●王知津　郑红军　张收棉

网络计量学的理论、方法及应用

摘　要　网络计量学研究的对象包括网络信息直接计量、网上文献信息及相关特征信息计量、网络结构单元的信息计量。网络计量学有其规律和研究工具。它具有内容新颖、广泛和大量运用统计学知识的特点。其研究方法有链接分析法、统计分析法和图论分析法等。网络计量学在网站评价与建设、信息资源开发与管理、网络信息检索,以及数字图书馆资源管理等方面,都有广泛的应用。参考文献 12。

关键词　网络计量学　基本理论　方法　应用　文献计量学　链接分析
分类号　G257

ABSTRACT　In this paper, the authors summarize the objects and characteristics of the studies in cybermetrics, and analyze the methods (such as link analysis method, statistical method and graph theory method) and the applications of cybermetrics. 12 refs.

KEY WORDS　Cybermetrics. Basic theories. Methods. Applications. Bibliometrics. Link analysis.
CLASS NUMBER　G257

一般认为,网络计量学的产生始于 1997 年 T. C. Almind 和 P. Ingwersen 发表的文章 *Informetric analyses on the world wide web: methodological approaches to "webmetrics"*。该文首次提出了"网络计量学"的概念来描述将文献计量方法用于万维网的研究,并且认为,"文献计量学的各种方法完全可以应用于网络信息的计量分析,只不过是将网络看成引文网络,传统的引文由 Web 页面所取代"[1]。由此可见,网络计量学是在文献计量学的基础上发展起来的。一方面,它是参照和借鉴文献计量学的理论和方法,用于对网络环境中的信息资源进行定量研究;另一方面,由于网络的动态化、无序化,不能简单地套用文献计量学的方法,而是不断开拓创新,形成自己的内容体系和研究方法。

1　网络计量学的基本理论

1.1　网络计量学的定义

关于网络计量学的定义有很多[2]。从网络计量学的研究现状及其发展趋势来看,笔者认为,网络计量学是综合采用数学、统计学、文献计量学等各种定量研究方法,结合计算机技术、网络技术,对网络空间上信息的组织、存储、分布、引证、利用等进行定量描述和统计分析,以便揭示网络信息内在规律和数量特征的一门新兴学科。网络计量学作为一门交叉性的边缘学科,在网络时代具有广阔的应用前景。通过对网络信息的计量研究,可以为网络资源的优化配置和有效利用,以及网络的规范化管理提供相应的依据。

1.2　网络计量学的研究对象

从网络计量学现阶段的研究来看,其研究对象主要有网络文献的计量、电子引文分析、数字图书馆、万维网的数据库分析、网络结构、电子资源、域名分析,等等。归纳起来,主要包括 3 个层次:网络信息的直接计量;网上文献信息及其相关特征信息的计量;网络结构单元的信息计量[3]。

在 1997 年召开的国际图联大会上,美国学者 T. 0'Neill 提交的论文 Characteristics of web accessible Information 率先深入探讨了万维网的统计指标、统计类型和其他问题,并集中分析了站点的数量、静止的网页数、交互式网页的数量、语言种类的分布、网页的平均寿命等问题[4]。这些是对网络信息外在特征进行测度和分析的,是对网络计量学有着开创意义的研究。

网络信息纷繁芜杂,不仅包括论文,还包括电子期刊、图书、报告、专利等各种类型的文献,对这些网络信息做好计量研究,不仅能促进对网络信息的开发利用,更能知道如何更好地去建设和管理网络信息。对网络上站点、聊天室、布告栏、电子邮件等结构单元中的信息增长、老化、分布和各种单元之间的引证关系进行计量研究,也是网络计量学的重要组成部分。

1.3　网络计量学的规律研究

在传统的文献计量学中，对文献分布规律、增长规律、老化规律等的研究趋于成熟。然而，在网络空间中，网络信息的特点决定了网络计量学规律与文献计量学规律既有相似之处，又有不同的地方。

