网络信息计量理论、工具与应用

孙建军　李　江　◎主编

科学出版社
www.sciencep.com

本书得到“985”二期工程“南京大学中国人文社会科学评价国家创新基地”的资助

网络信息计量理论、工具与应用

孙建军　李　江　◎主编

科学出版社
北京

内 容 简 介

本书在网络信息计量研究的基础上，系统介绍了网络信息计量的理论、工具和应用。具体内容包括：网络信息计量理论，如 Web 理论、网络链接分析理论、网络引文分析理论、网络日志分析理论、网络信息分布理论以及网络信息增长和老化理论；网络信息计量工具，如样本量控制工具、数据获取工具、统计分析工具、可视化工具等；网络信息计量学的应用，如网络环境下的学术信息交流、网络信息资源评价、网站健康度检查与网站设计，以及网络挖掘等。

本书可作为高等院校信息管理与信息系统等相关专业的本科生和研究生教材，以及信息服务人员、咨询人员、管理人员的参考用书。

图书在版编目(CIP)数据

网络信息计量理论、工具与应用/孙建军，李江主编．—北京：科学出版社，2009

ISBN 978-7-03-023961-7

Ⅰ．网… Ⅱ．①孙… ②李… Ⅲ．文献计量学－应用－计算机网络－信息管理 Ⅳ．G257 TP393

中国版本图书馆 CIP 数据核字（2009）第 012014 号

责任编辑：李 敏 林 剑 / 责任校对：朱光光

责任印制：钱玉芬 / 封面设计：鑫联必升

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

http://www.sciencep.com

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2009 年 2 月第 一 版 开本：B5（720×1000）

2009 年 2 月第一次印刷 印张：18

印数：1—2 000 字数：352 000

定价：56.00 元

（如有印装质量问题，我社负责调换〈明辉〉）

序

随着互联网的普及与广泛应用，越来越多的学科将研究对象拓展到互联网中，图书馆学、情报学也是如此：从传统文献资源建设到网络信息资源评价与利用、从传统信息组织与检索到网络信息组织与检索、从文献计量到网络信息计量……

网络信息计量学（webometrics）一词最早由 T. C. Almind 与 P. Ingwersen 在 1997 年提出。在构词方式上与 Bibliometric、Informetrics、Scientometrics 相比可知，webometrics 的计量对象为 Web 信息。Web 只是 Internet 中的一部分，对于整个 Internet 信息的计量，I. Aguillo 提出 Cybermetrics 一词，虽然同样译为“网络信息计量学”，研究对象却截然不同。一方面，在 Internet 中，Web 信息从 1.0 形式到 2.0 形式都备受关注；另一方面，E-mail、Telnet、FTP、Gopher 等类型的 Internet 信息难以计量，因此，Cybermetrics 远不如 Webometrics 那样受关注。

不过十余年的时间，Webometrics 领域已出现了丰富的研究成果。网络链接分析理论可谓最为成熟，应用也最为广泛。研究者们基于图结构研究链接的特征和链接的方向性，研究网站之间的学术信息交流模式，利用链接指标评价网络信息资源、筛选核心网页/网站，将链接分析算法应用于搜索引擎的检索结果排序，借助共链分析法挖掘潜在信息。

网络日志、网络引文本应与网络链接一样，成为网络信息计量学的核心研究对象之一，而事实却并非如此，直到近几年才有学者提出将网络日志与网络引文纳入网络信息计量学的学科体系。网络日志中包含丰富的 Web 用户行为信息，早已成为网络挖掘领域的研究对象，却迟迟未引起信息计量学领域学者们的关注，或许是因为技术上“挖掘”更易操作。对于网络引文的理解，目前仍存在分歧，有人理解为传统文献中的 http 类型的引文，也有人理解为网络环境中的文献引用关系。前一种网络引文可直接通过传统引文数据库（如 SCI，CSSCI）进行统计，而后一种网络引文可借助在线引文数据库（如 GoogleScholar）进行统计。

与文献计量学内容相似，在网络信息计量研究中，信息的分布与变化规律是核心内容之一。然而，这部分内容却未能得到深入的研究。当前信息计量学领域的学者研究网络信息分布与变化规律时，以“验证文献分布与变化规律在网络环境下是否适用”为主导思想，并未深入考查网络环境与网络信息自身的特征，如动态性、随意性（网络中的信息可以随意增、删、改），研究得出的结论自然也

