●邱均平 黄晓斌

WWW 网页的链接分析及其意义*

摘 要 WWW 网页链接分析的内容主要包括:链接和被链接量、链接网页的类型、链接的频次和变化、链接网页之间的关系和网络电子图书、期刊引证分析。WWW 网页链接分析为计量学开辟了新的研究和应用领域,对网页链接进行分析,可以使研究结果更加科学化和精确化。参考文献 10。

关键词 WWW 网页 超文本 超文本链接 网络信息计量学
分类号 G250.72

ABSTRACT The main contents of the analysis of links on WWW pages include the numbers, types, frequencies and citation of links. The analysis of links on WWW pages has opened a brand new research and application area for informetrics. 10 refs.

KEY WORDS WWW pages. Hypertext. Hypertext links. Network informetrics.
CLASS NUMBER G250.72

WWW 网页是利用超文本标记语言(HTML)编制起来并利用超文本链接而建立联系的一种信息组织方式。WWW 网页只有通过与其他的网页及其自身内容的链接,网页才能相互交换信息,扩大使用价值。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。对 WWW 网页链接进行分析是网络信息计量学一项重要的研究内容。

1 WWW 网页链接的结构和类型

目前 WWW 网页主要是采用超文本的组织方式,由许多不同信息节点和链组成。节点分链源和链宿,链源是链的开端,链宿是链的目标,它们是链形成的基础。链是特定节点之间的信息联系,它以某种形式将一个节点和其他节点联系起来。

1.1 网页链接的结构

(1)节点(node)。节点是围绕某一特殊的主题组织起来的数据信息单元,它是一种可以激活的材料。节点大小因主题不同而异。节点有许多类型:根据媒体划分,可分为文字、图片、音频、动态图像、程序、混合型等,不同的媒体有不同的属性和表现方法;根据结构分,可分为原子节点、复合节点、包含节点;根据动态分,有活动节点和静态节点;根据规范化程度分,可分成结构化、半结构化和非结构化。

(2)热标(hotspot)。热标是取得信息关联的链源,通过它激发链而引起向相关内容的转移。不同的媒体有不同的形式:如热字(hot-word),是文本中有特殊的涵义或需要进一步解释的字、词;热区(hot area),在图像的显示区指明一个敏感的区域,作为触发转移的链源;热点(hot point),在时基类的媒体如声音、视频等在时间点上的触发源;热元(hot-element),在图形媒体中,图元(例如一条线)是基本的单位,为了使这些独立的图形单位能够作为信息转移的链源,因此引入热元的概念;宏节点,是指组合在一起的许多节点群统一作为热标。

(3)链(link)。链是网页表现信息之间联系的实体,是将不同的节点联系起来的工具。链具有方向性,两个节点之间的链接具有单向和双向之分。单向是链源和链宿不可互换的关系;双向链是一种可互换的关系,链源和链宿可以互换。网页链接方式有:从一个节点到另一个节点;从一个节点到另外节点的内部;从一个节点的内部到另一个节点的内部,以及同一个节点的不同单元之间互相链接。链有不同的组合:一进一出,每个链只有一个链指向它,并且只有一个链从它出发;一进多出,每个节点只有一个链指向它的链,但它有多个从它出发的链;多进多出,每个节点有多个链指向它的链,也有多个从它出

*本文系教育部十五规划项目"网络信息计量学研究"(批准号:01JA870009)的成果之一。

发的链;网络链,节点之间由盘根错节的各类型链相互链接而组成链的网络。

1.2　网页链接的主要类型

根据功能和属性,网络链接的类型有:

(1)导航链:将相关的信息组织起来,引导用户参考相关的信息。导航链的类型较多,用途比较广泛。其中主要有[1]:目次链;注释链;实例链;索引链;扩展链(不同媒体的进一步扩展,如把有关"贝多芬"的生平介绍的文字链接到贝多芬的照片或音乐作品);相关链(在内容上关系比较密切,链接起来便于参考。如把有关"网络信息计量学"的网页链接到有关"文献计量学"、"信息计量学"的网页);应用链(链接不同的工具,便于利用);等价链(两个概念或信息的内容基本一致,但组织在不同网页或同一网页不同的地方,通过等价链把它们联系起来。如"武汉大学"链接到"武大");引用链(链接被引用的原文或其他内容);评价链(链接有关的评论、介绍文章或有关的信息);分解链(从整体链接到子成分);聚合链(从子成分链接到整体);版本链(不同版本之间的链接)。

