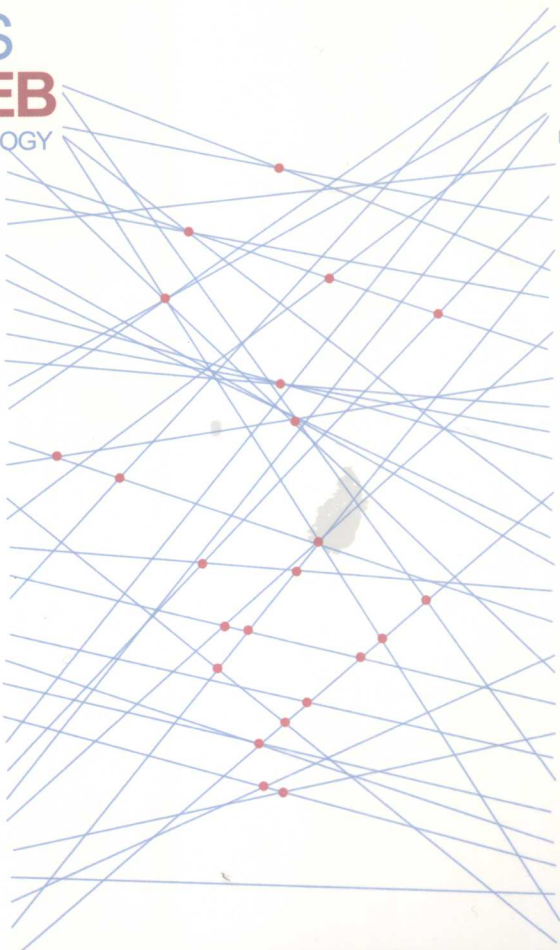
万维网的定律

——透视网络信息生态中的模式与机制

〔美〕伯纳多 A. 胡伯曼 著　李晓明 译

THE LAWS
OF THE WEB
PATTERNS IN THE ECOLOGY
OF INFORMATION



北京大学出版社
PEKING UNIVERSITY PRESS

万维网的定律

透视网络信息生态中的模式与机制

[美]伯纳多 A. 胡伯曼 著

李晓明 译

北京大学出版社
PEKING UNIVERSITY PRESS

版权登记号　图字：01-2008-3558

图书在版编目(CIP)数据

万维网的定律：透视网络信息生态中的模式与机制/(美)伯纳多A.胡伯曼著；李晓明译. —北京：北京大学出版社，2009.7

ISBN 978-7-301-15489-2

Ⅰ.万…　Ⅱ.①胡…②李…　Ⅲ.计算机网络－关系－社会行为－研究　Ⅳ.TP393　C912.68

中国版本图书馆CIP数据核字(2009)第114530号

ISBN 0-262-08303-5

书　　名：万维网的定律——透视网络信息生态中的模式与机制
著作责任者：[美]伯纳多A.胡伯曼 著　李晓明 译
责任编辑：王　华
封面设计：北京春天书装图文设计工作室
标准书号：ISBN 978-7-301-15489-2/TP·1037
出版发行：北京大学出版社
地　　址：北京市海淀区成府路205号　100871
网　　址：http://www.pup.cn　电子信箱：zpup@pup.pku.edu.cn
电　　话：邮购部62752015　发行部62750672　编辑部62752038
　　出版部62754962
印　刷　者：北京大学印刷厂
经　销　者：新华书店
　　787毫米×1092毫米　32开本　4.5印张　64千字
　　2009年7月第1版　2009年7月第1次印刷
定　　价：12.00元

The Laws of the Web

Patterns in the Ecology of Information

Bernardo A. Huberman

MIT Press

内容提要

亿万人在万维网上的各种活动，尽管是随机发生、随意展开、无人协调，但显现出了一些有趣的特征与明显的模式：从网络信息的结构，到网民冲浪的行为；从网络拥塞的现象，到电子商务的市场；从信息下载的策略，到社会网络的作用；等等。本书不仅以一种引人入胜的方式描述了那些模式，而且还令人信服地阐述了形成它们的机理；并且还深入浅出地介绍了获得那些机理的方法。本书所讨论的内容是学术性的，但展开的方式是通俗易懂的。本书的读者对象为一般知识分子，不仅受过科学技术训练的人员可以一口气将它读完，而且有一定自然科学知识背景的社会科学工作者也能轻松地欣赏到它的精彩。

