

国外网络引文研究的现状及展望*

杨思洛

摘 要 国外的网络引文研究可分为四个方面,侧重点各不相同:Print-Print 引文研究着重于网络环境下传统引文数据库的分析;Print-Web 引文研究大多集中在网络引文的可获得性方面;Web-Print 引文研究注重实证与传统引文的比较;Web-Web 引文研究则是将来引文发展的总趋向。网络引文数据的动态性和不可靠性,网络文献缺乏有效控制与规范,以及数据的精确性问题,都给引文分析带来一定困难。目前相关的研究还处于开创性和探索性阶段,未来的网络引文研究将重点从理论体系的构建,研究方法的完善与实践应用的深化三个层面展开。图 1。参考文献 67。

关键词 网络引文 Print-Print 引文 Print-Web 引文 Web-Print 引文 Web-Web 引文

分类号 G250.7

ABSTRACT Web citation analysis has become an important direction and trend of bibliometrics studies. Foreign researches of Web citation can be divided into four areas: Print-Print citation study focuses on the databases of traditional citation in the network environment; Print-Web citation study focuses on the availability of Web citations; Web-Print citation analysis focuses on comparison with the traditional citation; Web-Web citations research is the general trend of Web citation. At present, the Web citation analysis is in the first and exploratory stage. Future research will focus on three aspects: construction of theoretical system, improvement of research methods and deepening of the practice and application. 1 fig. 67 refs.

KEY WORDS Web Citations. Print-Print Citation. Print-Web Citation. Web-Print Citation. Web-Web Citation.

CLASS NUMBER G250.7

1 前言

引文是现代学术论文的重要组成部分,它表征了文献间的联系,记录了作者的学术行为特征。1955 年,E. Garfield 受 1870 年出版的法律工具书——谢拔德引文的启发,开创了科学引文索引,掀开了文献引文分析的新篇章^[1]。目前引文分析已成为文献计量研究的主要方法之一,广泛地应用于科学交流研究中,如对科学出版物、学者和机构的评价(如 H 指数),用于信息检索、知识组织及科学前沿发现(如知识图谱),科技发展过程模型的构建等^[2]。随着现代信息技术的加速发展与应用普及,互联网已成

为科学交流的主要媒体与中介,越来越多的学者通过网络来交流和获取信息、进行科学发现、保持与同行联系^[3]。互联网也促成了网络引文的出现和发展,近年来国外有众多学者关注网络引文的研究与探索^[4]。我们认为,传统引文是传统环境下纸质文献间的参考行为的结果。网络引文是在传统引文基础上发展而来,但与传统引文有显著不同:在引文的施引过程中,网络引文的作者主体、引用行为都存在于网络环境中,文献对象类型也有变化;另外网络引文分析通过各种网络数据库或搜索引擎进行。网络引文与一般网络链接也有区别:首先,网络引文表征了文献之间的引用关系,突出特点是其学术性;其次,网络链接通过超链接实现,而网络

* 本文系教育部人文社科基金“网络引文的相关规律及其应用研究”的阶段性研究成果之一。

引文虽可以超链形式出现,但大多为引文格式并通过引文分析方法进行研究。

为总体把握网络引文研究现状,本文以 ISI Web of Knowledge 为数据源计量相关论文外部和内容特征。浏览以“web citation”为主题的论文摘要,对以“web citation”为题名的论文详细分析,并通过引用关系扩展阅读,重点关注核心作者和高被引论文。我们发现,网络环境下文献发生了巨大变化,形成了传统文献(print, P)和网络文献(web, W)两大类型。一方面,传统的学术论文仍是人们进行学术交流的主阵地;另一方面,网络资源和网络交流手段以其无可比拟的优点受到人们的青睐,其关系见图 1。目前网络引文研究还处于初步阶段,国内研究大都认为网络引文是将网络资源作为学术论文参考文献的一种引文形式,其突出特征是引文中含有网址 URL,可称为 P-W 网络引文^[5]。国外网络引文研究分为四大类:①P-P 网络引文是指网络环境下的传统论文间的引文,主要集中在各类网络论文数据库中引文的研究;②P-W 网络引文更多的是探讨传统论文参考文献中的网络成分,突出表现是参考文献中有网址;③W-P 指从网络文献引用纸质文献(包括期刊论文、会议文献、手册指南、专论等)^[6];④W-W 引文是指源文献和引文文献都属于网络文献,但它们与一般的链接有显著不同。

