●邱均平 张 洋 赵蓉英

网络信息计量学方法论*

摘 要 系统全面地收集研究所需要的原始数据是开展网络信息计量研究的前提。数据收集方法的研究是网络信息计量学方法研究的重点问题和难点问题。网络信息计量学常用的数据分析方法有网络链接分析法、网络内容分析法和网络数据挖掘法。网络信息计量学的研究方法可分为 3 个层次:哲学方法、一般科学方法和特殊研究方法。参考文献 24。

关键词 网络信息计量学 方法论 数据分析 数据收集
分类号 G350

ABSTRACT Systematically and comprehensively collecting original data for researches is the premise for the studies in webometrics. The study on the methods of data acquisition is the key and one of the hard issues in webometrics. In this paper, the authors introduce some common data analysis methods in webometrics, such as network link analysis, network content analysis and network data mining. There are three levels for the research methods in webometrics: philosophical method, general scientific method and special research method. 24 refs.

KEY WORDS Webometrics. Methodology. Data analysis. Data acquisition.
CLASS NUMBER G350

1 数据收集方法

系统全面地收集研究所需要的原始数据是开展网络信息计量研究的前提,数据收集过程是为数据分析阶段准备原始数据的过程。这里的原始数据是指有关网络信息的数量描述,它的表现形式是一系列的数据整合,反映的是具体研究对象的数量特征。

1.1 网络数据源

目前用于网络信息计量研究的原始数据主要有两种来源:(1)结构化或半结构化的数据资源,主要是连接到互联网上的各种专用数据库,包括各种联机信息系统、引文索引、全文数据库、专题网站等。它们一般经过加工处理,属于高度组织化的信息资源,并配备有专用的信息检索工具。但专用于或者可直接用于网络信息计量研究的数据库还不多见。因此,对其中蕴涵的信息规律进行挖掘研究仍然是网络信息计量学的重要研究内容之一。这些数据库所属机构定期发布的统计资料或年鉴,同样可用于网络信息计量研究。(2)非结构化数据。网上拥有大量的自然语言文本、图像、声音等数据,它们无法用统一的结构来表示,被称为"非结构化数据"。这些数据里隐含着许多非常有价值的信息,对其进行开发利用是网络信息计量学的重要内容。非结构化数据数量巨大,形式复杂,现阶段的网络信息计量研究主要依赖"搜索引擎"来获取。此外,近年来出现的一些专门的网络信息统计网站,它们提供的统计数据类似于传统文献计量学中的"二次文献",可以作为开展网络信息计量研究的数据来源,例如 Alexa[1]。

1.2 网络数据收集工具

无论对于何种类型的科学研究,数据收集方式都可大致分为人工、计算机辅助两种。但在网络环境下,纯粹的人工收集既不经济也不可行,充分利用计算机技术和工具是进行海量数据收集的必然选择。"人机结合"的数据收集方法是目前网络信息计量研究中收集数据的主要方法。就目前所使用的计算机辅助工具来说,无论种类多么繁多,都可大致分为专用工具和通用工具两种。专用工具是针对特定的研究对象和研究目的而开发专门的数据收集工具,以实现数据的自动收集和筛选工作。例如,Alastair G Smith 和 Mike Thelwall[2]在研究中就使用了自己设计的爬行器。专用工具虽然有量身定做的优势,但开发周期过长,投入过多,技术门槛过高,有很大的局限性。通用工具则是指可以广泛地应用于多个领域的计算机软件或系统,成本较低,使用容易,成为研究者的主要选择,最典型的例子就是搜索引擎。

