基于布拉德福定律的Web被引频次分析

王晓芳 王 健 袁广林 陈 萍

(中国人民解放军陆军军官学院计算机教研室 合肥 230031)

摘 要 以布拉德福定律和学术论文被引频次作为研究对象,研究网络资源被引频次的分布规律,探讨信息计量法则在网络空间的适用性。采用区域分析法,对CNKI中数据库安全领域的论文进行被引频次分布规律研究,得到被引频次分布曲线,通过与布拉德福分布曲线和方程进行分析对比、曲线拟合、图像分析,得到网络空间中学术论文Web被引频次分布规律具有布拉德福定律特征的结论,得出布拉德福定律信息计量法则在互联网上具有适用性的结论。

关键词 网络信息计量,论文被引频次,布拉德福定律,曲线拟合

中图法分类号 TP391 **文献标识码** A

Web Citation Frequency Analysis Based on Bradford Law

WANG Xiao-fang WANG Jian YUAN Guang-lin CHEN Ping

(Department of Computer, Army Officer Academy, PLA, Hefei 230031, China)

Abstract This paper adopted Bradford law and academic papers citation frequency as researched object, researched about distributing rules of article's citation frequency, discussed applicability about information law in network space, got conclusion that Bradford law is applicable in networks. By adopting district analysis method on CNKI database within citation frequency about database security, distribution curve of citation frequency was got. Contrasting with Bradford curve, this paper accomplished curve imitatation and image analysis. This paper pointed out the conclusion that citation frequency on webometrics follows the Bradford law and Bradford law has certain applicability on network space.

Keywords Webometrics, Citation frequency of an article, Bradford law, Curve emulation

1 引言

信息计量学是研究各类信息数量特征的一门学科,具有广泛的应用价值,其中最重要的应用之一是研究学术传播的模式,例如学术成果的定量评价、学术传播的结构分析等等。随着网络技术的发展,学术传播系统从以学术期刊为中心的模式向更适用于网络环境的新模式发展,原有的信息计量指标是否继续适用于测度和评价网络信息资源,如何运用计算机技术、数学模型等方法,研究网络信息的内在规律,成为研究热点,也给我们带来了新的研究课题——网络信息计量学。这是一门新兴学科,即通过对网络上信息的组织、相互引证等进行定量描述和分析,研究揭示网络空间信息数量关系特征和规律,从而有效地利用网络信息资源。

本文所研究的问题是信息计量法则在互联网上是否继续适用,这属于网络信息计量学科下的子课题。鉴于网络计量学是信息计量学在网络环境中的应用,此研究对于核心网站的确定、网络结构挖掘和知识发现、机构研究能力评价、搜索引擎、数据处理、信息分析、科学决策、文献信息检索、海量数据库等广泛领域有重要而直接的应用价值。

被引频次是用学术论文发表以后被引用的次数来评价此学术论文的网络信息计量指标。被引频次评价是国际上公认的成果评价体系,它是一个"绝对指标",即某篇论文在发表后,有多少论文参考该论文进行研究。被引频次数字越大,则说明该文的学术影响力越大[1]。

本文以布拉德福定律和被引频次作为研究对象,来探讨信息计量法则在网络空间的适用性。国外的相关研究不均衡,定性研究比较多,定量研究和实际操作比较少。最早的是1997年英国Wolverhampton大学以Webometric为标题,对当前网络计量学研究的现状加以综述。在我国,2000年武汉大学的邱均平教授的《网络数据分析》一书的出版[2],标志着我国网络计量学这一学科体系的形成。网上信息资源的定量研究在国内外仍是起步阶段,没有实际操作的深入研究和成熟的研究方法,相关研究主要涉及网络链接研究的意义、存在的问题及发展趋势。将信息计量学的方法和技术应用到互联网中,必定会遇到新的问题,我们运用传统的信息计量学理论对网上信息资源的定量分析做探索性的研究,以期发现规律找出原因,实现科学的继承性。随着网络化的日益普及,加强网络管理已成为当务之急,网络信息计量学的研究成果必然会为网络管理的定量化和科学化提供理论指导和定量依据。