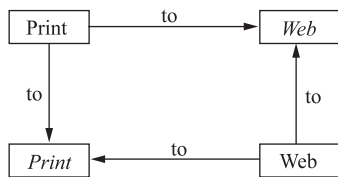


图 1 网络环境下论文间的引用关系模型

2 P-P 网络引文的研究

目前大多数传统文献引文分析都是基于网络数据库,引用行为也发生在网络环境下,所以都可认为是 P-P 网络引文。下面主要对网络环境下引文变化和数据库研究进行简述。

2.1 网络环境下引文的变化

互联网对学术交流和信息利用行为产生了重大影响,一些研究者从引文的角度对这些变化进行分析,由于样本、研究方法和角度的差异,得出的结论也不同。例如,芝加哥大学社会学家 Evans 在 *Science* 杂志发表的研究表明^[7]:随着越来越多的期刊可在线获取,论文更多地引用近期所发表的文章,同时越来越少的期刊和文章被引用。这一趋势产生的效应会更容易导致学术趋同性,并减少学术界活跃的争论声音。他的研究数据基于 Web of Science (WoS) 引文索引,统计包括 6000 种最突出的学术期刊。Larivière^[8]以自然和工程科学、医学、人文社科的论文为例,通过三项指标分析了 1900—2007 年 WoS 中论文引文的集中度变化,结果与 Evans 的结论相反,即引文的分散度明显增加。在另一个研究中 Larivière^[9]也得出与 Evans 相左的结论,当前的学术论文引用越来越多的老文献。Egghe^[10]用数学模型证实了 Larivière 的这一结论,表明参考文献年龄的均值和中位数都在增加,而普赖斯指数在下降。

2.2 引文数据库的对比分析

基于引文分析研究与评价的质量、合理性和可靠性很大程度上依赖于所用引文数据的精确性和全面性,不准确或不全面的数据往往造成不精确甚至错误的结果。直到 2004 年 Scopus 数据库发布前,四十多年内,绝大多数引文分析数据源来自 ISI 的三大引文索引:SCI,SSCI 和 A&HCI^[11]。目前引文数据库发生了很大变化,一方面随着科技的进展,三大索引经历了纸质、磁带、磁盘和光盘再到网络数据库 WoS 的变化;另外,也出现许多新的专业或综合性网络引文库,常见的就有十多种。众多数据库的出现增加了引文分析数据选择的余地,其竞争也促进了数据库质量的提高^[12]。国外对 P-P 网络引文研究主要集中在对各引文库的对比分析,系统总结针对具体领域的最佳数据源;对比分析各数据库的期刊覆盖度和检索功能;研究 WoS 引文的合理性,用其它数据库替代它的可能性;在验证或评价研究对象时,使用多种引文源的

价值与必要性等方面。

目前两大综合性的引文库为 WoS 和 Scopus, 在覆盖广度方面, 后者比前者的期刊数多; 在深度方面, 前者能检索的年段要长。两者各有优缺点, 且在不断地改进与完善之中。

在实践中要根据所选研究任务的内容范围和时间段来合理选择数据源^[13]。Lopez 分析了肿瘤学期刊在两数据库中的覆盖度^[14], 结果表明 WoS 包括影响因子较高的肿瘤学期刊, 其差异源自基本收录政策的不同: WoS 基于加菲尔德的理念, 即选用使用或引用频率最高的那部分期刊; Scopus 则有更大的期刊量, 是普通的科学期刊库。Gavel^[15]通过数据库商提供的收录期刊列表, 从总体上对比了 WoS 和 Scopus 中期刊的收录情况, 发现 WoS 所收录的期刊中 84% 被 Scopus 索引, 但在 Scopus 中仅有 54% 的期刊被 WoS 所收录。另外, Lokman^[16]随机抽取北美 80 名信息研究者, 对他们论著引用情况进行分析, 表明评价对象为微观实体, 如期刊、机构或会议集时, WoS 和 Scopus 产生显著不同的结果, 在具体评价时两数据库都需要; 在评估范围比较大的实践中, 如研究某领域或国家范围, 两数据库会产生非常相似的结果。Meho^[11]则分析了一个英国机构的 22 名人机交互研究者的引用情况, 结果显示 Scopus 包含有更多的人机交互方面的研究论文, 主要是因为它收录了许多 ACM 和 IEEE 同行评议的会议论文。在期刊引文方面, 两数据库没有显著不同。Scopus 虽然收集更广, 但并没有改变学者的引用排名; 针对单个学者, Scopus 则产生不同的引用关系图。

