

●于 光

二次文献发表延时过程的数学模拟及其分析^{*}

摘要 二次文献发表延时过程模型的建立,为研究利用二次文献查新的漏检问题提供了定量化工具。文献查新,不仅要检索二次文献数据库,还要检索学科范围内近期的核心期刊和相关期刊的一次文献。它也为二次文献检索数据库建立了量化研究方法。图2。参考文献1。

关键词 二次文献 出版延时 数学模拟

分类号 G253.2

ABSTRACT The mathematical modeling of index documents publishing process provides a quantitative tool for the study of hit rate of index documents. To check topics, we should search not only index databases, but also primary documents in core journals and related journals. It has also established a quantitative method for databases of index documents. 2 tabs. 1 ref.

KEY WORDS Index documents. Publishing process. Mathematical modeling.

CLASS NUMBER G253.2

二次文献是对分散的、无组织的已经出版的一次文献加工整理形成便于人们管理和使用的产品。

文献从产生到发表存在延时过程,文献的二次加工无疑又增加了一个环节,其延时乃是客观存在的。本文以著名的检索工具 *Engineering Index* (EI)、*Science Abstract* (SA) 为统计样本,分别统计出它们的发表延时数据,研究其内在机理,分析该过程与一次文献发表过程的相似性和相互差别,并利用文献[1]的研究理论和数学模型,对二次文献出版延时过程进行模拟,并对模拟结果进行分析。

1 二次文献发表延时过程的统计规律

本文所说的二次文献发表延时是指从被索引文献一次文献期刊出版日到二次文献检索期刊出版日之间的时差。我们首先以 EI、SA 为统计样本,统计数据为:一次文献发表时间(以期刊文献为样本,其他文献类型不作统计);二次文献出版时间。

数据处理过程如图1所示。

在二次文献出版延时数据统计中发现,该过程看起来杂乱无章,但当统计样本足够大时,即可发现该过程有着与一次文献发表延时过程同样可循的规律性,可以用数学的方式来描述,该过程的数学模拟是可行的。

2 二次文献发表延时过程机理分析及数学模型

2.1 二次文献的发表延时过程的机理

二次文献的发表一般要经过4个步骤:①根据二次文献检索工具的学科及文献类型的要求确定文献的来源出版物

(包括期刊、会议论文集、专利文献等)。②根据要求对一次文献进行整理、挑选,确定收录文献,可定义单位时间录用的文献为收录稿流量 X 。③对收录文献进行加工处理,包括标引、编辑、建立索引系统、建立数据库等,可定义编辑部内正在加工还未发表的文献量为存稿量 N 。④出版发行,包括排版印刷、CD-ROM 制作发行、网络版数据库的上网发行等,可定义单位时间内发表文献量为发表流量 Y 。如果把该过程看作一个连续的物理系统,可得到二次文献发表过程的物理模型,如图2。

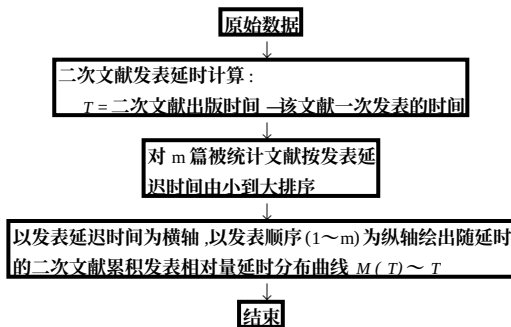


图1 数据处理程序框图

二次文献发表过程与一次文献发表过程很相似,但在出版运作上存在一些差别。

(1)文献发表策略上的差别。一次文献期刊的稿源来自作者投稿,稿源具有随机性,对于主编来说,稿源是不可控的,为了防止断稿,保证期刊的正常稳定出版,编辑部内要存有一定的具备发表条件的稿件。二次文献检索刊的稿

^{*} 本文系哈尔滨工业大学校基金资助项目(200081)研究成果。

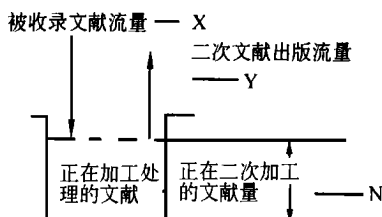


图 2 二次文献发表过程物理模型

源为收录的来源出版物上的文献,稿源是稳定的,只要文献具有发表条件均给予发表,不需要储备。

(2)造成延时的因素有差别。造成一次文献发表延时的主导因素有:评审、编辑论文的延时;编辑部中的存稿量导致的排队等候。造成二次文献发表延时的原因与被收录一次文献所需的加工工作量有很大关系,如一次文献传递方式、传递速度,语种、地域上的差别,这些均可导致一次文献加工成可发表的二次文献的时滞不同。

