

图书馆网站 Web 页面增长规律研究 *

——基于美国三种类型图书馆网站的典型调查分析

邱均平 丁敬达

摘 要 以美国公共、专业和高校三种类型图书馆网站的代表为典型调查对象,利用 Alltheweb 搜索引擎检索它们每年新出现的网页数量及建立网页数据的时间序列,然后运用统计学中的时间序列分析方法和 SPSS 软件对数据进行统计分析。得出结论:在图书馆网站的初始发展阶段,网页增长经历三个时期、符合普赖斯科学文献指数定律、增长指数与平均发展速度正相关。图 1。表 3。参考文献 8。

关键词 图书馆网站 网页 增长规律 指数回归方程 发展速度

分类号 G250.7

ABSTRACT The authors use three representative libraries' websites of the United States as the typical study objects, make use of the Alltheweb to search the quantity of their annually updated web pages and establish a list of web pages' data by year, then apply the time serial analysis method and SPSS software to analyze the statistical data. Finally, they obtain the conclusions that the increase of web pages experienced three periods and are in accordance with D. Price's exponential law for scientific literatures, and there exists a positive correlation between the growth index and the average development rate. 1 fig. 3 tabs. 8 refs.

KEY WORDS Library. Web page. Growth law. Exponential regression equation. Development rate.

CLASS NUMBER G250.7

在知识经济时代,作为信息和知识“汇聚地”的图书馆如何在网络环境下提高对广大用户的信息服务质量,是当前图书情报界亟待解决的一项重要课题,而不断加强图书馆网站建设和管理便是有益于解决这一课题的重要途径和方法。因此,开展图书馆网站 Web 页面增长规律的研究无论在理论上还是在实践上都具有重要的意义。通过探索图书馆网站 Web 页面随时间变化的规律,理论上,一方面可以丰富网络信息计量学的研究内容;另一方面有助于科学学和科学史对整个科学的发展趋势和规律的研究;实践上,一方面可以大致地揭示图书馆网站发展的某些特点和规律,优化网站的建设和管理;另一方面还可以预测广大用户未来的信息服务需求趋势,从而使图书情报机构可以合理

地确定经费的分配、信息搜集的原则以及新技术的应用等。

1 数据收集

1.1 收集工具

目前搜索引擎已成为搜寻和采集互联网上海量信息的基本工具。由于 Alltheweb 搜索引擎收录网页量大、支持语种多^[1],与其他免费搜索引擎相比,检索链接网页的功能最强^[2],并且具有按时间间隔计量、检索网页数目等功能,因此,我们选择 Alltheweb 搜索引擎作为本次实证研究数据收集的工具。

* 本文为国家自然科学基金项目“网上学术信息的分布与变化规律研究及其应用”(编号:70673071)的研究成果之一。

1.2 收集样本

统计学中,统计调查方法分为普查、随机抽样调查、重点调查与典型调查等几种方式^{[3]19}。为研究图书馆网站 Web 页面随时间增长的一般规律,对世界范围内所有图书馆网站进行普查显然是不现实的。另外,由于世界各地的经济发展水平不同、信息化程度不同、图书馆网站建设早晚不同等因素,使得能够检索到的不同图书馆网站网页数量差异很大,如使用搜索引擎 Alltheweb 对中国国家图书馆网站只能检索到 2005 年以来的网页,即使对美国国会图书馆网站也只能检索到 1995 年以来的网页,所以,采取随机抽样的调查方式推断图书馆网站 Web 页面的总体增长规律也是不可取的。重点调查要求重点调查单位的某一主要标志的标志总量在总体标志总量中占有绝大比重,对研究图书馆网站网页增长规律来说,没有哪些图书馆网站存在显著的标志总量,因此,亦不能采取此类调查方式。

典型调查是根据对调查对象的初步了解,有意识地从其中挑选有代表性的单位进行研究的一种非全面调查^{[3]20}。它根据唯物辩证法共性寓于个性之中、无个性即无共性的原理,从研究个别到了解一般,是人们常用的一种认识事物

的科学方法。由于美国国会图书馆^[4]、美国国家农业图书馆^[5]、美国耶鲁大学图书馆^[6]能够分别代表公共、专业和高校三种类型图书馆,且均可检索自 1995 年以来的网页,因此,就以这三家图书馆为典型调查样本来统计和推断图书馆网站 Web 页面随时间发展变化的规律。

