

●马建华

网络计量学研究进展

摘 要 网络计量学是随着网络技术的发展而涌现的一个新的研究领域,搜索引擎的性能、网页质量评价、网络影响因子概念、网上知识发现等都属其研究范围,目前在我国还处于试验和探索阶段。参考文献 23。

关键词 网络计量学 搜索引擎 网络影响因子 知识发现

分类号 G350

ABSTRACT Webometrics is becoming a new research field with the development of network technologies. It covers functions of search engines, evaluation of the quality of websites, Web impact factor and knowledge discovery on the Web. 23 refs.

KEY WORDS Webometrics. Search engine. Web impact factor. Knowledge discovery.

CLASS NUMBER G350

1 引言

1990 年, Paisley 提出将情报计量方法应用到电子通信领域:“现在的书、刊、报纸中的大部分文本信息将来都会被包含在电子数据库中,……所收藏的大量电子信息将来会成为文献计量学研究的主流。”^[1]网络计量学的开创性研究始于 Abraham 和 Larson^[2]。他们采用网络计量测度建构 Web 认知图

和数学模型,采用完整的模拟程序对 Web 页连续性进行测度,使用 AltaVista 去显示每个站点的网页数和网页间链接数目,获得连续性阵列原始数据,并记录下包含至少一个链接指向另一个 URL 的 HTML 文献的数目,显示出链接的相关网页数目。1997 年,丹麦皇家图书情报学院的 Almind 和 Ingwersen 首次提出“网络计量学”(Webometrics)一词。他们认为网络计量学包括了所有使用情报计量方法对网络通信

印张,将原来的 96 页内文增加到 104 页。改版后的《学报》将更加突出时效性和学术性。

图书馆学是一门正在发展壮大的学科,涉足在这门学科里的研究者数不胜数,每天都有大量的研究论文产生。源源不断的来稿给《学报》带来欣喜的同时,也使《学报》有一丝丝惆怅。原因是有的稿件篇幅太长,几万字的稿件屡见不鲜。长篇来稿占去很多版面,有的刊物发几篇文章所占的篇幅,在《学报》上只能发 1 篇,这主要是《学报》长期以来较多地发表长文形成的。今后《学报》将在不影响学术研究的前提下,尽最大努力减少长文,也希望我们的作者在投稿时,尽量把论文写得精练一些,以便使更多的优秀论文能在《学报》有一席之地。

《学报》迄今已经走过了 45 年的历程。45 年前《学报》还是一棵破土而出的小树苗,经过近半个世纪的风风雨雨,如今已成长为一棵根深叶茂的大树。《学报》的全体同仁清醒认识到,《学报》昨天的成就只是今天的起点。今后《学报》仍将不负众望,继续

坚持办刊方针,继续坚持理论联系实际的学风,继续坚持用稿质量的高标准,继续坚持和完善期刊编辑规范化,在继承的基础上有所创新,使《学报》产生更大的社会影响,赢得更多的读者,为我国的图书情报事业做出更大的贡献。

参考文献

- 1 周文骏. 老读者的心声. 中国图书馆学报, 1997(4)
- 2.6 李万健. 中国图书馆学期刊的发展现状及未来取向. 中国图书馆学报, 1999(1)
- 3 丘峰. 中国图书馆学报的过去、现在与未来. 中国图书馆学报, 1991(3)
- 4 彭斐章, 刘磊. 不负重望、再创辉煌——贺《中国图书馆学报》创刊 40 年. 中国图书馆学报, 1997(4)
- 5 全国第 4 届图书馆学期刊编辑工作研讨会会议纪要. 中国图书馆学报, 1994(1)

徐 菁 《中国图书馆学报》编辑。通讯地址: 北京中关村南大街 33 号。邮编 100081。(来稿时间: 2002-09-27)

有关问题的研究,而且认为“情报计量方法所使用的手段完全可以应用到万维网上,只不过是万维网看作引文网络,传统的引文由 Web 页面所取代。”^[3]

网络计量学与文献计量学、情报计量学及科学计量学研究存在着相似性,如对网页简单的记数及内容分析与传统出版物分析如出一辙;离开网页的链接(outlinks)及进入网页的链接(inlinks)分别与参考文献和引文分析存在相似性。与传统引文不同,网页是动态变化的;同时,由于缺少同行评议,网页上的信息质量很难保证。对网络查询行为模式的研究,可借用传统的信息查询行为的研究方法。网上的问题跟踪及知识发现也与通常的文献或文本数据库中的数据挖掘相似。