万维网作为一个分布式结构的网络，已逐渐成为科研人员发表论文的场所，论文交流也日益频繁和便捷。论文并不非要在规定的站点上发表，任何人只要稍作投资，就能很容易地申请到站点域名，可以将所撰写的文章发表在自己的站点上，这就使得科学论文分散在整个网络上，而不是像以往那样集中在若干核心期刊上，从而加剧了信息的分散性。此外，网络的飞速发展使得网络信息的增长和老化呈现出新的特征。一方面，信息量以指数级的速度增长；另一方面，知识信息的更新速度之快使得网络环境中的文献让人有稍纵即逝的感觉，故文献计量学对纸质文献的增长规律和老化规律的研究模式已渐趋过时。因此，建立网络环境下高效、准确的理论模型已成为迫切需要解决的问题，尤其是站点的增加与老化，以及站点所记载科学内容的增长与老化方面。

1.4 网络计量学的研究工具

网络计量学作为一门以定量分析为基础的学科，主要是利用计算机通过有关数据库或搜索引擎来获取文献信息的统计数据，并进行相应的整理和分析研究。网络信息浩如烟海，如果没有功能强大的工具作保障，对网络信息的计量研究也就无从谈起。

美国科学情报研究所编制的引文数据库（包括《科学引文索引》、《社会科学引文索引》、《艺术与人文科学引文索引》以及国内成功研制的《中国科学引文数据库》和《中文社会科学引文索引》等，在一定程度上提供了研究所必需的大量数据，有效促进了网络信息计量分析研究的全面展开。

搜索引擎的强大功能使得它已经成为人们在网上查找信息的首选工具，网络计量学的研究也要借助于搜索引擎来展开。如 google 的 link 检索，能够查找连接到某个站点的所有网站和网页数。功能最强大的莫过于搜索引擎 AltaVista 了，它能提供多种类型的限制检索，如域名限制、主机名限制、超链接限制、文件限制等；它还具备布尔检索、截词检索、范围限制检索、动态分类检索、制定语种检索等功能。

1.5 网络计量学的特征

1.5.1 内容新颖，范围广泛

网络计量学以数字化的信息为计量对象，研究网络空间中信息的分布、增长、老化、引用规律，并利用其中的新规律、新指标去认识并指导信息交流与管理。而网络上信息的激增、搜索引擎的功能与质量、网页的类型与质量、网络条件下的影响因子、超文本之间通过超链接引用、网上知识发现和问题追踪等[5]，已经成为网络计量学的研究热点。

1.5.2 统计学知识的大量运用

计量研究就是要科学地分析所收集的数据，从有限的数据中总结出一般规律。在网络计量学中，也需要大量运用统计学的知识对网络中的数据进行整理、分析，利用各种方法找出共性，创建网络计量学本身所适用的模型。例如，R. Rousseau 对网站的互联和进入网页的链接进行了洛特卡分布的分析，M. Thelwall 分析了网络影响因子和搜索引擎的覆盖率，这些都是运用统计学知识进行网络数据分析所得出的结果[6~7]。

2 网络计量学的研究方法

理论体系有如基础，而方法则有如手段。网络计量学的研究，需要采用恰当的研究方法，从它现阶段的发展来看，大致可以分为以下几种类型。

2.1 链接分析法

链接分析法是在引文分析法的基础上发展而来的。在传统的以纸质印刷型文献为研究对象的文献计量学中，引文分析可以用于研究文献资源分布、确定核心期刊、研究科学交流和信息传递规律等。引文分析所需的大量数据主要是由 SCI、SSCI、A&HSCI、CSCD、CSSCI 等数据库提供的，而这几乎是引文信息的惟一来源。在网络环境中，文献以网页的形式显示在网站上，站点、网页之间通过超链接相关联，因而，网络中的链接就可以看做印刷型文献中的引文。在数据的收集过程中，搜索引擎为网络信息计量研究提供数据来源。

正如评价印刷型研究资料时广泛采用影响因子那样，Ingwersen 提出用“网络影响因子”（WIF）来测评 Web 站点域的影响，即某一站点的链接数比率[8]。根据这种观点，我们可以为一个 Web 空间计算 3 个 WIF：自我链接 Web 影响因子，用于测量 Web 空间自身页面之间的链接；外部 Web 影响因子，测量外部链接到该 Web 空间的链接；整体 Web 影响因子，测量到 Web 空间的所有链接。