缺乏说服力。深入了解 Web 信息特征是从事网络信息计量研究的基础。具体而言，Web 信息特征包括 Web“链接—内容”关系、Web 五区结构特征、Web 图结构特征、Web 的小世界特性等。

上述各个方面的研究中，工具都是必不可少的。选取样本之前，需要根据多个因素确定样本容量，然后借助商业搜索引擎、爬虫、链接数据库或网络档案获取这些样本，进而借助统计分析工具对样本进行分析，最后可辅以可视化工具将样本间抽象的关系绘制成二维或多维图。准确地说，这一套流程中各个步骤使用的工具都是通用型工具。通用型工具在特定研究中存在不足之处，如商业搜索引擎用于获取链接数据时，存在“不一致性”。虽然在网络日志分析、链接流程度分析等方面有 WebTrends、PageStrength 等专门工具，但适用范围小。因此，有学者基于“引文索引”设计“链接索引”，专门用于链接分析研究；也有学者自行设计程序，用于自己的网络信息计量研究。

本书尝试从理论、工具和应用三个维度构建网络信息计量学的内容体系。信息计量本身含有方法的特性，因此，理论部分兼谈方法，如网络链接分析法、网络引文分析法、网络日志分析法等。工具部分则按照信息计量研究的流程依次介绍样本量控制工具、数据获取工具、统计分析工具等。应用部分以理论部分为基础，借助工具部分中的各种工具，以实例的形式展示网络信息计量学广阔的应用领域与应用前景。全书由孙建军、李江拟定提纲并共同撰写前言部分和第 1 章，李江撰写第 2 章，郑曦撰写第 3、4、10 章，曾雪娟撰写第 5、6 章，张玲玲撰写第 7 章，戴伟撰写第 8 章，邓中华撰写第 9、11 章，曾雪娟、邓中华共同撰写第 12 章，董珏撰写第 13 章，叶晓飞撰写第 14 章，董珏、叶晓飞共同撰写第 15 章。全书由孙建军修改定稿。

本书得到“985”二期工程“南京大学中国人文社会科学评价国家创新基地”的资助，特此致谢。

网络信息计量学的理论体系尚处于探索阶段，我们竭力搜集整理国内外网络信息计量研究的理论与应用成果，参考 P. Ingwersen、M. Thelwall 等著名学者的观点构建了网络信息计量学内容体系，但限于精力与学识，书中难免存在不足之处，恳请专家与读者批评指正。

孙建军

2008 年 9 月于南京大学

目　录

第一部分　理　论

第二部分　工　　具

第三部分　应　　用

引言　从文献计量到网络信息计量

文献定量化研究可以回溯到20世纪初。1917年F. J. Cole和N. B. Eales首次采用定量的方法，研究了1543～1860年所发表的比较解剖学文献，对有关图书和期刊文章进行统计，并按国别加以分类。1923年E. W. Hume提出“文献统计学”一词，并将其解释为“通过对书面交流的统计及对其他方面的分析，以观察书面交流的过程，及某个学科的性质和发展方向”。1969年，文献学家A. Pritchard提出用文献计量学代替文献统计学，他把文献统计学的研究对象由期刊扩展到所有的图书、期刊资料。

目前，文献计量学已经成为情报学和文献学的一个重要分支，同时也表现出重要的方法论价值，成为情报学的独有研究方法。在情报学内部的逻辑结构中，文献计量学已渐居核心地位，成为与科学传播及基础理论关系密切的学术环节。现在全世界每年发表的文献计量学学术论文为400～500篇[1]。

网络信息计量学最早由T. C. Almind与P. Ingwersen在1997年提出。学者们以“文献计量学在互联网中的发展”为思路，基于文献计量学中的布－洛－齐定律研究了网络中的布－洛－齐定律、基于期刊影响因子提出了网络影响因子、基于引文分析理论发展了链接分析理论，构建网络信息计量学的内容体系。目前，网络信息计量学已经成为图书馆学与情报学领域的研究热点。