随着网络的不断扩展和功能的不断增强,对大部分用户来说,搜索引擎已经成为其检索信息的主要工

* 本文系国家自然科学基金资助项目"网上学术信息的分布与变化规律研究及其应用"(70673071)的成果之一。

具。网络信息计量学既然以网络信息为研究对象，自然离不开搜索引擎。事实上，最早从 T. C. Almind 和 Peter Ingwersen[3] 所作的研究开始，一直到今天，相当多的网络信息计量学研究者都主要依靠搜索引擎来收集数据。所用到的搜索引擎也是种类繁多，特点各异，包括 AltaVista、AllTheWeb、Northernlight、Google、Excite、Lycos、HotBot、Infoseek 等等在内的众多搜索引擎都曾被应用于网络信息计量研究的数据搜集工作中[4]。其中，AltaVista[5] 由于检索功能强，检索途径多，能满足多种计量需要等优点，受到研究者青睐。事实上，迄今为止的网络信息计量学研究几乎都使用 AltaVista 来搜集研究数据。近年来，越来越多的搜索引擎也提供了强大的检索功能，覆盖范围更广，研究者们有了更多选择。可以说，没有搜索引擎，网络信息计量学就失去了有效的研究手段，不可能得到迅速发展，搜索引擎无疑是网络信息计量学研究最重要的数据收集工具。虽然搜索引擎在当前的网络信息计量学研究中具有如此重要的地位，但是在实际应用中，它们也表现出了查全率低、使用不便、效率低下、功能不足、稳定性差、缺乏客观性等许多局限，检索效果一直受到质疑，直接影响到研究结果的可靠性和合理性。因此，目前搜索引擎难以满足网络信息计量研究的需要，研究者们亟待功能更为强大的数据收集工具。

结合通用工具和专用工具两者的优势，在通用性的基础上加强专用性，将是未来数据收集工具发展的主要趋势。目前网络上已经开始出现一些专业搜索引擎，它们将检索范围限定在一定的专业领域内，加强了结果的有效性，有些还依托特定的数据库，查全率得到大幅度提高，对于某一领域的特定研究工作是很有效的工具。另外还出现了一些新的检索工具。例如，Bright Planet 公司开发的 DQM（Deep Query Manager）平台就是一个集信息发现、采集、管理和分析于一体的深层网络信息查询平台，不仅可以对位于“深层网络”数据库进行信息查询，还可同时对网络上成百上千个搜索引擎、目录索引和联网数据库中的信息进行自动采集[6]。但是这些工具的开发都还在起步阶段，作用有限，人们在相当长的时间内仍不得不以通用的商业搜索引擎为主要的网络数据收集工具。

2 数据分析方法

目前应用到网络信息计量学中的数据分析方法很多，其理论依据、来源出处、程序步骤、应用条件和使用范围都有很大的不同，而且大都源自于其他学科领域，在网络信息计量学中的应用尚处于摸索、起步阶段，尚未形成较为成熟的研究过程和实施步骤。但可以肯定的是，这些方法最终都将被纳入网络信息计量学的方法体系，成为真正意义上的网络信息计量学特征方法。下面简要介绍最常用的几种方法。

2.1 网络链接分析法

自 1997 年 T. C. Almind 和 Peter Ingwersen 首次提出“Webometrics”概念，用以描述文献计量学方法在网络信息计量研究中的应用以来，研究者们从不同角度将文献计量学的研究思想和特征方法应用于网络信息的计量研究中，取得了许多重要成果。其中，影响最大、成果最多、应用最为普遍的就是源自文献计量学引文分析法（Citation Analysis）的网络链接分析法（HyperlinkAnalysis）。

对于绝大多数 Web 页面来说，只有通过与其他的网页及其自身内容的链接，才能相互交换信息，扩大使用价值。目前互联网上的 Web 网页主要是利用超文本标记语言编制起来并利用超链接建立联系的一种信息组织方式。网页的不同链接体现了不同的信息功能，具有不同的特征和规律。网络链接与科学文献引文之间天然的相似性使文献计量学家们找到了文献计量和网络的契合点，他们创造性地将 Web 网页链接与传统文献中的“引用”联系起来，将文献计量学中引文分析法应用于网络信息计量研究中，由此产生了网络信息计量学的重要研究方法——网络链接分析法。1996 年，美国爱荷华州立大学图书馆的理论馆员 Gerry McKiernan 根据文献计量学中引文的含义，首次提出了“Sitation”的概念，来描述网站（Site）之间相互链接的行为，他指出：Cited Sites = Sitation，即所谓“Sitation”就是被引用的站点[7]。此后 Isidro F. Aguillo 在 1996 年比勒菲尔德召开的 4S/EASST 会议上引用了这一概念[8]。1997 年 Ronald Rousseau 发表的一篇论文中[9]，sitation 一词首次正式出现在文献题名当中，标志着网络链接分析法被学术界正式认可为一种真正意义上的科学研究方法。