当前 WoS 系列数据库包括光盘版和网络版 (SCI-E), 仍是引文分析最主要的数据源。Rodrigo^[17]详细对比了两大版本的收录内容和格式特征, 发现两者有许多不同, 例如收录范围、格式中标点符号、作者单位著录的差异以及这些不同影响版本间的自动分析和使用的方便性等, 并提出了 WoS 在标准化和一致性方面改进的建议。另外, Whitley^[18]对比了 WoS 和 CA 化学文摘网络版 (SciFinder Scholar), 分析了两数据库中特定化学领域学者的作品引用情况的异同, 认为虽然两数据库收录非常相似, 但在当前

的版本中, 两数据库产生的结果存在显著不同, 最佳办法是两数据库都使用, 下载所有结果合并去重分析。除了期刊外, 图书也是学术交流的重要载体, WoS 等数据库并没有对图书引文专门索引, 这使得图书引文研究较少; 现在 Google 新开发了图书搜索功能, 为图书引文分析提供了工具和数据源。Kousha^[19]分析了 10 个学科中, 图书引用学术期刊论文的情况, 结果显示不同学科的引文数不同; 图书引用与期刊引文有相似之处, 但也存在差异, 图书引用情况能为引文分析提供额外信息; 在人文社会科学中, Google 的图书搜索是一个有价值的引文分析数据源。

P-P 引文研究本质上属于传统引文研究的范畴, 只是在网络环境下, 其引用模式和分析方式都发生了变化。

以上这些实证研究总结了不同网络引文数据库的特征, 但是, 大多是专门研究某一特定学科领域的, 不同学科的差异, 使得结论的普遍适用性存在问题。另外, 各数据库自身也在不断地变化之中, 包括搜索功能与收集数据的范围, 目前 WoS 收录有 9000 多种刊物, 而 Scopus 则包括 15000 多种。前几年的相关研究表明, WoS 并没有收录会议论文, 但现在它已包括了多个学科会议数据库。所以在研究中要注意各数据库的最新情况, 使用数据库一定要注明使用的版本号和时间。

3 P-W 网络引文的研究

P-W 引文是指学术论文后面所附的、其来源出处是互联网资源的参考文献 (URL)^[20]。在互联网发展的初期, Zhang^[21]检验了 10 种图书情报期刊中网络资源的引用情况, 得出“对比印刷文献, 在科学研究中网络资源的影响非常小”。但是随着网络的发展, 他通过历时分析表明, 在论文中引用网络资源的作者数量和比例都显著增长^[22]。许多学科学者都参考网络文献, 包括计算机科学^[23]、法律^[24]、医学^[25]、跨学科^[26]期刊, 还有学位论文^[27]。

3.1 P-W 引文的利用情况

众多研究表明,有相当数量的网络引文存在于期刊论文中。Dellavalle^[28]选用 *Science* 等 3 种影响因子排名前 1% 的高质量美国期刊,统计论文中网络引文的频率和类型,结果显示 30% 的论文包括至少一个网络引文,网络引文占所有引文的 2.6%,比例最高的论文类型为 reviews,最低的为 letters。Oermann^[29] 对护理期刊的分析表明,平均每篇论文有 3.1 篇网络引文,大多网络引文与临床实践相关。Veena^[30] 则选择 9 种电子期刊 2000—2006 年的论文,研究表明在所有 25730 条引文中 43.52% 为网络引文,81.49% 的论文有网络参考文献。除了简单介绍网络引文使用率外,Brown^[31] 还较深入地分析其它方面。他以 8 种化学期刊为样本,发现虽然网络引文使用比例较少,但在稳步增长;网络引文在论文的参考文献中出现的数量最多(43%),其次是论文实验部分;网络引文的可获得性随时间的增长而下降;网络引文本身的类型中,纯文本的比例最大(85%),其次是文本和图像混合型。