1.3 收集方法

利用搜索引擎 Alltheweb 对这三家图书馆网站的网页进行分年检索,结果见表 1。检索方法是:利用 Alltheweb 的高级检索界面,在 Search for 项目选择“all of the words”,在后面空白栏输入检索式“site: 站点名称”,在 Language 栏选择“Any Language”,在 File format 栏选择“Any format”(即包括 .pdf、.doc、.xls、.ppt、.txt 和 HTML 等各种文件格式),在 Date 栏输入网页检索的起止时间。以 1995 年美国国会图书馆网站网页数量检索为例:进入 <http://www.alltheweb.com> 后选择高级检索,在第一空白栏输入检索式“site: www. loc. gov”,在 Date 栏设定“after 1995. 01. 01 and before 1995. 12. 31”,并在 Language 栏选择“Any Language”,其他的限定条件用 Alltheweb 的默认值,即可检索到该年网页更新数量。

表 1 美国三家图书馆网站每年网页更新和累积数量

年份	美国国会图书馆 www. loc. gov		美国国家农业图书馆 www. nal. usda. gov		美国耶鲁大学图书馆 www. library. yale. edu	
	网页更新数量	网页累积数量	网页更新数量	网页累积数量	网页更新数量	网页累积数量
1995	14	14	235	235	2	2
1996	31	45	492	727	19	21
1997	70	115	80	807	52	73
1998	75	190	103	910	94	167
1999	963	1153	83	993	107	274
2000	809	1962	26	1019	484	758
2001	998	2960	209	1228	701	1459
2002	2290	5250	522	1750	467	1926
2003	4360	9610	513	2263	378	2304
2004	6220	15830	406	2669	624	2928
2005	7430	23260	267	2936	632	3560
2006	9450	32710	974	3910	889	4449
2007	8820	41530	591	4501	841	5290

注:检索时间为 2008 年 12 月 14 日 8:05—11:28。

2 数据处理

通过前面的叙述,我们知道表1中检索得到的三家图书馆网站每年更新的网页包含了大量的电子期刊、论文、图书、报告等各种类型的网上文献,它们与传统文献的区别只不过是这些网上文献有可能包含在一个或多个网页中,因此,传统信息计量学理论对图书馆网站网页增长规律的考察仍具有普遍指导意义。

2.1 数据加工

根据信息计量学理论^{[7]45},我们以网页累积数据为依据,进行图书馆网页的增长规律研究。

这是因为各年更新网页数量逐年相加得到的网页累积数量总是增加的,就有可能趋于某种固定的规律,而非累积数据,即每年更新的网页数量,容易受到各种复杂的社会因素的影响,一般来说总是波动的,很难确定其是否近似地趋于某种固定的规律,给定量分析带来困难。如以美国国家农业图书馆为例,1996年新出现的网页数量是492;2000年新出现的网页数量是26;而2006年新出现的网页数量是974,波动较大。

事实上,当年网页累计数量就是当年网站所拥有的网页数量。把每年的网站网页数量按时间顺序排列,即形成了统计学上的时间序列。这样,就可以利用统计学中的时间序列分析理论来研究图书馆网站网页随时间发展变化的规律。

表2 美国三家图书馆网页加工数据

年份	美国国会图书馆		美国国家农业图书馆		美国耶鲁大学图书馆	
	环比增减速度	定基增减速度(%)	环比增减速度(%)	定基增减速度(%)	环比增减速度(%)	定基增减速度(%)
1995	--	--	--	--	--	--
1996	221	221	209	209	950	950
1997	156	721	11	243	248	3550
1998	65	1257	13	287	129	8250
1999	507	8136	9	323	64	13600
2000	70	13914	3	334	177	37800
2001	51	21043	21	423	92	72850
2002	77	37400	43	645	32	96200
2003	83	68543	29	863	20	115100
2004	65	112971	18	1036	27	146300
2005	47	166043	10	1149	22	177900
2006	41	233543	33	1564	25	222350
2007	27	296543	15	1815	19	264400

如果把1995-2007年每年网站网页的数量记为 $a_0, a_1, a_2, \dots, a_{13}$,则 $a_0, a_1, a_2, \dots, a_{13}$ 称为相应时间上的发展水平;如果以1995年作为基期(也可以其他年份,本文中环比增减速度和定基增减速度以1995年作为基期)对各时间的发展水平进行比较,则 a_0 称为基期水平,其余各项称为报告期水平。这样,时间序列中的一些分析指标就可表示如下^[8]:

