

文献信息增长模型 与网络信息增长模型

曾雪鹃

(南京大学信息管理系, 江苏南京 210093)

摘 要: 本文首先介绍了一般信息增长理论, 然后在此基础上分别介绍了文献信息增长模型, 如指数增长模型、逻辑增长模型等, 以及网络信息增长模型, 如 Mike Thelwall 的网络信息增长模型以及乘数扩张模型。最后, 从研究对象、研究方法、应用范围、规律性 4 个方面对文献信息增长模型与网络信息增长模型进行了比较。

关键词: 信息增长模型; 文献信息增长模型; 网络信息增长模型

中图分类号: G353.1 **文献标识码:** A **DOI:** 10.3772/j.issn.1674-1544.2008.04.003

Documentary Information and Web Information Growth Model

Zeng Xuejuan

(Information Management Department, Nanjing University, Nanjing 210093)

Abstract: This paper introduces the basic theory of information growth, and subsequently expounds respectively documentary information growth and Web information growth, including Index Growth Model, Logic Growth Model, Mike Thelwall's Growth Model, Multiplier Extension Model and so on. Finally, the author compares the two kinds of information growth model from the perspective of research object, methodology, application and regularity.

Keywords: information growth model, documentary information growth model, Web information growth model

文献信息计量学的研究对象是文献或文献情报, 范围较窄^[1]; 而网络信息计量学^[2]的研究对象则是网络信息。越来越多的学者认为, 网络链接和期刊文献引用虽然在形式上具有某些相似之处, 但两者存在的差异是显著的^[3]。本文将在一般信息增长理论的基础上, 详细介绍文献信息计量模型与网络信息计量模型, 并对二者从不

同的角度进行比较。

1 一般信息增长理论

一般信息增长理论主要针对泛指的信息, 并没有特定的信息载体。一般信息增长理论中最常见的定律是洛特卡定律, 此外, 比利时学者 Leo

第一作者简介: 曾雪鹃(1984 -), 女, 南京大学信息管理系情报学在读研究生, 研究方向是网络信息资源管理。

收稿日期: 2008 年 4 月 15 日。

Egghe 对信息增长理论有专门的研究。

1.1 扩展的洛特卡定律——Leo Egghe 的信息增长理论

洛特卡定律是信息计量学三大定律之一,由洛特卡提出。该定律指在某一时间内,写了 x 篇论文的作者数占作者总数的比例 $f(x)$ 与其所撰写的论文数 x 的平方成反比,即 $f(x) = c/x^2$,其中 c 为特征常数。洛特卡定律描述的是科学论文作者的分布规律。

学者对洛特卡定律进行了扩展,其表达式成为 $f(x) = c/x^n$ 。其中 n 为洛特卡指数,不是精确值,而是一个约数,大约在 1.2 和 3.5 之间; $c > 0, x = 1, 2, 3, \dots, x_{\max}$ 。 $x_{\max}$ 表示一定时间内作者的最大撰文能力。

Leo Egghe 在 2000 年的研究^[4]中分析了一次引文的分布和引文年限分布情况之间的关系,发现这两种分布情况之间的联系与洛特卡指数有关;当洛特卡指数为 2 时,这两者的分布情况是相同的。并且作者还通过数学研究说明, n 次引文分布与一次文献分布也是相同的。

在这项研究中,Leo Egghe 扩展了洛特卡定律的应用,不是只针对作者的分布,而是针对引文的分布。Leo Egghe 使用洛特卡定律分析文献的被引次数 A ,即 $\varphi(A) = C/A^\alpha$,其中 $\varphi(A)$ 表示 A 的分布。经过一系列论证得出,当 $\alpha = 2$ 时,一次引文的分布与总体引文年限分布相同。

之后,在 2006 年,Leo Egghe 扩展了 2000 年的这项研究^[5],使其不再仅仅针对引文分布,而是扩展到一般的信息资源分布情况。于是根据已有的洛特卡函数得出了一个随时间改变而变化的洛特卡函数,扩展了原来的洛特卡定律,将 n 次引文的累计理论概括为依时间变化的洛特卡信息计量规律:文献和引文被认为是信息资源和信息记录,累计 n 次引文的分布模型被深入表述为一个模型,其变量包括时间、规模分布和排列分布的函数,并且还得出随时间变化的 h 指数和 g 指数。 h 指数和 g 指数表示作者的出版物的最高被引次数。与洛特卡定律结合起来后,就能计算出这两个参数,这两个参数都随时间的变化而变化。

2 文献信息增长模型

2.1 指数增长模型

这一模型源自 1944 年赖德对美国有代表性的大学图书馆的藏书量所进行的研究,当时他发现藏书量平均每 16 年递增一倍。于是普赖斯将这一发现推广,一方面研究了单种科学杂志的数量随着时间变化而产生的变化趋势,另一方面还研究了杂志种类数目随时间变化的趋势以及期刊论文数量增长的特点。