2 数据来源

在研究过程中,我们发现论文发表后2~3年为引用高峰

本文受陆军军官学院科研学术基金(2011XYJJ-071)资助。

王晓芳(1975-),女,硕士,讲师,主要研究方向为计算机信息处理,E-mail:wxfxj520@yahoo.com.cn;王 健(1966-),男,硕士,副教授,主要研究方向为信息处理;袁广林(1973-),男,博士,讲师,主要研究方向为图像处理;陈 萍(1977-),女,硕士,讲师,主要研究方向为图像信息处理。

期[3]，为了得到最有效的数据源，我们在 CNKI 中以“数据库安全”为主题检索 2008－2009 期间发表的论文，检索到学术论文 435 篇，其中有被引频次的论文 103 篇，按被引频次递减顺序排列，如表 1 所列。

表 1　关键词为“数据库安全”的论文被引频次降序表

序号 j	论文题目	刊名	Lni	第 i 篇被引频次 x_i	$Ln(x_i)$	前 i 篇被引频次累积 $X(x_i)$
1	数据库安全技术研究与应用	计算机安全	0	10	2.30258	10
2	浅谈 SQL Server 数据库的安全设计与应用	电脑知识与技术	0.69314718	6	1.791759	16
3	数据库加密算法的分析与比较	科技情报开发经济	1.09861228	5	1.609438	21
⋮	⋮			⋮	⋮	⋮
101	嵌入式数据库 SQLite 加密方法分析与研究	计算机应用与软件	4.615121	1	0	154
102	基于角色的数据库安全访问控制的应用	通信技术	4.624973	1	0	155
103	Access 数据库的安全	软件导刊	4.634729	1	0	156

表中 i 为被引论文序号，x_i 为第 i 篇论文的被引频次，$Ln(x_i)$ 为论文被引频次 x_i 的自然对数，$H(x_i)$ 为前 i 篇被引论文的累积被引频次。经计算，被引论文总篇数 $A=103$，总被引频次 $R=H(103)=156$。

3　区域分析

设布拉德福分区数为 m，当 $m=5$ 时，把表 1 的数据空间分 5 个区，令每区被引频次累积数相等，得平均值为 $A/m=156/5=31.2$，按此平均值计算各区的论文累积数 $H_n(n=1,2,3,4,5)$，得到分区的结果，如表 2 所列。其中核心区划分以 $X(x_6)=30$ 为界限，即核心区[4]含 6 篇论文，该区论文被引频次累积数是 m 为 5 时的平均值。

表 2　$m=5$ 时被引频次区域划分表

区号 n	第 n 区被引频次累积数	第 n 区累积数 x_n	前 n 区累积数 $X(x_n)$	$Ln(x_n)$	前 n 区被引频次累积数 $H(x_n)$	布拉德福常数 (a=Lnn/Ln(n−1))
1	30	6	6	1.7917	30	
2	31	14	20	2.6390	61	2.3333
3	32	20	40	2.9957	93	1.4286
4	31	31	71	3.4339	124	1.55
5	32	32	103	3.4657	156	1.0323

布拉德福定律的文字表述为：如果将科技期刊按其刊载某学科专业论文的数量多少，以递减顺序排列，可以把期刊分为专门面对这个学科的核心区、相关区和非相关区。各个区的文章数量相等，此时核心区、相关区、非相关区期刊数量成 $1:n:n^2$ 的关系。其数学表达式表示的是“期刊载文量累积数”与“期刊累积数”之间的函数关系[5]。

由于网络下载频次具有一定的布拉德福定律特征[6]，下面我们假设被引频次也具有布拉德福定律特征，通过实证分析来证明我们的假设成立，最后得出结论。

假设 1　学术论文的被引频次可以代替传统期刊的载文量，一样反映其学术价值。

假设 2　如果假设 1 成立，那么学术论文被引频次服从布拉德福分布规律，即 Web 被引总频次累积数与论文累积数之间满足布拉德福函数关系。

由表 2 可知，$m=5$ 时，布拉德福常数取值范围为 1.03～2.33，平均值为 1.585。区域法分析的结果不具有明显的布拉德福特征，主要表现在 5 区，布拉德福常数偏小。我们从以下几点查找原因[7]：