其实,研究方法的相似性是一种表面现象。比如,在科学或专业数据库中的引文习惯在网络中并不存在,再者,网页内容随时间而动态变化,这与传统数据库完全不同。

下面就当前对网络搜索引擎的搜索范围及性能的研究作一回顾,并就万维网的引文分析、网络影响因子(Web-IF)、网上知识挖掘、问题跟踪的研究现状进行评述。

2 搜索引擎的覆盖范围及质量研究

1998年,Lawrence和Giles对商业搜索引擎的网络空间覆盖范围进行了全面而深入的研究,并首次引入了“可索引的万维网”(indexable Web)的概念——能被搜索引擎编入索引的万维网(不包括像Dialog这样的网络数据库)^[4]。他们在1997年12月对包括AltaVista, Excite, HotBot, Infoseek, Lycos, NorthernLight在内的6个主要商业搜索引擎进行了试验。结果显示可获得索引的网页达3.2亿个,单一引擎的索引能力为3%~34%不等,这可能与这些引擎的网络带宽、存储能力和计算能力有关。他们指出,为了克服单一引擎的局限性,从而获得全面的检索结果,一方面可将几个主要引擎结合起来使用,也可通过利用一些具有自动抓取功能的研究型搜索引擎来获得信息。另外,Clarke和Willet就搜索引擎评价的方法学问题作了讨论^[5]。除了对以前研究工作的评价外,还提出了现实可行的搜索引擎评价方法,并对AltaVista, Excite和Lycos作了比较,发现AltaVista的性能明显优于Excite和Lycos。

还有一些研究者对链接页的检索范围或网络结构作了分析。许多研究者发现基于时间序列的搜索

引擎的检索结果变化很大且缺乏一致性。Rousseau于1999年对AltaVista和NorthernLight进行了21周的连续跟踪研究,每天使用3个同样的单一词作为检索词对此二引擎的索引结果进行比较。正如预期的那样,NorthernLight的检索结果呈稳步增长,这与网络发展的步调是一致的^[6]。然而,AltaVista的结果在一特定日期前始终处于很大的变化之中,检索结果不稳定,直到AltaVista重新启动新的更为稳定的引擎结构,剔除了无效链接,才进入到一个相对稳定阶段。因此Rousseau建议利用中值筛选(median filtering)来降低这种不稳定性对检出结果的影响。

国内对搜索引擎的研究经历了介绍——引进——开发的基本过程。章琳等选取Yahoo, AltaVista, Excite, HotBot, Infoseek, Metacrawler等搜索引擎对5个提问的结果总数、结果的相关性、结果的可连续性等进行评测,发现不同搜索引擎的检索结果重复不多^[7]。任瑞娟等选取了10种中文搜索引擎进行评测,发现中文搜索引擎在数据更新、记录去重、相关性排序等方面存在问题^[8]。夏旭等对性能较好的10种医学搜索引擎进行检索和比较分析,并对10个搜索引擎两两之间索引相同站点情况进行分析,发现不同搜索引擎索引相同站点的频次较低,网页重复率不高,说明不同医学搜索引擎的覆盖范围有限,同时还发现医学搜索引擎检索结果的相关性不如通用引擎^[9]。

3 对网页类型和质量的研究

由于网络在学术交流中作用日益增大,因此网络计量学对网址或网页的本质、结构及性能的分析显得尤为重要。1997年Almind和Ingwersen用多种文献计量学方法对北欧国家的网页的链接种类及类型作了研究。其方法包括对网页的分等级取样以及为了对网页进行分析而采取的网页下载。分析结果表明每个网页平均提供9个链接。另外,他们还比较了北欧各国在科学类网页中所占份额与在引文索引中的分布,事实是这些国家在网络中的显示度远远高于在通常的印刷型文献数据库中的比重^[10]。

因特网上的信息良莠不齐,有些甚至是虚假、错误的。因此信息的可靠性、有效性、相关性以及真实性成为人们关注的一个重要方面,并由此引发了对网络质量的评估问题。有人甚至提出了“网络考古学”的新概念,并指出“网络考古学”将与网络计量学方法相结合而共同向前发展。最近中国医科大学的Lei Cui应用引文分析法对网上被高频引用的医学信

息网址作了分析。文中提到网络引文分析已成为衡量网页质量的一个基本指标。Cui 利用美国排名前 25 位的医学院(实际分析采用了 19 所)网页链接的统计分析,发现在共 1731 个链接中有 1/3 的链接集中在 74 个网址上,这一结果符合布拉福德分散定律,网址分布为 1 4 42。该方法对图书馆员及情报学家评价网页质量提供了借鉴^[11]。