最后,指数增长模型由普赖斯提出,并画出了著名的普赖斯曲线,该增长规律被描述为 $F(t) = ae^{bt}$ ($a > 0, b > 0$),其中 $F(t)$ 表示 t 年的文献累积量, a 为初始文献量, e 为自然对数的底, b 为持续增长率^[1]。这里要特别指出的是,在这个模型中,所使用的信息计量单元并不是研究中采用得最多的期刊或图书,而是文献。这样的信息计量单元使得该模型可应用的范围更加广泛了。

2.2 逻辑增长模型

前苏联科学家弗·纳里莫夫发现文献的增长是分阶段的,每个阶段的增长模式都是不同的。在与格·弗莱杜茨进行大量研究后,他们最终提出了逻辑增长模型,即 $F(t) = k/(1 + ae^{-bt})$ ($b > 0$),其中, t 为时间, $F(t)$ 表示 t 年的文献累积量, k 表示文献累积量的最大值, a, b 均为设定的参数, e 是自然对数的底^[1]。

纵观逻辑增长模型从萌芽到最后被正式提出的过程,学者们都是以科学文献为信息计量单元,并且是以某一学科内的文献或某一研究方向上的文献作为信息计量单元。在这个模型中,主要用到了文献数量的绝对值指标以及一段时间内的文献数量累积数。

2.3 其他文献增长模型

除了上面两个重要的文献信息增长模型之外,还有一些其他的文献信息增长模型,如线性增长模型、分级滑动指数模型、超越函数模型、舍-布增长模型。这几个模型都以科学文献作为信

息计量单元,并按照学科或某研究方向分别研究了科学文献的累积数。

3 网络信息增长模型

3.1 Mike Thelwall 的增长模型

引文分析是文献计量学的一个部分,情报学和计算机科学中很多对于超链接的研究都汲取了引文分析中的方法和思想,因此情报学的超链接分析直接将引文分析的方法用到了网络中。2003年, Mike Thelwall 和 Liwen Vaughan^[6] 共同建立了一个多重回归模型,用数学建模的方法分析了加拿大大学网站之间的超链接方式。

这项研究所采用的原始数据都是从加拿大的大学网站中获得的。作者首先获取加拿大各所大学的网址,然后采用一种情报学网络爬虫工具,从主页出发跟踪链接,从每所大学的网站上获取了链接信息。如果大学的主页上并不包含 html 链接,爬虫就会选择包含到达所有部门主页的链接的页面作为跟踪的起点。最后,爬虫工具总共获取了 3930113 张网页。

研究所采用的另一个数据源是大学的档案数据,提供了大学的各方面信息,包括院系质量、学生素质和大学图书馆质量等。这些数据是从加拿大首屈一指的《Maclean》杂志得到的,因此权威而可信。Mike Thelwall 最终所采用的数据有院系研究的考察数据,包括每 100 个院系中来自社会科学和人文研究学会的研究奖项数目(后面简写为社会学奖项数目)及每 1000 名国家级奖项的获奖者中全职教授的人数(后面简写为院系奖项人数)等。

Mike Thelwall 还采用了一个补充数据源,即从第 16 版的世界大学手册来获取其他的大学信息,以及从加拿大自然资源官方网站来获取每所大学的经纬度数据。

作者构建了一个模型:

$$\text{inlink} = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

模型中的因变量 b 是入链数(入链是指从其他网站链接到该网站的链接),自变量 X_1 、 X_2 、 X_3 、 X_4 分别是社会科学奖项数目、自然科学奖项数

目、学生获奖人数和大学使用的语言种类。作者采用多重回归分析的方法建立这个模型,用于表述所有变量之间的关系。需要说明的是,这个模型是建立在一定的假设基础之上的,即大学网站的入链数可以通过院系研究成果数据、学生素质数据和大学所采用的语言种类加以解释或是预测而得知。该模型是一个标准化回归模型,即不同的自变量组合都会被标准化,使得回归系数更加接近,通过对应某变量的系数在统计时是否重要也能显示出假设是否成立。

使用爬虫工具从大学网站中获取的链接有两种,即入链和出链。这两者密切相关,因此作者只考察了入链的数目。所使用文档模板不同(如信息计量单元是网页、文件目录还是域名),计算链接的方式也不同。为了选出最佳的文档模板, Mike Thelwall 使用了这 3 种计量单元分别对链接进行计量,每一种计量单元与 Maclean 的大学排名都相关,在考察了多种文档模板之后,他发现以域名为计量单元是相关度最高的,因此在域名层面上进行计量成为了最佳选择。