①网络空间与纸质文献空间有异构性，用户下载和引用论文行为不同，论文引用标准和网站引用规定也不同，这些环境的不同和约束都会导致布拉德福定律的特征表现不同。

②一篇学术论文在它发表数年后才可能被引用，引文分析具有滞后的特点，这会影响其特征的体现。下面我们采用更为直观的图像分析法来进行验证。

4　图像分析

为了验证假设即“论文累积数”与“论文的被引频次累积数”之间具有布拉德福定律描述的函数关系，我们参照布拉德福的做法[8]，以“论文累积数 x_n 的自然对数 $Ln(x_n)$”为 X 轴，以“论文的被引频次累积数 $H(x_n)$”为 Y 轴，以表 2 中的数据 $(Ln(x_n),H(x_n))$ 为坐标值作散点图，得到图 1。

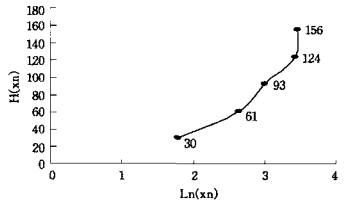

图 1　基于学术论文被引频次的 $(Ln(x_n),H(x_n))$ 散点图

从布拉德福分散特点上分析，此曲线具备一定的布拉德福分散特征[9]。我们再将曲线图的横坐标变换，以“前 n 区论文累积数 $X(x_n)$”为 X 轴，以“学术论文的被引频次累积数 $H(x_n)$”为 Y 轴，以表 2 中的数据 $(X(x_n),H(x_n))$ 为坐标值作散点图，得到图 2[10]。

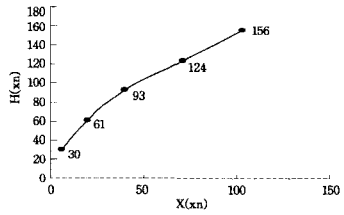

图 2　基于学术论文被引频次的 $(X(x_n),H(x_n))$ 散点图

将图 1 和图 2 中的曲线作比较，可以发现曲线的走向相似，曲线的直线部分相似，可以认为数据具有布拉德福分布特征，只是由于数据的不充分和网络空间数据的异构性，导致散点图特征不明显。下面，我们通过曲线拟合进一步验证。

5　曲线拟合

我们采用 MATLAB 对图 2 进行曲线拟合，曲线拟合的方法是利用线性、2 次多项式和 4 次多项式 3 种回归模型曲线对此散点图作曲线拟合[11]，通过拟合求得曲线方程和系数，根据系数的高低确定拟合的优度。

根据离散数据点确定 $(X(x_n),H(x_n))$ 散点图中坐标为 (6,30)、(20,61)、(40,93)、(71,124)、(103,156)，调用 polyift 命令对数据组进行多项式拟合[12]，拟合的多项式的最高阶数

为 n。

在 MATLAB 命令窗口输入如下的程序代码：

```
x=[6 20 40 71 103]
y=[30 61 93 124 156]
hold on
[p2,s2]=polyfit(x,y,2)
p2=-0.0069 2.0024 20.9692
s2=R:[3x3 double]
df:2
normr:6.8271
y2=polyval(p2,x);
[p4,s4]=polyfit(x,y,4)
p4 =0.0000 0.0000 -0.0218 2.7563 14.2381
s4=R:[5x5 double]
df:0
normr:6.0292e-014
y4=polyval(p4,x);
plot(x,y,'ro')
plot(x,y2,'g-')
plot(x,y4,'m--')
xlabel('x')
ylabel('y')
legend('原始数据','2次拟合','4次多项式拟合')[13];
```

图 3 所示为拟合的多项式的曲线图像，'o'为原模型，'—'为 2 次多项式模型，'- - -'为 4 次多项式函数模型。s4 的均方误差为 normr:6.0292e-014，s2 的均方误差为 normr:6.8271，s4 的均方误差小，说明提高多项式的次数可以提高拟合精度，图 3 中 4 次多项式曲线拟合最优[14]。

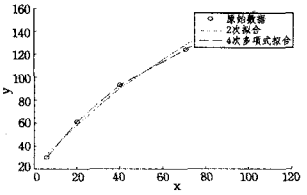

图 3 "论文被引频次"的多项式曲线拟合图

我们再在 MATLAB 中用图形用户界面进行曲线拟合。操作如下：

①在命令窗口中输入要拟合的数据，用 Plot 画图；

②在 Figure 窗口选中 Tools 菜单的 Basic Fitting 选项；

③在 Plot fits 复选框中选择 linear、quadratic、4th degree polynomial 选项，进行线性、2 次和 4 次多项式的拟合，窗口右部分为图像和均方误差信息。