Rousseau 1999 年所做的研究中,对网页提供信息内容的准确性做了评估。他用了 3 个非常简单的提问(evolution; genetically modified organism; endangered species),通过搜索引擎检索,然后分别请两位专家对检索出的前 500 个网址做出筛选,找出 60 个与检索内容相关的网址,然后按照这些网址提供信息的准确性程度分类:将含有不正确信息的网址划分到“不准确”类;将对科学做出不正确解释的网址列入“误导”类;将没有引用经过专家评议的参考文献而提供信息的网址归入“未引用”类。结果这 3 个提问的“误导”率分别达到了 87.8%、82.8%、73.6%;“未引用”率均超过了 48%^[12]。这一结果使人们对网络信息的可信性产生怀疑,也使人们相信通过建立和发展经科学家评议的能提供即时的、准确的链接网址的门户网站或数字化图书馆才是解决这一问题的最可行办法。而且这些网站的建立不仅能提供准确有效的信息,也将为网络计量学分析提供一个研究平台。

4 网络影响因子研究

McKiernan 于 1996 年首先提出了 citation 这一新术语,用以研究网页之间的引用关系^[13]。Rousseau 认为对网页链接关系的研究与对发表文章的引文研究相似,但又不尽相同^[14]。他对网址的分布模式和进入网页的链接作了分析,发现在他所研究的 343 个网址中最高层域名服从洛特卡分布,而且对这些网址的引用也符合洛特卡分布,自引比例约为 30%。1998 年,Ingwersen 发表了《网络影响因子计算》的文章,对计算网址的影响因子的可行性和可靠性作了研究^[15]。他对网络影响因子的定义为:在一定时间,某个国家的网址(或某个网址)被其他网址和其自身所链接的网页数目的逻辑和除以该国家(或该网址)的所有网页数。这里需要注意的是定义中提到的“网页数目”的和,不是指链接次数的和而是链接页数目的和。通过对 7 个国家(挪威、英国、法国、丹麦、瑞典、芬兰和日本)网址和 4 个主要域名(gov, org, com, edu)的研究,

Ingwersen 得出了 3 个有趣的结论:(1)由于用作搜索引擎的 AltaVista 只能统计链接到某个网页的网页数目而不是链接的次数,因此自链接不会影响网络影响因子的结果。4 个主要域名的自链接 Web - IF 值平均为 0.5,“com”的自链接 Web - IF 值为 0.59,平均来自外部的链接页的 Web - IF 为 0.39。(2)对于非主要域名的个人网址,其 Web - IF 的可信度要小得多。(3)在做 Web - IF 计算时可以利用计算结果发生的变化作为评价网络引擎性能的一个测度。对于第 2 个结论,Smith 和 Thelwall 也分别作了研究,但遗憾的是他们几人的研究使用的均是不稳定版的 AltaVista^[16,17]。因而,他们均对引擎的覆盖范围和检索性能产生了怀疑。Thelwall 认为,现存的网络影响因子的概念在实际应用中还只是一个相对粗浅的提法,因而所得出的结论也存在很多问题。

5 网上知识发现和问题跟踪

数据库中的知识发现(knowledge discovery in databases, KDD)作为一个多学科研究领域,近 10 年来逐渐为人们所重视。Frawley 将 KDD 定义为对数据中隐含的、以前不知道的、潜在有用的信息的有价值的提取^[18]。KDD 技术包括:信息检索、统计学、机器学习、模式识别等。KDD 在目标及方法上与文献计量学有许多共同点,例如,文献计量学中的聚类分析技术可以看成是 KDD 在文献和引文数据库中的一个应用。由于 KDD 是针对数据库中呈指数增长的内容的开发方法,因而,对于网络这个呈指数增长的分布数据库,KDD 也可用于网络环境下的知识发现。但由于网络的分散性、多元性及动态性,在网络环境下的知识发现或网络挖掘显得较为困难。

网上知识挖掘主要包括 3 个方面,即对网页内容、链接结构以及网络使用者的信息使用行为的研究。其中对链接结构的研究是知识挖掘的重点。对链接结构的拓扑学研究关系到人或计算机能否实现网上知识挖掘。在对链接结构的研究中,图形理论法(graph theoretic methods)是一个极好的研究工具。Broder 等利用两亿网页和 15 亿个链接给出了看起来像“领结”形状的网络结构曲线图,构建了网络曲线的数据库模型^[19]。他们的研究结果表明,网络的宏观结构比以前人们在小规模范围内所构建的模型要复杂得多。这一模型无论对人们更好地了解网络结构的特性还是在机器的知识发现方面都具有一定价值。