网络链接分析法是网络信息计量学的重要内容和研究方法，它的产生和迅速发展普及，极大地促进了网络信息计量学的发展。目前，有关网络链接分析的研究主要包括网络链接分布规律、同链聚类分析、网络影响因子、网络链接分析工具等几个方面的内容[10]。近年来，网络链接分析法已被成功应用于分析和评价网站的质量，评价期刊和大学、图书馆等学术机构的权威性，指导网络资源的组织建设，指导网络资源检索和

开发利用,分析和掌握学科发展状况,开发和应用智能超文本链接,企业管理和情报分析,评价社会科学方面的成果等众多的领域。但在另一方面,网络链接作为一种新的信息组织方式,它在有很多优点的同时,也有容易迷失方向、漏掉一些信息内容、认知负担过重、使用效率较低等缺点[11]。无论在外部特征还是内在机理上,网络链接都要比文献引用复杂得多,这使得源自引文分析法的网络链接分析法有许多先天的缺陷和问题。此外,还有许多不利因素,包括网络信息覆盖范围广、动态性强,信息量巨大、不确定性、缺乏合理组织、难以预测等特点,信息过载和信息污染现象严重,引用与被引用关系也变得十分复杂,网站的被链接数量还与它的商业推广有着密切的联系,缺乏客观性,商业搜索引擎作为目前主要的数据搜集工具,其公正性和可靠性都深受质疑等等,使得网络链接分析法作为一种科学研究方法有一定的局限性[12]。

网络链接分析法源自引文分析法,但决不仅仅是引文分析法在网络上的简单应用和推广。我们在进行网络链接分析研究时,一方面可以将引文分析法作为研究起点,为我们提供新思路和可参考的经验。另一方面我们要充分意识到引文分析与链接分析的差异性,避免盲地机械地将引文分析法直接应用到网络上。正如 Ronald Rousseau 所言,"由于网上信息往往不是科学论文,而且链接要比一般文章多得多,所以其复杂性也大,这也是此项研究的必要性和开创性所在"[13]。

2.2 网络内容分析法

内容分析法(Content Analysis)是自第二次世界大战时期逐步发展形成的一种新兴的社会科学研究方法,目前被广泛地运用于新闻学、传播学、社会学、政治学、图书情报学、教育学、经济学、人类学、心理认知科学等广泛的社会科学领域。它产生于新闻传播领域,发展又受到新闻传播研究的有力推动,但其应用范围决不仅限于新闻传播领域。由于大众媒体传播的内容不仅是社会信息的重要组成部分,也是情报学的主要研究对象,来自传播学的内容分析法的研究观点在情报学研究中自然也同样适用。因此,早在 20 世纪 60 年代末,西方图书情报学领域的学者就在研究中引入了内容分析法。经过多年的发展,内容分析法已经逐步被纳入了情报学的方法体系,被广泛应用于情报研究工作,指导情报学规律的摸索和理论的建设[14]。