一般检索刊收录的文献在加工时间上的差别很大,有几种情况:在源期刊编辑部等待发表的一次文献,其二次发表可能在一次文献之前;已发表的一次文献不用加工处理即可发表;需标引和修改文摘;需翻译文摘等。二次文献检索刊的发表延时分布主要由文献加工时间的分散性形成,编辑部存稿主要是正在加工中的文献。

2.2 二次文献发表过程的数学描述

由二次文献发表过程的物理模型及其内在机理分析可得出,该过程与一次文献(期刊)发表过程具有相同的物理模型,可类比地建立收录稿流量、存稿量、发稿流量等变量之间的关系。我们利用文献[1]中建立的发表延时的数学模型——偏微分方程组来描述该过程:

$$\begin{cases} \frac{dN(t)}{dt} = X(t) - Y(t) \\ N(t) = \int_0^t n(T, t) dT \\ Y(t) = \int_0^t y(T, t) dT \\ \frac{\partial n(T, t)}{\partial T} + \frac{\partial n(T, t)}{\partial t} = -y(T, t) \\ n(T, t) = X(t) \\ y(T, t) = P(T) \cdot n(T, t) \end{cases} \quad (1)$$

以上方程组的初始条件:

$$\begin{cases} N(t) |_{t=0} = N_0, \\ n(T, t) |_{t=0} = n_0(t), \\ X(t) |_{t=0} = X, \\ Y(t) |_{t=0} = Y. \end{cases} \quad (2)$$

变量含义: T : 为二次文献发表延时(存稿年龄),

t_0 : 一次文献发表时间, 其中 $T = t - t_0$,

$X(t)$: t 时刻被收录的文献流量,

$N(t)$: t 时刻正在被编辑加工等待发表(可称存

稿)的二次文献数量,

$n(T, t)$: t 时刻存稿的延时分布,

$Y(t)$: t 时刻二次文献发表流量,

$y(T, t)$: t 时刻二次文献发表流量的年龄分布。

2.3 方程定解条件讨论

2.3.1 二次文献发表过程的稳态假设

二次文献检索期刊具有期刊的一切特点,即创刊一定时期后进入稳定状态,年或月收录的文献量和报道文献量(发表流量)在一定时期内保持稳定,即稳态过程。在模拟之前可对该过程做稳态假设,即录用稿流量、发表流量、存稿量、发稿概率等变量与时间 t 无关,录用稿流量与发表流量相等。该过程稳态假设条件下的数学模型为:

$$\begin{cases} X = Y \\ N = \int_0^T n(T) dT \\ Y = \int_0^T y(T) dT \\ \frac{dn(T)}{dT} = -y(T) \\ n(T) = X \\ y(T) = P(T) \cdot n(T) \end{cases} \quad (3)$$

2.3.2 二次文献发表的等概率特性

一次文献发表与二次文献发表在存稿特点和发稿策略上的差别,可用“零等待”的概念来描述。稿件经审稿、编辑加工后立即发表,没有等待发表时间,我们称之为发表过程中零等待概念。对任何期刊编辑部来讲,“零等待”都是追求的理想目标。由于一次文献期刊稿源的不可控性,一次文献发表过程达到全部零等待是不现实的。二次文献的发表均可认为是零等待发表,即收录的文献只要具备发表条件会立即发表。将该过程发稿概率函数近似为等概率形式,可得数学模型(3)的定解条件:

$$P(T) = P = \text{const} \quad (4)$$

这样二次文献发表过程的定解条件与文献[1]中的特解相同,可求解出稳态假设下的二次文献发表流量的延时分布方程:

$$y(T) = P \cdot Y \cdot e^{-P(T-T_0)} \quad (5)$$

同时由[1]中结果 $P = Y/N$,代入式(5)得:

$$y(T) = \frac{Y^2}{N} \cdot \frac{T-T_0}{e^{N/Y}}$$

设时间常数 $T_s = N/Y$,可化简上式:

$$y(T) = \frac{Y}{T_s} \cdot e^{-\frac{T-T_0}{T_s}} \quad (6)$$

式(6)左端的意义是发表的年龄为 T 的二次文献数量, T_s 是时间常数。

为了减少采样统计误差,提高模拟精度,本文做累积发表流量延时(年龄)分布的模拟研究。累积发表流量通常定义为延时区间 $[0, T]$ 内发表的总文献量,由式(6)积分得:

$$Y(T) = Y(1 - e^{-\frac{T}{T_s}}) \quad (7)$$

采用年做时间单位,方程(7)左端的意义是每年发表的延时(年龄)为 T 以下的文献总数, $T_s = N/Y$ 为时间常数,该参数从物理意义上反映了该延时过程的物理平均延时。方程(7)是二次文献发表过程累积流量延时(年龄)分布的稳态解。设累积发表流量的相对量: $M(T) = \int_0^T y(t) dt / Y$, 那么有

$$M(T) = 1 - e^{-\frac{T}{T_s}} \quad (8)$$

3 二次文献出版延时过程的数学模拟

本文对 EI 1997(11) 中 2290 篇文献、SA 2000(3) 中 8500 篇文献二次发表延时过程进行模拟。

EI 1997(11) 发表过程的模拟使用累积流量的相对量延时(年龄)分布模型,即式(8)。模拟结果图略。其模型参数: $T_s = 2.674$ (月); $\tau = 2.8891$ (月)。