0.1　文献计量学

按照C. L. Borgman与J. Furner的观点，文献计量学主要是对文献及其相关过程的属性的计量[2]。其计量对象主要是文献量（各种出版物，尤以期刊论文和引文居多）、作者数（个人、集体或团体）、词汇数（各种文献标志，其中以叙词居多），其最本质的特征在于其输出务必是“量”[1]。

0.1.1　文献计量学的内容体系

0.1.1.1　基础理论

文献计量学是按照几个文献经验规律展开的，其中出现最早、影响最大的是普赖斯（D. J. Price）提出的累积优势分布。

（1）普赖斯模型

普赖斯模型主要是运用了一个重要原理——成功产生成功，这也是产生布－洛－齐分布的重要原因。普赖斯的累积优势分布就是建立在“成功产生成功”这一思想上的。为了建立累积优势分布，普赖斯采用了概率论中常见的缸模型。在初始条件下，单缸中有 r 个红球和 b 个黑球。现在连续地从缸中随机摸球：如果摸到一个红球，就算作一次成功，此时便往缸中加入 s 个红球；如果摸到一个黑球，就算作一次失败，此时则不必向缸中补充球。显然，当某种出版物或者引文，在其后的一段时间内进一步出版或被引用的概率增加时，单缸模型是适宜的。可以用它来考虑一个作者的文章是否将再次发表，一种期刊是否将包括某一课题的论文。多缸模型是指假设有 N 个缸，摸球的方式是：如果一个红球被摸出，加入 s 个红球到缸中；如果一个黑球被摸出，停止从这个缸中摸球，转入下一个缸，然后重复上述过程。多缸模型对于描述作者、期刊、论文等一组现象在成功产生成功作用下每一次试验是有用的。例如，一篇论文被引用，在其后的过程中这篇论文还可能被引用。

（2）伯利亚分布的单缸模型

伯利亚（Polya）等提出了描述统计后效或传播的单缸模型。当试验进行 n 次时，每次摸球之初和下一次试验之前，将被摸出的球同 c 个该种颜色的球和 d 个相反颜色的球一起放入缸内。由概率论可知，当 $c=0$，$d=0$ 时，红球被摸出的总次数 x 分布被称为负二项分布。每一次出现红球或黑球都会进一步增加这种颜色的球的出现概率，这样，在伯利亚一般模型中，成功的结果增大了进一步成功的机会。从另一角度来看，失败也增加了再次失败的机会。对于 $c>0$，$d=0$ 来说，x 的极限分布趋向于负二项分布。

伯利亚分布的单缸模型科学地描述了“成功产生成功”，奠定了文献计量学的理论基础。许多情报学家也充分利用此原理对文献情报的规律进行了分析。

（3）布－洛－齐分布系

布拉德福（S. C. Bradford）－洛特卡（A. J. Lotka）－齐普夫（G. K. Zipf）分布系是一系列经验定律，揭示了某一具体对象在其主体来源中的集中与分布规律。这些定律不仅在文献学、情报学的理论与应用中占有十分重要的地位，而且在社会科学研究的其他领域中也有着广泛应用。布－洛－齐分布系的结论从表面上看不尽相同，但采用的统计方法是相似的，而且在实质上也同属一类。

布拉德福定律描述了期刊分布的集中与离散特性，B. C. Vickery 通过对布氏定律的研究，将载有某一学科论文的期刊按刊载相关论文数量的递减顺序排列，并按照每个区的相关论文数量相等的原则划分出 m 个区，则各区中的期刊数量成等级分布，其图像分析描述了累积频次——对数等级分布。

洛特卡定律指出了科学领域里论文的作者频率和论文数量之间存在着规律，

定律中的两个变量是作者数和与之相关的论文数。当按其著作生产增长的顺序排列作者之后，便得到作者数与相应的论文数的分布率。

齐普夫定律刻画了不同词汇的使用和出现频率之间的规律，定律中的两个变量是正文中不同的词汇数量和它们出现的频率。

从统计方法上看，这三大定律都研究两组数据中的某一具体排列：观察值和等级，并将二者进行特殊排列得到简单模型；从结论上看，这三大定律经过适当的数学变换后，都可以用反比函数表示，即 $F = c/\gamma^{\alpha}$。其中，F 为某种事项（如论文、词汇等）出现的频次；γ 为相应的等级数；c 为与样本量有关的常数；α 为与统计对象有关的常数。方法与结论的相似性使得人们将这些类型的分布规律称为布 - 洛 - 齐分布系。