20 世纪 90 年代末,网络作为"第四媒体"得到迅速发展,网络传播已经成为包括新闻传媒领域在内的社会各界关注和研究的焦点,内容分析法也随之进入一个新的发展阶段,从而产生一种新的研究方法——网络内容分析法(Network Content Analysis),也称为基于网络的内容分析法。它有两层含义:一是对网络信息内容进行分析,二是基于网络技术和网络环境来研究内容分析法。它具有传统内容分析法的基本特征,但又不是内容分析法应用范围的简单扩大。它缩小了定性分析和定量分析的差距,提高了内容分析的效率,扩展了内容分析的层次[15]。网络内容分析法有十分重要的价值和美好的应用前景,越来越多的学者对其开展积极研究。在网络信息计量学领域,将内容分析方法应用到网络空间,对网络信息的内容特征及其变化进行定量分析和深入揭示,同样成为一个重要的主题。例如,Bar-Ilan[16] 曾选择"Informetrics"为主题在 6 个主要搜索引擎中进行检索,然后采用内容分析法分类了近 800 个独立的网页,结果发现约 40% 的网页可以粗略地归类为参考书目,至于其他类别尚有国际科学与信息计量学会、虚拟图书馆以及网络信息计量学,此外,他还将网络环境下的问题跟踪模式与文献数据库和引文数据库的模式作了比较。张红燕则通过对我国 10 个图书馆的网站进行内容分析和分类,指出其存在的问题,并对网上图书馆的建设提出了一些建议[17]。

总之,网络内容分析法作为一种计量和揭示网络信息的数量特征和内在规律的有效方法,已成为网络信息计量学的重要研究内容,将在网络信息计量学中发挥越来越重要的作用。网络信息计量学最终将把网络内容分析法完全纳入到自己的方法体系当中,成为本学科的有力工具。

2.3 网络数据挖掘方法

"数据挖掘"就是从大量的数据中提取知识[18]。它是信息技术尤其是数据库技术自然演化的结果。近年来,数据挖掘无论在理论上还是实用技术上都取得了长足进步。原则上讲,数据挖掘可以在任何类型的信息存储上进行。数据挖掘既可用于各种基于不同数据模型的数据库,也可用于数据对象通过超链接集合在一起所构成的复杂巨系统——Web,即网络数据挖掘。网络数据挖掘是指从与网络相关的资源和行为中抽取用户感兴趣的、有用的模式和信息。它针对包括网页内容、页面之间的结构、用户使用记录、电子商务信息等在内的各种网络数据,从中可以发现有用的知识来满足人们从网络中提取知识、改进站点设计、更好地开展电子商务等需求[19]。从研究现状来看,它涉及的分析处理对象有很多不同类型,由此可

大体上分为网络内容挖掘、网络结构挖掘、网络使用记录挖掘三个大类，每个大类还可进一步细分为若干小类[20]。

我们认为，网络数据挖掘属于网络信息计量研究的范畴，它是网络信息计量学的特征研究方法。因为，首先，两者的研究对象是一致的。数据是按照一定规则排列组合的记录信息的物理符号。它的原意是以数字形式表达的信息，数据库就是以数字形式表达的信息的集合。此外，还存在着大量非数据的信息，如模拟信息、文本信息、语音信息、图像信息、图形信息等。在计算机领域，文本信息、语音信息、图像信息、图形信息也被习惯性地称为文本数据、语音数据、图像数据、图形数据。随着计算机的应用日益普遍，这一领域的用语习惯也随之流传开来。在这种情况下，数据的概念便已经泛化了，再去区分"数据"与"信息"就失去了实际的意义[21]。从这个角度来看，网络数据挖掘的处理对象网络数据的实际含义与网络信息没有本质的区别，网络数据挖掘实际上就是网络信息挖掘。其次，两者的目的是一致的。与数据挖掘相似，我们可以将网络数据挖掘看做是从网络上发现具有潜在价值的或事先未知信息的全过程，显然，这一过程的目的是揭示网络信息数量特征和内在规律，而这也正是网络信息计量学研究的最终目的。

3 网络信息计量学的方法论体系

网络信息计量学方法论属于个别学科的专门方法论，是网络信息计量学在探索其研究对象时所采用的理论、方法和技巧的总和。网络信息计量学方法论是联系网络信息计量学理论与实际应用的桥梁，它既是情报学理论体系的组成部分，也是科学方法论体系的组成部分，对于学科建设有十分重大的意义。注重方法论的研究，将网络信息计量学方法的研究上升为方法的理论高度并建立网络信息计量学方法论体系，必将对网络信息计量学的发展产生积极的推动作用。