(2)执行链:通过热标和应用程序相连,可以激发一个操作的执行。如通过点击 E-mail 的热标,便可打开和运行 E-mail 程序。

(3)类型链:有些系统允许用户描述两个节点的关系,可以定义链的类型。这种类型链必须有一个独立的数据实体来描述。

(4)推理链:通过智能化的链和节点,引入计算方式,可在多个目标中动态地确定目标和表现方式。例如,"把网页 x 链接到包括支持 x 条件的网页上","条件存在,网页 a 链接到 x,否则,链接到 y 网页上"。推理链主要包括有:is-a 链,指明对象节点中某类成员;has-a 链,用于描述节点具有的属性;蕴涵链,用于链接推理树中的事实,它相当于正在触发或已经触发的规则。

(5)自动链:允许系统把当前的节点与相似主题的节点或满足条件的其他节点自动连接在一起。如可以在文本文件中搜索关键词和所在的位置,或通过基于内容的检索确定某些特征,或通过通信协议和其他服务器中的内容建立联系等。

网页的链接类型存在如下几种关系:自我链接,指网页自身的链接。同被链接,指两个网页共同被另外的一个或一个以上的网页所链接,以链接它们的数量的多少可以测定同被链接强度。链接耦合,两个网页如果同时链接一个或多个网页,则称它们在链接上是耦合的。

2　WWW 网页超文本链接的功能和作用

WWW 采用客户机/服务器的体系结构,通过超文本技术,将许多网页链接起来,提供给用户利用。超文本系统一般分为三个层次:①表现层,即用户接口,由运行在用户计算机上的客户浏览程序管理。②抽象机器层:存储节点和链,服务器提供客户的数据采用超文本标注语言,网络采用的通信协议标准是超文本协议(HTTP)。③信息库层:由因特网上的各种服务器组成,负责提供各种各样的信息资源。WWW 的客户软件(Web 浏览器)在用户端提供统一管理各种媒体的界面,负责向服务器提出请求,解释和定位资源,利用统一资源定位器管理有关信息资源。

网页超文本链接为信息的组织和利用提供了便利途径。其优点是:

(1)界面友好。网络环境下,由于异构数据库较多,存在有不同的检索命令方式,用户难以掌握。而超文本系统采用窗口形式,利用各种图标来描述选项,用户只需要点击图标就可以执行各种选择,简单易学,人性化强,方便利用。

(2)兼容性强。因特网信息资源中存在不同的文件格式和媒体类型,网页能连接和处理不同的格式和媒体。

(3)扩充性强。超文本的节点和链可以动态地改变,各节点中的信息可以随时更新,网页的补充不受限制,可以通过新的链接来反映新的关系,只要按照超文本协议就能连接起来和互相利用。

(4)交互性强。采取人机对话,用户自主性较大,可自由地选择链接方式。

(5)灵活性强。链接把信息知识有机地联系在一起,用户可以顺着这些链接寻找自己需要的信息知识。这种链接可以是封闭的,也可以是开放的。超文本的链接使它不局限于前后翻页式的线性操作,这种非线性特征使它变得更为灵活。随机通达信息是超文本的又一特点,由于有了信息知识点之间的超链接,同一信息可以由不同的路径得到,这种接近信息的方式不是先前决定的,它近似于随机,用户可以根据自己对信息知识点的掌握情况选择通达的道路。

(6)可以追踪有关的参考资料。通过链接找到有关资料的网页,然后根据新的网页线索,进一步链接其他网页,获得更多相关信息。

(7)能深入到信息知识的单元片段,进行更深层

次的检索。超文本检索不仅可以找到整篇网页的内容,也可检索到网页中的个别信息单元。而传统的信息检索通常以题录、摘要或全篇文章为主。

(8)联想式检索符合人们的思维习惯。人们在思考和阅读时,有时往往是跳跃的、发散的,网页超文本的链接方式适应了这种特点。

网页的超文本链接也存在一些缺陷和问题。主要是:①容易迷失方向(disorientation),具体表现为有时用户不知道身在何处,较难返回原处和到达相应的方向,尤其是在信息空间太大,使用者不了解导航设施时。②容易漏掉一些信息内容。非线性的顺序使用户误入歧途,把一些重要的信息错过。③认知负担过重。由于网页链接错综复杂,用户容易疲劳。④设计时花费的工夫要比组织线性文本多。⑤效率较低。浏览链接的网页往往要花较多的时间,而且不如线性文本系统性强。为了解决这些问题,新的浏览器往往采取了一些导航措施,如提供浏览的历史清单、书签、可视化组织器、跳转、索引、在线帮助、搜索引擎、分层导航等。美国斯坦福大学研究使用一种簇化技术使超文本中某些具有共同特征的节点归结为一个节点的集合,这种从原始的信息系统提取出一种高层次的结构做法,可减少链的数量,提高使用效率。