由于学生对网络信息等新资源接受的程度较高,而且他们没有形成较为稳定的学术规范,故这一领域受到国外学者的重视。Davis 对同一类样本进行持续研究,首先他分析了 1996—1999 年研究生微观经济学课程学期论文的引文情况,发现在 4 年中图书类引文从 30% 下降到 19%;报纸类从 7% 增加到 19%;网络引文从 9% 上升到 21%^[32]。2002 年,Davis 发现由于网络和报纸类引文的增加,篇均引文持续增加;因为要求学生递交论文电子版,网络引文的可获得性显著增加^[33]。他建议教师或图书馆员对论文内容和格式提供更详细明确的要求,最好提交电子论文。2003 年他又检验了 2001 年同类学生论文的情况,发现当教师提供了明确而具体的指南后,网络引文的准确性和持久性也显著增加了^[34]。在 Davis 研究基础上,Robinson^[35] 以加拿大某高校政治学专业学生的论文为例,通过制定详细分类表,把论文中网络引文分为学术型(如图书、期刊、政府出版物)和非学术型(包括新闻、杂志和其它类型),发现大多

数学生的网络引文并没有指向学术资源,互联网没有导致学生研究质量的下降,而是提供了可获取更广泛资源的途径。相类似的还有 Herring、Gillette、Nicholas G 和 Grimes 等人的一些研究^[36]。

3.2 P-W 引文的可获得性

网络引文的可获得性是发挥其作用的基础,引文的不可打开或不可获取成为愈益重要的问题。Harter^[37] 是此领域最先的研究者之一,他对 74 种电子期刊(1993—1995 年)的 83 篇网络引文进行检验,发现在 1995 年仅有 43 篇可获取。此后有众多学者从不同年段、对不同学科的期刊论文网络引文进行验证分析,例如:分析在引文库中计算机论文网络引文的无效率^[23];在两种计算机期刊中网络引文的半衰期^[38];在高影响力肿瘤期刊中,网络引文可获取性^[39];在 MEDLINE 数据库中网络引文的稳定性与持久性^[40];D-Lib 杂志论文中网络引文的不可获取性^[41];在 6 种传播类期刊中 1600 个网络引文的稳定性^[42];6 种新西兰期刊在 2002—2005 年内网络引文的不可获取性;生物医学期刊网络引文的不可获得性^[43]。为解决网络引文的不可获取,学者提出了一些建议,包括技术、政策和法律诸多方面。如建议作者和读者重视网络引文及引文格式的标准化, Digital Object Identifier (DOI) 和 PURL 技术的推广^[44], WebCite、Internet Archive 和 Google 服务的广泛应用等^[45]。例如 F. Casserly 在对 1999—2000 年的 500 个引文样本调查中,直接可打开的网址为 39%;对不可获得的引文用 Google 检索,又可获得其中的 30.7%;对用搜索引擎还不可获取的网络引文,又用 Internet Archive 进行两次查证,分别可获得其中的 47% 和 50.8%^[46]。Lawrence 随机选择了 300 个不可直接打开的引文,发现通过搜索引擎分别可获得其中的 44% 和 53%^[23]。

P-W 网络引文引起了国外学者的广泛关注,虽然已有研究对网络引文的可获取性十分重视,从不同学科数据源进行验证,但所选的数据样本一般都较少,有关网络引文的总体图景和引用机理不明;目前大多数研究都集中在网

络引文的可获取性,对其它方面研究很少,如网络引文是否和传统引文一样具有评价和检索功能;网络引文的接受和使用对科学交流有怎样的影响。目前存在系列问题还有待深入研究:①P-W引文一般判断引文中是否有URL,但往往人们使用网络资源时并不著录网址或者网址著录很随意;②在实践中获取和处理大量的网络引文数据比较困难,像SCI和LISA等许多数据库没有在记录中包括网络引文;③网络信息本身也在快速变化之中,例如Web2.0的发展,网上信息知识大量增长,出现了博客、书签和Wiki等众多网络资源,但目前几大标准如APA、IEEE和NLM等在著录Web2.0资源时都存在问题,这需要各方共同努力来应对^[47]。