（4）文献交流过程中的两种效应

文献交流过程中的两种效应分别指文献变化规律中的马太效应与用户需求特征形成的波敦克效应。

马太效应是指成功导致成功，形成累积优势。文献作为科学社会运行的载体，具有自组织特征。根据哈肯的协同理论，文献群作为一个开放系统，不断与外界交换能量，对自身进行组织和更新，形成系统特有的规律，即文献的集中与离散分布规律、增长与老化规律、文献作者分布规律（例如，10% 的科学家发表的论文数是所有科学家发表的论文总数的一半）、文献信息浓缩与冗余规律（例如，《化学文摘》中 75% 的文摘条目来自 9% 的期刊），这些规律都体现着马太效应。

波敦克效应是由美国科学社会学家 J. Gaston 提出的，用来阐释因地区、名望、技术差异而导致信息用户选择文献中形成行为习惯趋向集中的现象。信息用户的信息需求特征多种多样，这与用户的心理有一定关系，也与用户所处的外在环境和自身素质密切相关，可以归纳为：省力特征（齐普夫提出的最小省力原则）和可靠性特征（经典著作的可靠性更高）。

0.1.1.2　方法与工具

（1）文献数量统计

文献数量统计（simple document counting）主要统计某一学者、机构、地区、国家发表的学术论文数量，按其在时间轴上的分布来分析其发展趋势；也可以统计某一学科、主题领域的文献数量，按其在时间轴或地域轴上的分布来分析其发展趋势或进行横向比较。

文献回顾或综述可以初步判断某一领域的研究现状及发展趋势，文献数量统计方法是最常用的定量方法之一。这种方法所借助的工具包括各种学术数据库（全文数据库、文摘数据库、引文数据库等），如 CNKI、LISA、SCI 等。统计文献数量时，只需按学者、机构、地区、国家、学科、主题等统计即可。

（2）词频分析

词频分析（word frequency analysis）是图书馆/情报与文献学常用的研究方法，主要以词频的高低揭示科学研究中各研究主题的受关注程度。这里的“词”指能够表征文献主题特征的关键词，而不包含泛指的词，也不包含无实际意义的助词等。一般而言，词频越高，表明这一主题词受关注程度越高。受关注程度高的关键词可代表该领域的研究热点。

学术论文一般要求在摘要之后给出 3～7 个关键词，这些关键词成为词频分析所需数据的主要来源。全文数据库、文摘数据库、引文数据库都含有论文关键词，所以各种学术数据库都可作为工具，为词频分析提供数据。在具体统计过程中，EXCEL 可作为工具统计各关键词的词频。

（3）引文分析

引文分析（citation analysis）即引文分析法，就是利用图论、模糊数学、统计学等方法，对引文年代、内容、类型、语种、国别、引文集中与离散规律、同引与耦合现象等进行的分析，以便通过文献的变化揭示来反映更广泛、更深刻的社会、科学、经济、文化的多种动态与走向[3]。

引文分析主要以引文为数据来源，以引文索引为工具，如 SCI、SSCI、A&HCI、CSCD、CSSCI 等。

（4）共词分析

共词分析（co-word analysis）就是通过统计一组词两两出现在同一篇文献中的次数，将其进行聚类分析，以反映出这些词之间的亲疏关系，进而分析这些词所代表的学科和主题的结构变化[4]。两个词的“共词强度”（指两个词同时出现于一篇论文中的次数）越高，则这两个词之间的关联越紧密。J. Law 指出“共词分析法”最早起源于 20 世纪 70 年代中后期[5]，属于内容分析法的一种。