Mike Thelwall 的这个模型,通过链接分析计量了网络信息,即以链接为单元计量了网络信息。从模型中可以看到,作者的研究重点放在了入链数目上面,并且认为可以通过入链数目反映出链数目,所以计量的对象可以简化为一种链接。这种方法避免了繁琐的计量工作,更重要的是避免了重复计量带来的误差。但是这种研究方法无法解决共同作者带来的问题,因为在计量过程中无法将不同地域的研究进行比较,目前可能的方法就是将数据均分,但这种方法的有效性和正确性暂时还没有得到验证。

3.2 乘数扩张模型

2003 年,国内学者侯经川等^[7]根据信息传达的不守恒性和网络信息传播的全达性,在“信息转发”假设和“信息创新”假设的基础上,将银行货币乘数机制与网络信息增长机制进行类比研究,从而提出了网络信息增长的乘数扩张模型。这个模型对网络信息总量的增长、网络真实信息的增长、网络泡沫信息的增长分别进行了计量和总结,最终得出乘数扩张模型。

网络信息总量增长模型: $I_m = n[(m-1)n+1]A$

网上真实信息量的增长模型: $I_r = mnA$

网上泡沫信息量的增长模型: $I_f = I_m - I_r = n[(m-1)n+1]A - mnA = n(n-1)(m-1)A$

网络信息的乘数扩张率: $I_r/I_f = n(n-1)(m-1)A/mnA = (n-1)(m-1)/m \approx n-1$

这个模型与其他模型的不同之处在于,它并不是直接对信息进行实证计量研究所得出的结果,而是通过类比,将货币银行学中的理论类比到信息中。作者对这种类比方法所提出的理论依据是网络信息和货币的相似性,即两者均为流通的物质,且可以重复使用,还采用了分布式网络结构的形式组织起来。这种通过类比而得出的模型创新性地对网络信息进行了计量,并且通过作者的论证得出该模型对于文献信息的计量也有一定的适用性。这个模型独特地将计量重点放在了泡沫信息上,因此还可以用于网络垃圾信息的控制。

4 文献信息增长模型 与网络信息增长模型的比较

4.1 研究对象

文献信息增长模型的研究对象仅限于文献或图书等,如普赖斯的指数增长模型,正是针对文献累积量进行研究,借此反映信息的情况。网络信息增长模型的研究对象则是多种多样的,针对前面提到的3个层次的网络信息进行研究,它们包括链接、网络影响因子、网络文献等。

4.2 研究方法

文献信息增长模型中的研究方法与研究对象对应,如逻辑增长模型的提出就是计量了科学文献的数量。网络信息增长模型则是采用链接分析法等,借助特殊的统计工具对信息进行分析,如通过网络日志搜集数据、监测网站,以及利用链接分析法、数据挖掘法等,从这些渠道得到的数据都可以用于网络信息分析。

4.3 应用范围

文献信息增长模型作为网络信息增长理论的基础,在网络信息计量研究中常被借鉴。另外,

文献信息增长模型本身就反映了文献信息的发展情况。网络信息的形式多种多样,因此网络信息增长模型的应用范围也很广,如评估搜索引擎、研究网站关系、控制网络信息等。

4.4 规律性

因为文献信息的复制存在一定的困难,所以它的复制率相对较低,文献数量信息的变化相对也较慢。由于网络信息资源与文献信息资源不同,可以无限制地快速复制,所以它的数量变化规律更加复杂,并且变化的速度也更快,网络信息增长模型也就更富于变化且更加复杂。

5 结 语

网络信息增长规律的研究,从最初将传统信息文献计量学的方法应用到网络中,发展到现在已经有了很多独立的方法,学者们能够从链接和网站等方面对网络信息进行分析。而文献信息计量方法在网络中的应用研究,仍是网络信息计量研究的一个重要方面,随着研究的不断深入将会产生更多的成果。

参考文献

- [1] 邱均平. 信息计量学[M]. 武汉: 武汉大学出版社, 2007.
- [2] Almind T. C., Ingwersen P. Information Analyses on the World Wide Web: Methodological Approaches to 'Webometrics' [J]. Journal of Documentation, 1997, 53(4): 404-426.
- [3] 赵蓉英, 段宇锋, 邱均平. 网络信息计量学研究 (I)——网络链接研究的现状及趋势 [J]. 情报学报, 2005, 24(2): 181-192.
- [4] L. Egghe. Theory of First - Citation Distributions and Applications [J]. Mathematical and Computer Modelling, 2001(34): 81-90.
- [5] L. Egghe. Item - time - dependent Lotkaian informetrics and applications to the calculation of the time - dependent h - index and g - index [J]. Mathematical and Computer Modelling, 2007(45): 864-872.
- [6] Liwen Vaughan, Mike Thewall. A modeling approach to uncover hyperlink patterns: the case of Canadian universities [J]. Information Processing and Management, 2005(41): 347-359.
- [7] 侯经川, 赵蓉英. 网络信息的增长机制研究 [J]. 情报学报, 2006, 22(3): 267-272.