4 W-P网络引文的研究

W-P引文主要是指传统纸质文献被网络文献所引用。网络文献是作者的智力或创造活动的成果,可以是XML和HTML网页或网站,也可以是PDF、DOC或PPT类型文档,还可是图像、视频或声音等。W-P引文可通过新兴的引文索引如CiteSeer、CiteBase、Scirus、Google及Google Scholar等获取数据;另外,传统引文库也重视网络资源,开发相应的功能,如在Scopus的检索结果中全面集成网络信息,拥有超过4亿科技信息网页,可以检索论文被网络文献引用的情况^[48]。

4.1 W-P引文与传统引文的比较

国外对网络引文的研究很大部分是对W-P引文的研究,而且集中在W-P引文与传统引文的对比分析上。因为引用中的一方是印刷型文献,所以与网络链接存在明显不同。Steven专门探讨了网络链接和W-P引文的异同^[49]:①网络链接局限在网络空间内,而W-P引文一方为传统文献,指向网络之外,这是网络链接无法连接到的。②W-P引文是持久的,虽然有时获取比较困难,但永远不会消失;而网络链接是动态的、脆弱的,当它链接指向的资源消失的时候,就不能再发挥作用。③一旦网络文献被打

印出来,网络链接就失去作用;而W-P引文会转换成传统引文。④W-P引文与传统引文一样,具有作者、出版年、题名、杂志名称等题名项;而网络链接简单地由锚来指示,相应元数据仅由目标URL地址来组成。

已有许多人讨论网络引文的作用,一些人认为在对论著的影响力评价中,网络引文可与传统引文相当、甚至替代传统引文,但也有些人对此存在异议。Vaughan^[50]对比了46种图情期刊论文的网络引文与传统引文。她使用Google提取网络引文,对传统引文和网络引文的异同展开研究后发现:57%的期刊的网络被引与SSCI和JCR中的传统被引数显著相关;许多网络引用出于学术目的(30%为传统论文张贴在网上,12%为课程阅读书目);对于同一篇论文,通常网络引文数要比传统引用数多;对于有自己网站的期刊,其网络引文往往更多;从1992年至1997年,期刊论文网络引文数逐年增加。尽管网络引文有许多优势,但她认为在近期不可能替代传统引文分析。在后续研究中Vaughan^[51]又选择代表美国图书情报研究者学术成果的1483种出版物,用WoS、Google和Google Scholar检索,结果显示:在WoS中各类型论著(除图书外)被引的中位数为0;在Google Scholar中期刊论文和图书的被引中位数分别为1和3;而在Google中,中位数最大的为图书,达41,最小的为会议论文,为9。通过对论著的网络引文分类分析,发现Google Scholar中有92%的引文是学术性引用(主要来源于期刊论文);在Google中书目服务(非学术性引用)是最大的引文源;开放获取期刊通常有更多的网络引文,但传统期刊的网络引文往往代表着学术性引用。尽管Google Scholar存在问题,如引文和源文混在一块儿,由于信息的多种复本,一个引文往往重复出现多次。但是它有潜力提供有用的学术评价数据,特别是要对一个领域快速、详细分析的时候。

Steven^[49]详细定义了W-P引文的含义,研究了传统文献的网络影响,即网络引文的数量,探讨了在人文学科网络引文替代传统引文索引的可能性,结果显示了网络引文的重要性,但在

选择的样本中网络引文数太小而不能得出普遍意义的结论。Bauer 选取 1985—2000 年 JASIST 刊登的 2000 篇论文作为样本,采用作者同被引分析法、t 配对检验法,发现 Google Scholar 中的引文数量远高于 Scopus 或 Web of Science^[52]。Van^[53]利用 Google 进行网络共引分析,发现它能有效地发现科学论文之间的联系。在最近的研究中,Kousha^[54]综合多种网络引文数据,提出学术成果评价的新指标 IOI (Online Impact Indicator)。