最早提出共词分析假设前提的是 Whittaker。他指出选择文献作为共词分析的假设前提主要有：①作者都是很认真地选择他的技术术语；②当在同一篇文章中使用不同的术语时，就意味着它们之间有一些关系并不微不足道，它们一定是被作者认可或要求的；③如果有足够的不同作者都对同一种关系认可，那么这种关系可以认为他们所关注的科学领域具有一定的意义；④当关键词被用于分析时，第四个论据被提出来，即经过培训的标引者选择出来的用来描述文章内容的关键词，事实上是可以信赖的一个指标。只有这些前提都成立，共词分析法利用文章中词语对的共现频次来反映包含在文章中的概念结构才会成为可能[6]。

0.1.1.3 应用

M. Thelwall 把文献计量学的应用分为两个方面，即评价与挖掘，并提出“评价式文献计量学”（evaluative bibliometrics）与“挖掘式文献计量学”（relational

bibliometrics）两个概念[7]。

（1）评价式文献计量学

评价式文献计量学源于一个基本假设前提：引用代表推荐或认可[8]。基于此，被引用次数（times cited）便成为科学价值的一个指标，因此引文便成为评价式文献计量学的主要数据来源。事实上，引用动机非常复杂，按照1965年E. Garfield的研究，文献引用的动机大致有“对开拓者表示尊重”、“对有关著作给予荣誉”、“核对其所用的方法及仪器”等15种情况[9]。1986年，A. T. Brooks根据前人的研究，将引文的动机分为7类[10]：新颖性、负面证据、操作型信息、说服、正面评价、提醒和社会认同。1994年，L. M. Baird和C. Oppenheim在E. Garfield的15种引文动机的基础上提出了“受作者的师长影响而引用”、“在不慎重的情况下引用”等17种引文动机[11]。此外，科学领域中存在的马太效应使得被引用次数并不能客观地反映科研成果的价值，但是在没有更合理的指标出现之前，被引用次数这一指标仍被各个科学领域采用。

被引用次数被应用于评价学者、机构、地区科研成果的学术影响力。20世纪60年代，E. Garfield在被引用次数的基础上提出了“期刊影响因子”（journal impact factor，JIF）这一指标，用于评价学术期刊的影响力。

（2）挖掘式文献计量学

早期的挖掘式文献计量学受限于计算能力与可视化技术。尽管如此，从早期的研究中我们仍可以获得一些有用的线索，如借助简单工具（关键论文集中引文流的网络图），我们可以粗略地挖掘出潜在的科学结构[12]。作者共被引分析（ACA）可以测度文献之间的相似性，可用于构建学科知识地图（图0-1）。

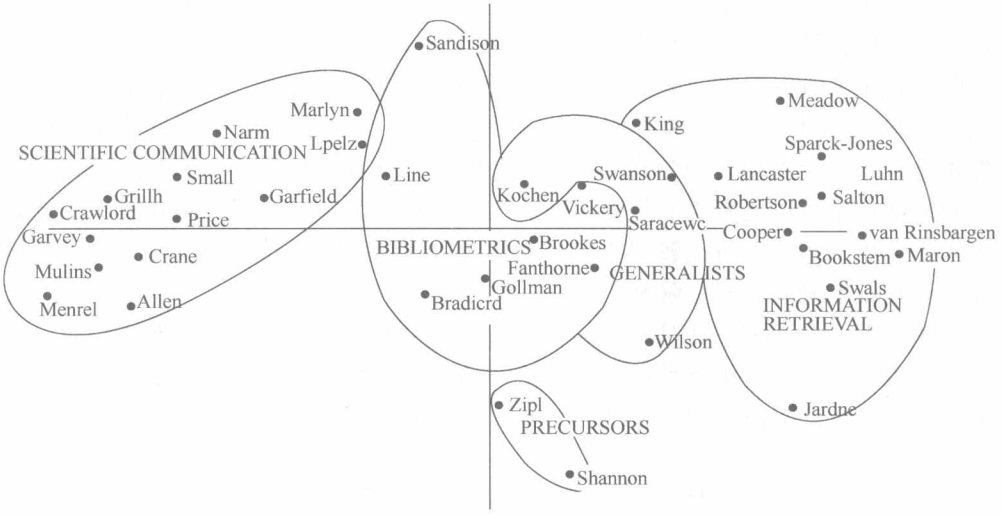

图0-1　1981年H. D. White与C. G. Belver绘制的信息科学领域知识地图[13]