中文版序言

李晓明教授热心地担起这项劳作，将我的书译成了中文，令我感到十分高兴。万维网的信息生态中不断显现着各种模式，有了这本书，在中国对那些模式有兴趣的同仁就可以来分享我的一些认识了。此外，看到中国在网络的发展方面有大量的研究活动，我希望本书的读者会发现书中所陈述的内容是有用的，能激励他们继续探索的旅程。

自从我写完这本书后，万维网技术已经有了许多新的发展，但我很高兴地发现，我在这本书中描述的那些定律今天依然是适用的，如同若干年前一样。主要的区别在于互联网泡沫消失了，人们广泛参与网络内容生产的现象出现了，网络信息的消费者与生产者的区别不再那么明显了。鉴于这种情况，我特别为此中文版增加了一个新的章节（第9章，网络时代的社会注意力），描述随着千百万的内容生产者正将万维网变成一个极其丰富和具有创造力的环境，注意力新经济如何变成了前沿课题。

伯纳多 A. 胡伯曼

2009 年 5 月

前　言

在短短一个时期里，网络（World Wide Web[①]）不仅成了亿万人获得信息的媒介，而且它的巨大规模、各组成部分之间复杂的关系，以及遍及全球的空间特性也引起了许多科学家的热切兴趣，成为他们研究的一个对象。这种兴趣植根于这样一个事实，尽管网络的成长是自发随意的，但它背后蕴藏着很强的规律：从它的信息之间的链接结构及其组织的方式，到亿万人使用它的模式。一些实验室正在努力发现和解释这些规律。尽管这些研究活动目前主要还只是试图解释那些现象，而且也只能部分地说明一些问题，但已经对实践有了指导意义。人们在从事这些探索研究中所获得的新知识能用来设计有巨大潜力的新颖的机制，提高人们使用网络的效率。

---

① 英文“World Wide Web”，现在比较正式的译法是“万维网”。但一般百姓谈到“网络”的时候，多数情况下指的就是它，比较通俗上口。因此在本书中我们也用“网络”对应“World Wide Web”，在有些强调或需要区别的场合也用“万维网”。（译者注）

在所发现的若干规律中，包括网络上的信息存储和链接的方式，个人利用它们的模式，以及人们在这样一种大规模的新媒体中寻找信息时相互作用的性质。有些结果的建立，采用标准的在线调查而得，与其他社会研究领域没什么不同。另外一些研究则有赖于统计学和经济学方面的方法。基于物理学和统计力学的理论模型，人们预测到了一些网络的规律或者法则，这很有意思，因为以前很少有人想到那些理论能应用于社会领域。

过去十三年来，在大规模分布式系统的研究中，从经济系统到互联网，我一直在应用统计力学和非线性动力学的方法。合作的人数虽然不多，但形成了一些重要的认识和有趣的应用。不幸的是，可能由于表达它们所采用的技术性语言，那些结果没有得到广泛关注。不过，有些人已经看到了我们的工作成果在其数学形式之外的意义，这是几年来敦促我花些功夫来阐释这种方法的动力，特别是针对非技术背景的读者。

这本书就是这样一种努力，它以网络作为讨论那些方法的载体，或者说是关注的焦点。网络已经是无处不在，它是大多数人都熟悉的话题，而且也是一个方便且独特的实验室，使人们可以来研究它自身的成长和结构。但是，比较容易研究它，不等于也容易将用数学公式表达的概念转换为普通的

语言。然而，我坚信从任何领域产生的概念都应该能用普通语言来解释。本书中，依靠日常生活中的经验，我试图通过简单的例子来传递其中一些主要的思想。

这样做的结果，一些本来用几个公式就可表达的内容可能就变成了一段长长的解释，有技术背景的读者会难以忍受，或者屈尊将就我为了清楚地传递有关概念而不得不包括的那些细节。对此，我表示歉意，但我也觉得我不知道有什么更好的方式来传递这些规律的品味和力量，能引起读者的注意，而不要求他或她事先花些时间来学一些技术领域的知识。更重要的是，我相信，虽然探索和解释这些规律会涉及某些技术上的难度，但任何人，只要他对这个主题有足够的兴趣，都可以认识到它们的重要性和影响。关于这类话题，我做过多次报告，听众也是非常多样化的。那些经验教育了我，非专业人员提出的问题和评论，是不断产生新想法的源泉，挑战我继续把工作做得更好。

针对非技术性读者写一本关于这些方法的书，最初是巴尔都·费耶塔(Baldo Faieta)向我提出的建议，他很早就被书中描述的那种方法吸引了。尽管他的建议很诱人，我还是有些畏缩不前，感到需要付出很大的努力，才有可能形成一个有良好逻辑的内容，并且足够内聚来保持读者的注意力。后