科学研究方法论告诉我们，科学研究方法按照普遍性程度和适用范围可以划分为哲学方法、一般研究方法、特征研究方法等3个层次，层次越高，方法的概括和抽象程度就越大，其适用范围就越广。网络信息计量学的研究方法同样可从这3个层次来考察。

哲学是理论化、系统化的世界观，是自然、社会、思维科学的总结和概括，哲学方法论是为所有科学研究提供研究方法的总原则，具有最广泛的普适性。因此，哲学方法是具有最大普遍性和最高概括性从而具有世界观意义的最一般方法，适用于一切领域。哲学方法具有高度的抽象性、强烈的间接性和深刻的本质性，具有认识功能、启迪功能、批判功能和预见功能。马克思主义哲学倡导和坚持科学实践观原则、唯物辩证法原则和历史唯物主义原则，主张并实践着从实践的、唯物辩证的和历史唯物主义的视角去理解和变革现实，促成现实合乎规律的发展。自觉地以马克思主义哲学作为科学研究的世界观和方法论的指导，不仅有助于科学研究的成功，而且必将促进科学研究的健康发展[22]。马克思主义哲学的核心——唯物辩证法，是认识与改造世界的方法，具有丰富而深刻的方法论意义，因此，哲学方法的运用，也可以说是唯物辩证法的运用。网络信息计量学只有在辩证唯物主义思想的指导下，才能把学科建立在可靠、严密、科学的基础之上。

所谓"一般科学方法"是指在所有科学活动中具有普遍适用意义的科学方法。它是将各个学科研究方法的共性抽出来而形成的泛化方法，具有横断性与综合性特征，包括数学方法、逻辑方法、系统方法、控制方法、信息方法等应用于各个领域的一般方法。与其他许多学科和研究领域相比，数学方法、系统方法和信息方法的广泛应用，是网络信息计量学研究的显著特征。首先，网络信息计量学作为一门揭示网络信息数量特征和内在规律的学科，定量描述和统计分析是其基本的研究手段，数学无疑是展开定量研究的最重要的研究方法。其次，网络作为网络信息的载体是一个复杂宏观系统，为探索网络信息的特征和规律，只有运用系统分析的思想才能将其从复杂的体系中剥离出来，并兼顾宏观体系的影响。第三，网络信息计量学以网络信息作为研究对象，属于信息科学领域的分支学科，必然强调用信息的观点和方法来认识、综合和变革高级复杂系统，只有这样才能有效地解决问题，这就意味着信息方法是开展网络信息计量研究的不可或缺的研究方法。

网络信息计量学的特征性研究方法是具有网络信息计量学学科意义的特定方法，是应用在网络信息计量学这一门学科内部的个别方法。从总体上看，网络信息计量学的特征研究方法有3个来源：首先，改造上位学科的特征研究方法。网络信息计量学以"三计学"作为其学科基础，从某种意义上来说，网络信息计量学就是"三计学"在网络上应用的一门学科，因而在"三计学"中得到广泛应用的文献信息统计分析法、数学模型分析法、引文分析法、书目分析法、系统分析法、关键词统计分析法、（下转第41页）

11　麦敏华.公共图书馆与义工组织合作运作模式的探索.图书馆论坛,2004(4)
12　孙孝诗.义工服务于图书馆而引发的思考.河南图书馆学刊,2004(2)
13　刘彦方.试论图书馆义工的引入.图书馆杂志,2002(9)
14　林坤明.图书情缘——谈谈我从义务图书馆员到馆长之路.图书馆论坛,2000(6)
15　禹莲辉."义务图书管理员"与高职院校图书馆管理.科技情报开发与经济,2006(9)

韩　芸　中国青年政治学院社会工作学院讲师,博士后。通讯地址:北京海淀区西三环北路25号,中国青年政治学院社会工作学院。邮编:100089。

(来稿时间:2007-05-23)