环比发展速度:报告期水平与前一期水平

之比,说明现象报告期水平较前一期水平发展的相对程度。

$$\text{环比发展速度} = \frac{a_i}{a_{i-1}} (i=1, 2, \dots, n)$$

环比增减速度:报告期水平较前一期增长量与前一期水平之比,说明现象报告期水平较前一期水平增长的相对程度。

$$\text{环比增减速度} = \frac{a_i - a_{i-1}}{a_{i-1}} = \frac{a_i}{a_{i-1}} - 1 = \text{环比}$$

发展速度 - 1, (i = 1, 2, \dots, n)

定基增减速度: 增长量与基期水平之比, 说明现象报告期水平较基期水平增长的相对程度。

$$\text{定基增减速度} = \frac{a_i - a_0}{a_0} = \frac{a_i}{a_0} - 1, (i = 1, 2, \dots, n)$$

平均发展速度: 是各期环比发展速度的序时平均数, 说明现象逐期发展的平均程度。

$$\text{平均发展速度} = \sqrt[n]{\frac{a_1}{a_0} \times \frac{a_2}{a_1} \times \dots \times \frac{a_n}{a_{n-1}}} = \sqrt[n]{\frac{a_n}{a_0}}, (i = 1, 2, \dots, n)$$

平均增减量: 逐期增减量的序时平均数, 说明现象在一段时期内平均每期的增减量, 其中 n 为逐期增减量的个数。

$$\text{平均增减量} = \frac{\sum_{i=1}^n (a_i - a_{i-1})}{n} = \frac{a_n - a_0}{n}, \text{其中 } n \text{ 为逐期增减量的个数。}$$

按上述指标加工得到的数据如表 2 所示。

2.2 数据分析

我们以年代为坐标横轴, 利用表 2 的数据, 绘制美国国会图书馆、美国国家农业图书馆和美国耶鲁大学图书馆网页数据散点图, 如图 1 所示。

表 2 和图 1 数据表明, 从 1995 年到 1999 年, 美国国会图书馆、美国国家农业图书馆和美国耶鲁大学图书馆网页数量环比增减速度变化起伏较大, 定基增减速度缓慢增长, 每年网页平均增减量分别为 228、152、54, 这一时期可以看作是三家图书馆网页数量处于较不稳定的缓慢增长时期; 从 2000 年到 2004 年, 三家图书馆网页数量环比增减速度仍有较大起伏, 但相对趋缓, 定基增减速度在快速增长, 每年网页平均增减量分别为 2773、330、434, 这一时期可以看作是三家图书馆网页数量处于不稳定的快速增长时期; 从 2005 年到 2007 年, 三家图书馆网页数量环比增减速度趋向稳定, 定基增减速度在快速稳定增长, 每年网页平均增减

量分别为 6090、522、577, 这一时期可以看作是三家图书馆网页数量处于相对稳定的快速增长时期。总体上, 三家图书馆网页数量随时间的增长呈指数曲线上升趋势, 但美国国会图书馆、美国耶鲁大学图书馆、美国国家农业图书馆网站的平均发展速度依次递减 ($\sqrt[13]{\frac{41530}{14}} > \sqrt[13]{\frac{5290}{2}} > \sqrt[13]{\frac{4501}{235}}$)。另外, 通过对美国纽约州立图书馆、美国国家医学图书馆、美国麻省理工学院图书馆和加拿大国家图书馆等图书馆网站网页数据的统计分析, 发现也存在同样的指数增长趋势。

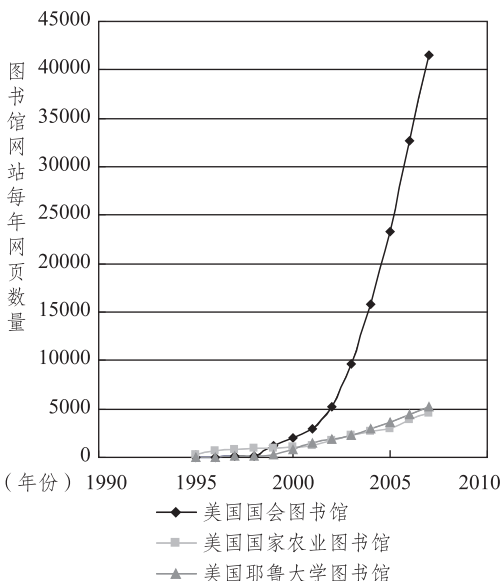


图 1 美国三家图书馆网站每年网页数量散点图

2.3 增长趋势拟合

根据前面的数据分析和图 1, 我们初步判断图书馆网站网页随时间增长的趋势大致为指数增长类型。下面我们利用 Spss13.0 for Windows 软件对表 1 中数据拟合分析出我们所判断的指数增长类型的曲线回归方程。选取 Spss 曲线回归中的 Exponential 模型对表 1 样本数据分别进行拟合, 结果见表 3。