3　WWW 网页链接分析研究的主要内容和意义

网页链接分析的内容主要包括:链接和被链接量;链接网页的类型;链接的频次和变化;链接网页之间的关系;网络电子图书和期刊引证分析。

对 WWW 网页链接进行研究,具有重要意义。

(1)为文献计量学开辟了新的研究和应用领域。许多学者利用文献计量学的一些原理去分析网络电子出版物的有关情况,取得较好效果。1996 年,Mckiernan 根据文献计量学引文(Citation)含义,提出"Sitation"的概念,意思是对网页(Website)的引用行为分析,进一步扩展文献计量学的研究范围。Website. net 仿照《科学引文索引》的做法,编制了一个"网页引用分析工具"(Web Citation Indexs, WCI),可以用来统计分析网页引用情况,研究网页链接之间的关系和规律,监视网页链接的变化情况等。它还提供一个叫"Citeseer"的自动引文索引系统,可以用来查找和了解网页的引用和被引用情况,评价网页、网络杂志、有关作者以及有关研究课题的情况[2]。1998 年,Ingwersen 提出可以把文献计量学的期刊影响因子应用到网页的评价中去[3]。网页的影响因子(Web impact factor,WIF)是指某一类型的域名或网页被链接之和与有关域名或网页之和的比例。网页影响因子可以用来分析在一定的时期内相对关注的网页情况。应当指出的是,网页的链接机制与文献计量学研究的引文机制有许多相似之处,但也存在一些区别。与传统的引文分析方式相比,网页链接分析的数据已经数字化,并利用计算机进行,可以自动操作,交互性强,并能对有关的数据进行多方面的深度分析。WWW 网页链接分析范围更广,除了引证分析外,还包括参考、应用、相关等,有时甚至是一些意义不太大的广告;链接涉及的载体类型多,包括文本、声音、图像、动画等;网页变化大,链接的动态性强,常常处于不断的变化中。链接数量多,数据量大。这些特点使得网页的链接分析带来许多新的问题,值得今后进一步研究和探讨。

(2)WWW 网页链接分析是网络信息计量学的一项重要的内容。网络信息计量学是新出现的一个研究领域。它采用现代技术手段,利用定量分析法,专门对网络的资源配置、组织、利用等问题进行定量的研究分析,以揭示网络信息的数量特征和内在规律[4]。研究的主要内容包括网络资源的分布规律、更新老化规律、传输和利用的效率、用户行为和规律、成本与效益规律等问题。目前国外利用定量方法分析网页链接主要有两种算法:一种是 Brin 和 Page 提出的"Pagerank"算法,另一种是 Kleinberg 提出的"HITS"算法。两者都是根据一个网页链接其他网页的数量和质量来判断一个网页的质量和权威性。"Pagerank"直接采用链接模式的矩阵。其基本思想是:一个网页被多次链接,则这个网页可能是重要的;一个网页虽然没有被多次链接,但被一个重要的网页链接,则这个网页也可能是重要的;一个网页的重要性被均分并被传递到它所链接的网页上。而"HITS"则增加了一个转换的矩阵,既分析某一网页被多个网页链接的"权威性"质量,又分析网页链接其他网页的"集中性"情况。如一个网页本身并不突出,很少被链接,但它提供了最为突出的链源。通常"集中性"好的网页是指向"权威性"好的网页,结合起来可以发现有关的结构规律。当然,这种算法的不足之处是都是静态的,单纯用已经存在的链接模式去分析,显然不能适应 Web 不断变化的要求。如何对网页链接进行动态的分析仍需要进一步探讨。