来，一些人——特别是斯考特·科里尔瓦特(Scott Clearwater)，伊坦·埃达(Eytan Adar)，纳塔丽·格兰仕(Natalie Glance)，我的妻子梅蒂(Mette)，我的孩子拉拉(Lara)和安德鲁(Andrew)，还有马克·鲁丽尔(Mark Lurie)和克里斯多夫·劳克(Christoph Loch)——坚信这样一本书的潜在价值，并多少向我保证了至少他们是读者。最后，在MIT出版社的罗伯特·普莱尔(Bob Prior)的耐心和有说服力的激励下，我在法国巴黎开始写作了，当时正在访问INSEAD，即欧洲商学院。在那个学院的讲学以及和同事们的讨论，帮助我进一步提炼了一些论点，而早期与纳塔丽·格兰仕和拉达·阿达米克(Lada Adamic)在一些科普刊物发表的文章，教会了我如何用平实的英文来阐述最初是用技术术语表达的论点。

尽管这本书是我的努力所成，我对它负有责任，但它的内容却是长期与许多同事和学生合作的结果，没有他们，我将无从所写。从我最初从事这方面的研究开始，泰德·哈格(Tad Hogg)在许多方面就一直是一路同行的伙伴。纳塔丽·格兰仕，拉詹·鲁克斯(Rajan Lukose)，拉达·阿达米克，伊坦·埃达，吉米·皮特考(Jim Pitkow)，塞巴斯蒂安·莫雷尔(Sebastian Maurer)，马特·弗兰克林(Matt Franklin)，亚历山大·阿奎斯特(Alessandro

Acquisti)，皮特·派罗利(Peter Pirolli)和阿米特·普尼亚尼(Amit Puniyani)在我研究的过程中曾给予许多帮助，与我分享他们的见解和兴奋。最后，塞巴斯蒂安·多尼亚克(Sebastian Doniach)以一种真正学者的健康的怀疑态度，让我对每个新结果一一进行解释，从而使书中所表述出来的内容得到了进一步的推敲。

如果没有一些支持和客观的读者能够对这本书的内容和风格进行评论，以及提出所需的改进建议，它是不可能完成的。我感谢罗伯特·普莱尔非常好地发挥了他编辑的作用。我爱的感激献给梅蒂，她不仅审读初稿，告诉我哪些需要修改，最要紧的，是她理解了写这本书所需要的付出。

伯纳多 A. 胡伯曼

2001 年 1 月，于帕拉奥尔托(Palo Alto)

目　录

第1章　E生态(网络生态)

在旧金山城风景如画的普雷西迪奥(Presidio)区的一个小楼里,一群人正在进行着相当于一种庞大规模的生态调查,但他们却不需要离开他们的办公桌。计算机工作站中的一个程序在互联网的疆域里"爬行"[①],为互联网档案馆(Internet Archive)的工作人员们源源不断地获取网页并将它们存储起来。为了未来的研究,他们要收集并存储整个网络上的文本内容,从硅谷的网站,到地球另一侧遥远的服务器上的网页。在某种意义上,他们不仅像是生态学家,而且也是在构造一个图书馆,其规模不久就要使世界上那些最大的图书馆,例如美国国会图书馆或法国国家图书馆,相形见小了。到2000年7月,他们已经收集了10亿网页,占33.5 TB存

① 英文词"crawl"被用来形象地描述一个程序从网上不断获取网页的行为,它的直意是"爬行",表示该程序顺着网页中的超链,递进地获得一篇篇网页的过程。那个程序叫"crawler"或"爬取器"或"爬虫"。(译者注)

储空间，并且其收藏的规模在以每月10%的速度增长[①]。为理解这个收藏的规模，我们可以对比书籍的情况。一本书的文字内容经数字化后大约对应1 MB(兆字节)数据，1 TB(万亿字节)是100万个1 MB，美国国会图书馆有2000万册书，数据规模因此大约是20 TB(不包括图片)。

尽管网络的规模以及它成长的速度是惊人的，但它的内容并不需要一个很大的建筑空间设施才放得下。不同于那些古老的机构，例如位于纽约的美国自然历史博物馆，或者伦敦的大英博物馆，互联网档案馆项目用一个冰箱大小的存储服务器就将其完整的收藏装在其中了。现在还不清楚整个档案最终会有多少，也不清楚有多少目前在网上看得见的东西几年后还会依然存在。互联网档案馆的目的是要为未来的研究，将这种流逝的数据保存到一个永久的介质中。他们感到科学家、历史学家，以及新闻记者等都会对那样的研究有兴趣。进而，为了看到这些收藏起来的数据内容，人们不需要专门来到普雷西迪奥这样一个美丽的地方。只要有合适的授权，从世界上任何计算机上都可以浏