链接分析法已经成为当前网络计量学研究中适

用最多的一种方法。通过分析站点被其他站点“引用”的情况，以及相应的网络影响因子，就可以确定核心站点，能够帮助用户快速查找和选择利用相关的信息，这也有利于推动网站建设。

2.2　统计分析法

运用统计分析方法进行研究是网络信息定量研究的基础，网络信息的收集、整理、分析都离不开统计学的方法。在传统的文献计量学中，对文献信息的统计分析业已形成相对完整的体系；而由于网络结构的分散和复杂，以及网络信息的丰富，网络计量学需要重新构建针对网络信息测度的统计指标体系，并在实际应用中将指标与各因素建立起对应关系，从而构建出数学模型，再不断地进行修正，使之趋于合理。

运用统计分析进行研究，最重要的就是如何收集数据。网络环境下，一方面可以使用搜索引擎来收集数据，这已经被人们广为采用；另一方面，采用网上日志文件和网上调查来收集数据也日渐流行[9]。网上日志文件有的是由 Web 服务器直接生成的，有的是第三方统计机构在服务器端加入的模块生成的。这种方法收集的数据能够保证真实可靠，便于认证度量，如度量访问者、网站访问量、访问者的特征及其度量等。网上调查是通过监测软件来记录网络的安全状况、网民的上网行为模式、网民对网站的评价、网民的分布情况等，从而为推理统计提供必要的数据。很多研究机构，如中国互联网信息中心、中国网络研究与发展中心等，都是运用这种方法来对网站数量、网络用户特征以及网络发展的增长进行统计分析的。

2.3　图论分析法

在图论中，图是网络的一种数学表达。网络是由结点和边所组成的，结点之间通过边相连接。在有向图中，边表示结点之间的定向联系，Web 就是有向图的一个例子，其中的网页对应于结点，而超链接则表示边。

近些年来，网络计量学的许多研究工作已经开始从图形的角度来对网络进行研究，分析网页间超链接的拓扑结构，以直观反映网页之间的连接关系。在众多将图论应用于网络计量学的研究中，较有影响的是 A. Broder 等人的研究[10]。他们利用搜索引擎 AltaVista 收集了 200 兆的网页和 15 亿个链接，并采用图形分析法对本地和全球网络图形结构进行分析，得出了看起来像“领结”形状的网络结构图，并在此基础上构建了网络图的数据库模型。

A. Broder 研究得出的“领结”形状的网络结构由 4 部分组成。第 1 部分是一个中间核心，每个网页都具有离开网页的链接和进入网页的链接，其中的所有网页都可以沿着有向链接到达另一个网页；第 2 部分称为“IN”，仅具有离开网页的链接，“IN”中的网页可以到达核心部分的网页，但不能由核心部分的网页到达“IN”中的网页；第 3 部分称为“OUT”，仅具有进入网页的链接，从核心网页可以到达“OUT”中的网页，但反之则不行；第 4 部分称为“TENDRILS”，其中的网页不具备任何链接，不能到达核心部分的网页，也不能由核心部分的网页到达该部分的网页。这种运用图论进行分析的方法，可以使人们更好地理解网络的错综复杂的结构。

3　网络计量学的应用

3.1　网站评价与建设

用户在享受网络所带来的海量的信息资源的同时，也在为如何快速、高效地找到所需信息而困惑。网站作为网络上承载信息资源的主体，质量良莠不齐，并且缺乏权威的认证，而用户处理信息的能力又是有限的，这就使得网络的利用受到极大限制。鉴于此，有必要从实际应用出发，创建一整套科学、合理的评价指标体系来对网站进行评价，确定出“核心站点”。

像引文分析可以确定核心资料源那样，核心网站的确定可以通过链接分析来实现。通过计算网络站点的被链接率，可以确定各网站的网络影响因子，从而对站点进行评测。网页的质量、网站的总体架构、网站的点击率等，也都是评价网络站点的重要指标。

“核心网站”的确立，一方面便于用户快速查找和选择利用网络信息资源，扩大网站的知名度，增加网站的利用价值；另一方面，在评价的过程中，能够找出网站当前存在的缺陷和不足，在以后的开发和建设中改进；再者，核心网站相对于其他站点来说，质量较高，可以作为今后其他网站建设的一个标准。