表 3 美国三家图书馆网页增长趋势拟合及比较

名 称	回归方程	判定系数 R ²	修正判定 系数 R ²	方差分析 F 值 显著水平	拟合效果
美国国会图书馆	$\hat{y} = 17.529e^{0.663t}$	0.955	0.951	<0.001	很好
美国国家农业图书馆	$\hat{y} = 343e^{0.202t}$	0.925	0.918	<0.001	很好
美国耶鲁大学图书馆	$\hat{y} = 9.611e^{0.568t}$	0.856	0.843	<0.001	较好

其中, $\hat{y}$ 代表对应年的网站网页累积数量,t代表年(1995年为1,1996为2……2007年为13)。美国国会图书馆、美国耶鲁大学图书馆、美国国家农业图书馆网页增长指数依次递减(0.663>0.568>0.202),与平均发展速度($\sqrt[13]{\frac{41530}{14}} > \sqrt[13]{\frac{5290}{2}} > \sqrt[13]{\frac{4501}{235}}$)正相关,符合实际情况。

3 结论

(1)限于目前的网络环境和技术条件,本文对公共、专业和高校三种类型图书馆网站网页的研究时间仅限于1995年至2007年,但从图书馆网站未来发展的长远历程来看,我们认为这一时期和未来几年都是图书馆网站发展的初始阶段。

(2)根据共性寓于个性之中、个性体现共性的原理和前面的数据分析,我们推断在图书馆网站发展的初始阶段,图书馆网页的增长规律仍然符合普赖斯科学文献指数增长定律^{[7]47-48}: $F(t) = ae^{bt}$ ($a>0, b>0$)。其中,t为时间,以年为单位;a为条件常数;b为持续增长率。由于受社会多种因素的影响,不同网站的网页增长指数即b值有可能不同,但与网页的平均发展速度呈正相关。

(3)我们得出的图书馆网页指数增长规律与普赖斯后来提出的科学文献“指数型规律终将成为逻辑型”^{[7]51}并不矛盾,因为我们是在图书馆网站当前处在发展的初始阶段这个前提下得出的,从而验证普赖斯科学文献指数增长定律适用于图书馆网页。

(4)我们认为在图书馆网站发展的初始阶段,图书馆Web页面的增长经历了较不稳定的

缓慢增长、不稳定的快速增长和相对稳定的快速增长等三个时期,此外,用普赖斯科学文献指数增长定律预测未来5年图书馆网站网页的数量是可行的。因此,图书情报机构可以据此指导网站建设,例如预测新增网页数量、确定合理的网页建设经费、制定信息收集范围等。

(5)随着时间的推移、网络技术的进步、世界各地图书馆网站的不断发展,未来可从多语种的角度、在更长时间范围内对图书馆网站进行随机抽样和统计推断,对图书馆网页增长规律的指数模型进行不断的检验和修正。

参考文献:

[1] [2008-12-14]. <http://www.alltheweb.com/advanced?advanced=1&>.
[2] 袁毅. 核心网站评选的理论与方法[M]. 北京: 北京图书馆出版社,2005: 172-173.
[3] 谢启南,韩兆洲. 统计学原理[M]. 广州:暨南大学出版社,2005:19-20.
[4] [2008-12-14]. <http://www.loc.gov/index.html>.
[5] [2008-12-14]. <http://www.nal.usda.gov>.
[6] [2008-12-14]. <http://www.library.yale.edu>.
[7] 邱均平. 信息计量学[M]. 武汉:武汉大学出版社,2007:45-51.
[8] 袁卫. 统计学[M]. 北京:高等教育出版社,2004:185-202.

邱均平 武汉大学信息管理学院教授,博士生导师。通讯地址:武汉。邮编430072。

丁敬达 武汉大学信息管理学院2008级情报学博士生。通讯地址同上。

(收稿日期:2008-11-16;修回日期:2008-12-18)