① 在中国，北京大学网络实验室从2001年开始了类似的工作，到2009年已经搜集了30多亿中国网站上的网页，占超过50 TB存储空间，并且以每天约200万网页的速度增长，见 http://www.infomall.cn。(译者注)

览它们。

欢迎来到信息时代,这是一个几乎不用任何代价就可以访问最多样化、无奇不有的知识的时代。那些电子信号正从价值上取代工业生产出来的物品,用来体现国家、公司和个人的财富。而且,这才是刚刚开始。用不了几年,我们现在所看到的,将会显得陈旧、有趣和过时,网络信息空间将带给这个地球上的人们无与伦比的奇观。

互联网档案馆的工作人员并不孤立。在美国的另外一边,波士顿大学计算机科学系的一群研究人员在从事一项活动,从某种意义上讲与互联网档案馆的工作互补。通过在一些网络浏览器中安插一些检测程序,他们能记录用户访问过的网站,在访问过的网页上停留多长时间,等等。他们试图通过分析总结出一种使用模式,形成对人们在网上寻找信息行为规律的一种理解,从而可以指导人们设计更好的网站。其他相关的活动还有许多,例如一些科学家们开始针对网络的内容和结构进行研究,他们所表现出来的强烈兴趣,与生物学家们研究热带雨林时类似。往下走到新泽西,我们发现位于普林斯顿的NEC实验室的研究人员在对网络的内容进行抽样,希望能理解它的规模,确定最终有多少内容有机会被人们常用的搜索引擎列出来。结论是,其实并不多。

回到加利福尼亚，普雷西迪奥往南几英里，在环绕斯坦福大学的山丘上，施乐研究中心（Xerox PARC）互联网研究组的科学家们一直在好奇地探究和分析一个巨大的网络资源贮藏库。他们的工作，综合利用互联网档案馆和许多其他来源的数据，揭示了在过去三年里，网络中有若干隐形的模式，反映出人们在信息空间中交流和寻找信息的状况。除了他们外，其他研究人员也发现了一些模式，既令人惊奇，也十分有趣。例如，我们已经确立了，每个网站中的网页数，以及网页中的链接数的分布显现出一种普遍的、规律性很强的现象，即少数网站有巨量网页，而许多网站只有少量网页。同样我们也已经了解了，由网络冲浪者在网上造成的拥塞能导致一种可预测的互联网“风暴”，它会突然出现，并以一种有统计意义的模式消退。

这些结果令人惊奇的原因，是因为看到网络成长的那种混沌无序的方式，以前没有人预见会发现什么规律性。没有什么中央计划者告诉互联网的用户如何设计网站，寻找信息，或者如何组织可从一个网页或网站浏览到另一个的超链接。网络的结构和内容是千百万人自发和自由行动的结果，他们很少会想到在他们的网站中添加一个网页或链接会对全局造成什么影响。这样，自然就会期望对互联网的研究只能看到一种没有任何特征可言的

网络,而不是现在所发现的强规律行为现象。

这些很强的规律性有意思的另一个方面与它们的来源有关,反过来会关系到它们被解释的方式。我们不只是可以利用这些有规律的模式来设计更好的网站,或者设计应对网络拥塞的策略,适当地考虑网络用户和网站设计者的行为机制细节①,我们可以形成能够解释这些模式的理论。随着这些理论的发展,将个人在网上的行为与网络的累积效应关联起来,显现了一些在网络之外也适用的社会机制。它们通过一些法则或定律表现出来,解释人们如何冲浪,如何通过一种拥塞模式相互影响,以及如何确定给定网站的流行程度。从这种视角看,网络变成了一个巨大的信息生态系统,能用来定量地评测人类行为和社会性互动作用的理论。

在本书中,我将提供关于网络上显现出来的若干规律②的描述,以及它们对于理解某些社会现象、设计更好的信息访问机制的影响。我也想通过这样做来解释获得这些理论的方法,它们提供了一种理解社会动力学中复杂问题的新颖方式。理论的原始形式是通过数学形式化描述的,因而可以有比

① 即网络用户行为是形成那些有规律的模式的基本根源(译者注)。

② 根据上下文的方便,我们称它们为"法则"或"定律"。(译者注)。