SA 2000(3) 发表延时过程的模拟使用累积流量延时(年龄)分布模型,即式(7)。模拟结果图略。其模型参数: $Y = 8500$ 篇/月; $T_s = 1.98$ (月); $\tau = 0.34$ (月)。

SA 2000(3) 的时效性较好, $T_s = 1.98$ (月), 小于 EI 1997(11) 发表过程的 $T_s = 2.674$ (月) 值, SA 2000(3) 发表的速度比 EI 1997(11) 快。SA 发表过程的纯延时项 $\tau < 0$, 即有一少部分二次文献发表在一次文献发表之前, 说明了由于现代通讯技术和网络的发展, 带来了二次文献加工过程的缩短, 预示着未来二次文献传播的发展趋势。

这两个模拟结果都比较理想, 仅在过程初始端有较大的误差。这说明:

(1) 该数学模型能反映二次文献出版延时过程的实质和内在规律, 可以作为研究二次文献发表延时过程的一种定量化工具, 即从不同二次文献检索工具发表过程模拟结果定量比较其时效性, 分析影响出版过程的各种因素等。

(2) 该模拟过程证实了文献[1]中的研究成果和数学模型对二次文献发表延时过程是适用的, 证明了文献[1]中的研究理论具有实际意义。

4 二次文献发表延时模型的应用前景分析

科学研究的第一步是通过大量的文献查新检索, 来证

明所要研究的项目是否具有创新性。由于全球的期刊文献的发表本身就存在延时现象, 而且文献查新使用的一般为二次文献检索工具(数据库), 而二次文献出版存在着不容忽视延时现象, 无疑又增加了一个延时, 其结果就是查新中的“新”字不新, 即近期出版的一次文献的漏检, 导致的直接严重后果是全球科研项目的重复研究和研究经费的大量浪费。二次文献发表延时过程模型的建立为研究利用二次文献查新的漏检问题提供了定量化工具, 其研究结果对于查新检索效果评价具有实际意义。

利用二次文献数据库查新检索, 由发表延时造成的漏检率与检索技术无关。漏检带来一系列问题, 例如专利发明权认定时所做的文献查新, 由于延时导致的漏检, 可带来一种法律状态的不稳定; 科学研究立项至项目批准确认的时差将增加延时影响。文献查新检索应伴随科学研究项目申报、立项、确认、完成等过程的始终, 要求除检索二次文献数据库, 还要检索学科范围内近期的核心期刊和相关期刊的一次文献, 利用二次文献发表过程模型指导查新检索, 确定补充检索一次文献年限, 将查新漏检损失降至最小。

二次文献发表延时过程模型的建立, 为二次文献检索数据库建立了量化研究方法。该研究成果对文献查新检索具有指导意义, 也为建立检索期刊(数据库)质量评价指标提供了定量工具。由于二次文献发表延时同样会给文献传播、引用、老化带来直接影响, 其研究成果应成为文献计量学研究内容。进一步深入研究可建立检索数据库的定量评价指标, 这对文献研究者和文献使用者都有重要意义。

参考文献

- 1 Yu Guang, Yu Daren, Rong Yihong etc. The Mathematical Models of the Periodical Literature Publishing Process. Information Processing & Management, 2000(3)

于光 女, 副研究馆员。通讯地址: 黑龙江省哈尔滨工业大学图书馆信息服务中心。邮编 150006。

(来稿时间: 2001-06-27)

(上接第 74 页) 要内容的大教育体制对图书馆经济活动产生的影响是巨大和深远的, 知识经济、信息经济、网络经济的发展对图书馆经济活动的影响也至关重要, 它们共同构成了左右未来图书馆经济活动的主因素。

从现在众多的因素来看, 图书馆经济活动在将来大有增强的趋势。未来的图书馆经济活动主要体现在如何为网络经济、网络社会、信息社会的服务上。今后, 图书馆的经济活动将以先进的管理理念、先进的管理体制为先导, 以先进的设备、高素质的员

工为基础, 制造先进的产品, 创造响亮的品牌, 符合市场经济发展的规律, 符合图书馆服务的宗旨。

参考文献

- 1 周秉澄. 校园经济论. 北京: 人民教育出版社, 2000
- 2 马陵. 新太阳——教育经济化. 北京: 新华出版社, 2001

王东波 曲阜师范大学图书馆办公室主任, 副研究馆员。通讯地址: 山东曲阜师范大学图书馆。邮编 273165。

(来稿时间: 2001-06-20)