(上接第32页)关联数据分析法(包括聚类分析、共词分析、同域分析等)、计算机辅助文献信息计量分析法等特征研究方法都能应用到网络信息计量研究中[23]。当然,由于网络环境的特殊性,研究者们在应用过程中对上述方法进行了不断的调整和变化,使其最终成为网络信息计量学的专用方法。其次,移植其他学科领域的特征研究方法。网络信息计量学作为"网络技术、统计学、文献计量学理论三合一的产物"[24],涉及计算机、人工智能、拓扑学、社会学、图论等众多学科和研究领域,来自这些学科领域的研究方法和技术手段丰富了网络信息计量学方法体系,促进了网络信息计量研究工作的发展。例如,新闻传播领域的内容分析法、数据库技术领域的数据挖掘方法、理论物理学领域的复杂网络理论、计算机科学领域的Web信息检索算法、社会学中的社会网络分析方法等都已经应用到网络信息计量研究中。第三,移植一般研究方法。现代科学技术的一个重要特点是相互转移和渗透,某些基本原理和思维技巧,是大多数类型的科学研究所共同适用的。将网络信息计量学研究工作的特殊性与一般方法相结合,把一般研究方法科学地、合理地、创造性地加以移植,是建立网络信息计量学特征方法的重要手段之一。例如数学中的图论方法和拓扑学方法就已经被应用到网络信息计量研究。

参考文献

1　http://www.alexa.com
2　Smith A G, Thelwall M. Web impact factors and university research links. In: Proceedings of the 8thInternational Conference on Scientometrics and Informetrics, Sydney, Australia, Jul 16~21, 2001, 657~664.
3　Almind T C, Ingwersen P. Informetrics Analyses on the World Wide Web: Methodological Approachesto "Webometrics". Journal of Documentation, 1997 53(4): 404~426.
4,20,23　邱均平,张洋.网络信息计量学综述.高校图书馆工作,2005(1)
5　http://www.altavista.com.
6　http://www.brightlanet.com/products/dqm_benefits.asp/
7　McKiernan G CitedSites(sm): Citation Indexing of Web Resources. [2005-04-09]. http://www.public.iastate edu/~CYBERSTACKS/Cited.htm.
8　刘东贤.信息计量学的新进展:从Webometrics谈起.情报杂志,2002(10)
9,13　Rousseau R. Sitations: an exploratory study. Cybermetrics: International Journal of Sciemometrics, Informetrics and Bibliometrics, 1997(1)
10,12　张洋,邱均平,文庭孝.网络链接分析研究进展.图书情报知识,2004(6)
11　邱均平,黄晓斌.WWW网页的链接分析及其意义.中国图书馆学报,2002(6)
14　邹菲.内容分析法的理论与实践研究.[硕士论文].武汉大学,2004
15　周黎明,邱均平.基于网络的内容分析法.情报学报,2005,24(5)
16　Bar-Ilan J. The Web as information source on Informetrics? A content analysis. Journal of the American Society for Information Science, 2000, 51 (5)
17　张红燕.对我国图书馆万维网站点内容的分析.图书情报工作,1998(11)
18　Jiawei Han, Micheline Kamber 著;范明,孟小峰等译.数据挖掘概念与技术.北京:机械工业出版社,2001:3
19　邱均平,黄晓斌,段宇峰,陈敬全.网络数据分析.北京:北京大学出版社,2004:180
21　钟义信.信息科学原理(3版).北京:北京邮电大学出版社,2002:58
22,23　王晖.科学研究方法论.上海:上海财经大学出版社,2004
24　徐久龄,刘春茂,刘亚轩.网络计量学的研究.见:张力治.情报学进展(第3卷).北京:航空工业出版社,1999

邱均平　武汉大学信息资源研究中心,教授。通讯地址:武汉。邮编430072。
张　洋　中山大学资讯管理系工作。通讯地址:广州。邮编510275。
赵蓉英　武汉大学信息资源研究中心工作。通讯地址:武汉。邮编430072。

(来稿时间:2007-02-25)