据与模型理论计算绘制的曲线，可以看出辨识效果很好。其中各参数分别为 $T_1 = 5.921$，$T_2 = 0.6381$，$\tau = 0.5496$，最大误差为 0.5 %。

对 1964～1992 年《科学引文索引》累积引文分布数据，辨识得到的模型如式⑬，图 4 为 SCI 统计数据与理论模型计算结果对比曲线。其中参数分别为 $T_1 = 12.148$，$T_2 = 4.4488$，$T_3 = 0.97339$，$T_4 = 7.759$，$\tau = 0.52159$，最大误差为 0.29 %，拟合误差已达到很小的程度。进一步的模型检验表明，拟合残差完全满足零均值白噪声的要求，说明辨识结果是严格满足辨识建模要求的。以图 4 的辨识结果为例，给出残差的自相关函数图 5，结果非常接近于脉冲函数。

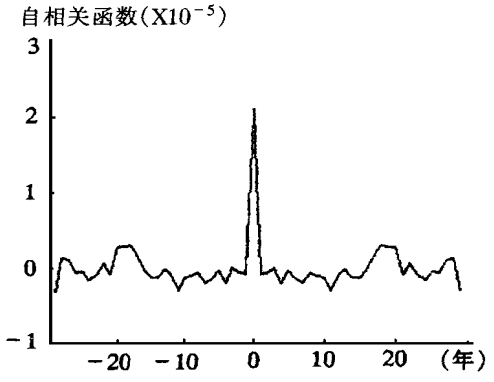

图 5　SCI 引文数据拟合残差的自相关函数

4　结论

通过对以往科技文献老化模型的讨论，指出了在建立模型时引文频次存在着采样误差。本文采用辨识方法，建立了科技文献老化过程的累积引文频次数学辨识模型，并首次引入了纯延时项 τ，减小了初始段拟合误差。通过对典型科技文献老化过程进行的模拟，得到了非常理想的拟合结果。

参考文献

1，10，12　邱均平. 文献计量学. 北京：科学技术文献出版社，1986

(下转第 88 页)

目前我国部分图书馆已经开始使用由全国情报文献标准化委员会第五分会等单位编写的《通用汉语著者号码表》编制中文图书著者号码,如果"杰克·伦敦"和"伦敦"不能选取统一标目,这不但会导致排检混乱,影响书目检索查全率,而且还会因为对同类书进行区分的著者号无法一致,而使文献排架也会相应产生混乱现象。

3　不断完善规范控制标准,早日推广使用规范数据

近年来,我国图书馆编目事业发展很快,特别是 1996 年文化部正式出台我国文化行业标准《中国机读目录格式》后,各级各类图书馆利用计算机编制机读书目数据不断向规范化、标准化靠拢。由于实现了书目数据交换和联机编目,许多馆之间书目资源共享力度加大,图书分编速度加快,书目质量控制得到一定程度保障。但也应看到,我国的规范数据建设还处于起步阶段。由于还未正式出台标目规范控制的国家标准,对个人责任者、团体名称等标目的选取方法还有待进一步完善。国家图书馆采编部自 1989 年以来开展了规范数据标准化的研究,1995 年国家图书馆成立"名称规范组",至今已积累和编写了 10 万余条名称规范记录,并早已灌制成光盘。他们的辛勤耕耘为我国的规范数据的发展奠定了良好基础。希望国内图书编目软件公司的科技人员尽快研制规范数据库与书目数据库的链接功能,使国家图书馆编制的名称规范记录早日运用到各种 CNMARC 的编目系统中,同时也希望国家图书馆名称规范组能在现有的基础上不断听取在各图书馆第一线从事编目工作人员的意见,使国家图书馆名称规范组几年来编写的《中国机读规范格式使用手册》、《中文图书名称规范数据款目著录规则》、《中文图书主题规范数据款目著录规则》日趋完善,早日得到国家有关部门的批准,在全国范围内得以推广使用。

参考文献

1　中国图书馆学会《西文文献著录条例》编辑组. 西文文献著录条例,1985
2　中国大百科全书·外国文学卷. 北京:中国大百科全书出版社,1982
3　国家图书馆图书采编部. 中文图书书名名称规范数据款目著录规则,1999

邵小鸥　中国社会科学院图书馆编目部主任,副研究馆员。通讯地址:北京建国门内大街 5 号。邮编 100732。

(来稿时间:1999-11-17。编发者:刘喜申)

(上接第 84 页)

2　刘文. 科学文献老化方程新探. 情报学刊,1991(5)
3～5　朱西传. 关于引文年代分布数学模型的探讨. 情报学报,1994(6)
6　于光等. 文献发表滞后过程的动态数学模型识别. 情报学报,1996(2)
7　丁学东. 文献计量学基础. 北京:北京大学出版社,1988
8　于光等. 文献引用延时效应及文献老化模型修正. 情报学报,1998(1)
9　方崇智等. 信息控制与系统. 北京:清华大学出版社,1988
11　SCI Guide and Lists of Source Publications. Science Citation index (SCI) 1992 Annual. ISI 1993,65～65

李洪喜　于　光　荣毅虹　哈尔滨工业大学图书馆工作。通讯地址:哈尔滨市南岗区司令街 17 号。邮编 150001。

(来稿时间:1999-12-02。编发者:翟凤岐)