3.2　网络信息资源管理与开发

网络的开放性使得网络信息资源缺乏规范的表达和组织管理，这在一定程度上造成了网络信息的混乱无序。利用网络计量学的某些指标来统一表达

和组织网络信息，使之以某种规范的形式存在，更加便于用户使用，充分体现出网络信息的价值。

网络计量学对于用户需求和上网习惯的研究，对于网络信息资源的管理与开发来说都极为重要。一方面可以使用专用的软件进行动态的跟踪，另一方面可以在网上进行交互式的调查，这样研究得出的用户上网规律，可以用来指导网络建设。

3.3 网络信息检索

在网络的海量信息面前，人们通常借助于搜索引擎来检索信息。虽然搜索引擎越来越多，功能也越来越完善，但查准率、网页的可到达性等并不尽如人意，往往存在许多空链接，这主要归咎于数据采集和组织不够完善。这个问题可以借助于网络计量学的分析结果进行改善。譬如，收集Web页面时可以直接抓回权威的网页。收集权威网页是一个原始评价过程，即先给出一组源网页，然后按深度优先或广度优先的算法进行遍历，再利用链接分析技术来计算出权威页：网页的质量是由指向它的页面的数量决定的；把收集回来的网页看做一个有向图，然后采用递归定义法就可以算出权威页[11]。此外，在对所收集的网页进行加工和入库时，要根据网络计量学所需的计量指标来挖掘潜在信息源，从而提高查准率和查全率。

3.4 数字图书馆资源管理

数字图书馆赖以生存的基础是数字化、网络化的资源，网络数据库和网络信息资源已逐渐成为数字化资源的主体。网络资源的激增一方面满足了人们的信息需求，另一方面又给数字图书馆的建设和管理带来了困难。网络计量学的研究对象、方法、内容体系都符合这一要求，能对数字图书馆的资源管理进行科学指导，并提供定量的依据，提高数字图书馆的管理水平。数字图书馆既要进行动态馆藏的维护，在探明信息资源数量特征的基础上判断其价值和实效性，又要对各类信息源的分布进行定量分析评价，确定出核心信息来源，有效地指导信息收集，充分满足用户的需求[12]。

4 结语

网络计量学研究尚处于试验和探索阶段，理论体系也尚未成形。将文献计量学的方法应用于对网络的研究，尽管产生了一些积极结果，但网络研究中的实际操作并没有人们预期的那样好。如何在理论和方法上进一步探索，是今后亟待解决的问题。

参考文献

1 T. C. Almind, P. Ingwersen. Informetric analyses on the World Wide Web: methodological approaches to "webmetrics". Journal of Documentation. 1997,54

2 徐久龄,刘春茂,刘亚轩. 网络计量学的研究. 情报学进展1998-1999年评论,第3卷. 北京:航空工业出版社,1999

3 邱均平,陈敬全. 网络信息计量学及其应用研究. 情报理论与实践. 2001,24(3)

4 T. O' Neill. Characteristics of web accessible Information. http://www.ifla.org/IV/ifla63/63onee.htm

5 L. Björnebom, P. Ingwersen, Perspective of Webometrics. Scientometrics, 2001,50

6 R. Rousseau. Sitation: an exploratory study. Cybermetrics. 1997,1(1)
http://www.cindoc.csic.es/cybermetrics/

7 M. Thelwall. Web impact factors and search engine coverage. Journal of Documentation. 2000,56

8 P. Ingwersen. The calculation of web impact factors. Journal of Documentation. 1998,54(2)

9 An introduction to the webmetrics: The national science digital library. http://webmetrics.comm.nsdl.org/strupp_intro.pdf

10 A. Broder. Graph structure in the web. http://www.almaden.ibm.com/webfountain/resources/GraphStructureintheWeb.pdf

11 J. Kleinberg. Authoritative sources in a hyperlinked environment. http://www.cs.cornell.edu/home/kleinber/auth.pdf

12 S. Jones. A transaction log analysis of a digital library. http://www.cs.waikato.ac.nz/~stevej/Research/PAPERS/ijodllogs.pdf

王知津 南开大学国际商学院教授、博士生导师。通信地址:天津。邮编300071。

郑红军 张收棉 南开大学国际商学院情报学硕士研究生。通信地址同上。

(来稿时间:2004-12-30)