命令行拟合和图形界面拟合的结果如图 4 所示。

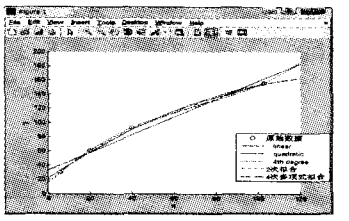

图 4 对原始数据进行命令行拟合和图形界面拟合的结果

结束语 本次研究说明针对被引频次这一网络信息计量指标，布拉德福定律仍然适用，证明了此研究是可行的。由于网络结构的复杂性和易变性，需要将数据进行深加工，找出其特有的共性，构造出适用于网络计量学本身的、通用的模型[15]，相关研究在国内尚不成熟，还有很长的路要走。

参 考 文 献

[1] 王召兵，陈燕. EndNote 在网络信息计量分析中的应用[J]. 情报探索，2011(1)：95-96

[2] 王召兵，王标，孔繁超，等. 网络信息计量在高校图书馆绩效评估中的应用[J]. 山东图书馆学刊，2011(4)：68-70

[3] 张洋，张淑玲. 中美医学院网络信息计量指标的比较分析[J]. 图书情报工作，2011(4)：26-29

[4] 殷之明，冷熠. 网络信息计量实证研究——中国社会科学院研究所网站评价[J]. 科技情报开发与经济，2009(19)：106-108

[5] 万锦望，花平寰，孙秀坤. 期刊论文被引用及其 Web 全文下载的文献计量分析[J]. 现代图书情报技术，2005(4)：58-62

[6] 张洋，弋云. 应用网络信息计量指标测定我国图书情报学核心网站的实证研究[J]. 图书情报知识，2011(1)：82-87

[7] 张洋. 期刊 Web 下载总频次的布拉德福分布研究[J]. 图书情报知识，2006(6)：38-42

[8] 沙勇忠，阎劲松. 网络著者分布规律实证研究：以 Python. cn 论坛为例[J]. 图书·情报·知识，2006，114(6)：17-21

[9] 崔旭，邵力军. 揭开布鲁克斯公式 K、N 关系之奥秘[J]. 情报杂志，2003(09)：42-43

[10] 赵隽. 基于布拉德福定律区域法的学术论文分布研究[J]. 现代情报，2007(05)：26-28

[11] 申红莲. Matlab 中曲线拟合的方法[J]. 福建电脑，2010(7)

[12] 马卫东，李幼平，周明天. 万维网无尺度特征与主动服务网格[J]. 计算机科学，2005(9)：31-34

[13] 赵丹群. 试论引文分析方法的网络化发展与应用[J]. 图书情报工作，2009(8)：40-43

[14] 马晓佳. 网络引文分析与传统引文分析的比较[D]. 南京：南京大学，2011

[15] Wallace D P. The relationship between journal productivity & obsolescence[J]. Journal of the American society forinformation science，1986，37：135-136

[16] Goffman W，Morris T G. Bradford's law and library acquisitions [J]. Nature，1970，226：922-923

(上接第 314 页)

[10] Fowlkes C，Belongie S，Chung F，et al. Spectral grouping using the Nyström method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2004，26(2)：214-225

[11] 史卫亚，郭跃飞，薛向阳. 一种解决大规模数据集问题的核主成分分析算法[J]. 软件学报，2009，20(8)：2153-2159

[12] Ng A Y，Jordan M I，Weiss Y. On spectral clustering：Analysis and an algorithm[C]// Advances in Neural Information Processing Systems. Cambridge，MA，MIT Press，2002，14：849-856

[13] Sander T D. Optimal unsupervised learning in a single-layer linear feedforward neural network[J]. Neural Network，1989，12：459-473

[14] Kung S Y，Diamantaras K I. A neural network learning algorithm for adaptive principal component extraction(apex)[C]// Proc. of IEEE Conf. on Acoustics，Speech，and Signal. Albuquerque，1990，2：861-864

[15] Weng J，Zhang Y，Huang W S. Candid covariance-free incremental principal component analysis[J]. IEEE Trans Pattern Analysis. Machine. Intelligence，2003，25(8)：1034-1040