除了对网络结构的探索,情报学家通过应用情

报计量学方法在网上进行问题跟踪,从而获得了以前未知的有用知识,认为数据和文本挖掘技术为情报分析提供了极大的可能性。Bar-Ilan(1999)选择 informetrics 为主题在 6 个主要搜索引擎进行了一段时间的搜索和跟踪。由于网络的动态性,随着时间的变化,有些文献消失了,而又有新的文献产生了,表现出研究对象的不确定性及研究结果的不可重复性^[20]。2000 年,他用同样的方法对同样的问题做了更为详细的跟踪,并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^[21]。

6 结果与展望

目前我国,网络计量学还处于试验和探索阶段,它是随着网络技术的发展而涌现的一个新的研究领域。将文献计量学方法应用到对网络的研究中,尽管产生了一些积极结果,但由于网络结构的分散性、多元性、动态变化性以及搜索引擎功能的不足,文献计量学方法在网络研究中的实际可操作性并没有人们预期的那样好。这也就是为什么基于搜索引擎数据的网络影响因子未能得以实施的原因所在^[22]。另外,由于网络文献及链接缺乏元数据以及搜索引擎不具备元数据开发能力,进而影响了网络的过滤和知识发掘功能,而传统数据库却非常支持 KDD。但这并不意味着在网络上无法进行 KDD,“横向链接”(transversal links)概念的引入(Björneborn),对实现网上 KDD 指出了一个新的可行的途径^[23]。对网络计量学的研究者而言,如何进一步从理论及方法上做新的探索,从而使网络计量学成为网络资源建设及评估的有效工具,是今后必须面对的一个挑战。

参考文献

- 1 P. William. The Future of Bibliometrics. In: Borgman C L, ed. Scholarly Communication and Bibliometrics. [s.l.]: Sage, 1990
- 2 吴华香. 网络计量学:互联网上的文献计量学. 图书馆杂志, 2001, 20(1)
- 3, 10 T. Almind, P. Ingwersen. Informetric analyses on the World Wide Web: Methodological approaches to "Webometrics", Journal of Documentation, 1997, 53:404-426
- 4 S. Lawrence, C. L. Giles. Searching the World Wide Web. Science, 1998, 280:98-100
- 5 S. J. Clarke, P. Willet. Estimating the recall performance of Web search engines, Aslib Proceedings, 1997, 49:184-189
- 6, 12 R. Rousseau. Daily time series of common single word searches in AltaVista and NorthernLight. Cybermetrics, 2/3, paper2. 1999, Issn:1135-5019. (<http://www.cindoc.csic.es/cybermetrics/articles/v2ilp1.html>)
- 7 章琳,张保明. WWW 检索工具的比较研究. 情报学报, 1998, 17(4)
- 8 任瑞娟,李洪建. 中文 WWW 搜索引擎比较研究. 大学图书馆学报, 1999(5)
- 9 夏旭,方平. WWW 网络生物医学信息资源搜索引擎的比较研究. 医学信息, 2001, 14(1)
- 11 L. Cui. Rating health Web sites using the principles of citation analysis: A bibliometric approach. Journal of Medical Internet Research, 1999, 1(1)e4
- 13 G. McKiernan. CitedSites (sm): Citation Indexing of Web resources. <http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm> (1996)
- 14 R. Rousseau. Situations: an exploratory study. Cybermetrics, 1997, 1(1) (<http://www.cindoc.csic.es/cybermetrics/articles/v1ilp1.html>)
- 15 P. Ingwersen. The calculation of Web Impact Factors. Journal of Documentation, 54:236~243, 1998
- 16 A. G. Smith. A tale of two web spaces: Comparing sites using web impact factors. Journal of Documentation, 1999, 55:577~592
- 17 M. Thelwall. Web impact factors and search engine coverage. Journal of Documentation, 56:185~189
- 18 W. J. Frawley, et al. Knowledge discovery in databases. Menlo Park, Cal: AAAI Press, 1991
- 19 A. Broder et al. Graph structure in the Web. <http://www.almaden.ibm.com/cs/k53/www9.final>
- 20 J. Bar-Ilan. Search engine results over time: A case study on search engine stability. Cybermetrics, 2/3, paper1 (<http://www.cindoc.csic.es/cybermetrics/articles/v2ilp1.html>)
- 21 J. Bar-Ilan. The Web as an information resource on informetrics? A content analysis. Journal of the American Society for Information Science, 2000, 51:432~443
- 22 L. Björneborn, P. Ingwersen. Perspective of webometrics. Scientometrics, 2001, 50(1), 78~79
- 23 L. Björneborn. "Small world" Web and Possibility Space: outlining a conceptual framework for transversal links on the WWW. Master's Thesis. Royal School of Library and Information Science, Copenhagen

马建华 北京大学信息管理系博士研究生。通讯地址:北京大学信息管理系。邮编 100871。

(来稿时间:2002-04-01)