(3)分析和评价网页的质量。点击率是目前判断网页的访问次数的一个指标,受欢迎的网页点击率一般较高,其链接率也较多。利用链接的有效性可

判断网络的生命力,如发现某个网页链接经常出现死链,说明其已经修改或删除。分析链接设计科学与否,可以用来评价使用效率。还可利用 ZIPF 曲线分析网页的受欢迎程度[5],利用 WIF 评价网页的权威性。

(4)有利于网络资源的组织建设。通过对节点和链路的统计分析,有助优化网页的链接设计,减少不合理的链接,如悬空链或死链。通过语义距离测量分析,有利于聚集相关的网页,自动建立超文本链接。英国南安敦大学的"开放期刊计划"(Open Journal Project)开发了一个自动链接工具,根据语义的相似性定量分析,可将电子期刊有关的内容和有关的网页进行自动链接,并可以对有关文章的引证关系进行定量研究[6]。根据网络链接的结构,可分析站点的联系程度和集中度,进行网络结构的布局分析,合理配置资源。WebQuery 通过网页的链接关系和内容进行检索,然后再把结果用三维图像可视化手段表现出来,使相关节点之间的关系一目了然[7]。

(5)应用于网络资源检索和利用。美国斯坦福大学的数字图书馆计划开发的 Google 搜索引擎,利用超文本链接的定量分析来确定信息的重要性。IBM 公司开发出第二代的搜索引擎,综合利用了几种算法,可根据网页超文本的结构来发现高质量的信息,使检索结果更准确。目前还出现一种叫做"文献的书目示例计量检索工具"(Bibliometric Retrieval of Documents, BIRD),它先给出读者一篇感兴趣的文章,然后再根据其引文链找到相关资料[8]。

(6)网页链接分析有利于分析和掌握学科发展状况,如学科的独立性、吸收能力、渗透性、地位、发展动态和趋势等。如可通过利用网页的同链接或网络电子期刊的同引情况,了解一个学科的知识结构。美国伯克利加州大学信息管理与系统学院的 Ray R. Larson 教授曾利用 Alta Vista 搜索引擎收集到有关地球科学文献的同引情况数据,用同引频率矩阵分析了地球科学、地理信息系统、卫星遥感等学科相互关系以及发展趋势[9]。

(7)有利于开发和应用智能超文本链接。智能超文本链接是在链和节点中嵌入知识或规则,允许链进行推理和计算。有的学者把 WWW 看成是人类的大脑,网页之间的不同链接看成是人脑的神经[10]。神经网络理论认为,神经网络是由大量处理单元广泛互连而成的网络,是人脑功能基本特性的某种抽象、简化与模拟。网络的信息处理由神经元之间的相互作用来实现。人工神经网络具有良好的自组织、自学习和自适应能力,特别适用于处理复杂问题或开放系统,因而可以利用神经网络理论指导建立网络自动链接的机制。通过对用户的行为习惯如链接路线、书签等记录的统计分析,可建立一个自适应的链接系统,根据需要,自动链接有关网页,方便用户利用。

由此可见,网页的链接分析是网络信息计量学研究的一项重要内容。随着网页链接规模的不断扩大,网络结构的复杂化,利用数据挖掘和知识发现的方法和技术对网页链接进行分析,可以使研究结果更加科学化和精确化。

参考文献

1 刘连方等. 超文本/超媒体技术. 北京:国防工业出版社,1998
2 The mission of website. net: Building a Webcitation-index. http://www.website.net.html
3 Ingersen, P. The calculation of Web Impact Factors. Journal of Documentation, 1998(2)
4 邱均平. 信息计量学(一). 情报理论与实践,2000(1)
5 Zipf Curves and Website Popularity. http://www.dlib.org/december981/12hitchcock.html
6 Hitchcock. Citation Linking: Improving Access to Online Journals. http://journal.ecs.ac.uk/amend97.html
7 Jeromy Carriere. WebQuery: Searching and Visualizing the Web through Connectivity. http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html
8 Bird Resources: Bibliometric Systems. http://ai.iit.nrc.ca/html
9 Ray Larson. Bibliometrics of the WWW. http://sherlock.berkeley.edu/asis96/asis96.html
10

邱均平　教授,博士生导师,武汉大学信息资源研究中心室主任、《图书情报知识》副主编。通讯地址:武汉大学。邮编 430072。

黄晓斌　中山大学信息管理系副主任、副教授,武汉大学信息管理学院博士生。通讯地址同上。

(来稿时间:2001-11-20)