4.2 W-P 引文的类型

网络引文所包含的文献类型多于传统引文。2003 年,Vaughan 将网络引文分为 7 类:期刊(引用期刊的发行者或主办者的网址)、作者(包括文献作者的通邮地址或网址)、服务(主要指文章中列出的文献网络地址)、课程(一门课程中所开列的阅读书目)、论文(网络文献的正式参考引文)、会议(会议文献中的被引文献)以及其它^[50]。2005 年,Vaughan 使用该分类方法,研究生物学、遗传学、医学和跨学科的网络引文和传统引文^[6]。她从学术层面将网络引文分为:①对研究产生影响的,类似于传统的参考文献;②其他与智力相关的(课程提纲、学术网站、学术性的问答系统中提到的参考文献等);③与智力无关的(表格中、在线书目和作者的个人主页中提到的参考文献)。研究发现,30% 的网络引文属于第 1 种类型。2008 年 Vaughan^[51]将网络引文分类扩展到 10 类:期刊论文、会议论文、图书、技术报告、学位论文、课程论文、课程阅读书目、博客文章、在线文章和其它。基于网络论文的 W-W 引文研究中也涉及到引文的分类问题,Kousha^[55]将其分为 15 个小类,并合并成 4 个大类:学术交流目的,非正式交流目的,导航及其他目的;在 2007 年的研究中他又将分类扩展到 6 个大类,21 个小类^[56]。

4.3 W-P 引文数据的获取

对于 W-P 引文数据,目前必须依赖于引文库或搜索引擎间接提取,这些方式在 W-W 引文研究中也普遍使用。在基于引文库的数据获

取中,CiteSeer 和 CiteBase 等较常用。CiteSeer 由美国 NEC 集团公司的研究所创建,该索引基于互联网,自动在网上收集各类型的对公众开放的计算机科学领域的学术文献,对收集到的文献进行引文标引,其特点包括:收录的文献使用互联网作为传播媒体;包含代表着学术传播不同阶段的文献类型;比 ISI 引文库提供更多的被引用文献的信息;制作和维护高度自动化;提供引用文献的全文信息;对论文后的参考文献按其被引次数排序;提供许多 ISI 引文索引所没有或根本不可能提供的有价值信息。Goodrum^[57]和 Zhao^[58]都以 CiteSeer 为引文数据获取工具,分别对不同学科的网络引文情况进行了分析。

利用搜索引擎获取引文数据,目前主要通过 Google 和 Google Scholar 获得。早在 2002 年,Wilson 就利用搜索引擎,检索一组他本人论文的网络引文情况。Vaughan^[50]使用 Google 提取网络引文,对传统引文和网络引文的异同展开研究。具体的 Google 网络引文提取方法为:利用 Google 搜索引擎键入每篇文章的题名,检索引用情况以及被引次数,并在必要时加入文章副标题及作者项检索,以减少误检率和漏检率。还有一些学者采取综合提取方法,如 Kousha 结合网络引文获取方法和链接提取方法,综合使用 Google 和 Google Scholar 及 URL 搜索定位法获取网络引文数据,被称之为 Web/URL 引文法^[59]。

目前对 W-P 引文的研究也存在许多问题。大部分研究必须依赖于网络引文库或搜索引擎。以 Google Scholar 为例,由于商业机密等原因,其收录与检索内在机理还没公开,对人文社会科学的信息收集不足,而且存在语言方面的偏见;逻辑组配方式也不完全,没有提供截词检索;检索结果排序方式单一,仅按论文被引数排列;系统对检索词的同义词提供不够,检索结果中“相关文章”信息不全等。另外,由于网络的开放性和便捷性而易导致恶意点击。作者可以利用网络薄弱的审查机制,大幅度增加自引次数,从而达到提高网络引文数的目的。另外一个问题是如何区分网络引文中的“学术引用”与“仅仅涉及名称的引用”。E. Garfield 在“*Impact*

factors, and why they won't go away"中提到,希望网络引文能够标准化,形成一系列的参考标准从而适于精确计算,否则网络引文是不能与 WoS 等的引文分析相提并论的。

5 W - W 网络引文的研究

W - W 引文指原始文献和引文文献都属于网络文献的情况。随着网络学术交流模式越来越普及,W - W 引文分析很可能是将来的研究趋势。网络引文与传统引文和网络链接有着较密切的联系,并逐渐从链接分析的研究中独立出来,但网络引文的功能更倾向于传统引文。

5.1 基于链接关系的 W - W 引文

基于链接关系的 W - W 引文,即引用与被引之间存在链接关系,但与一般链接不同,它们具有学术性和规范性,最重要的是利用传统引文分析理论与方法进行研究。

Rousseau^[60]是较早对网络学术信息的引用情况深入研究的学者之一。他积极推荐采用 Sitation 术语,系统比较了网络链接关系与传统引用关系。他认为从用户动机角度分析,两者有相似之处,但又不尽相同,网络引文的情况要复杂得多。他对 343 个网址的分布模式和进入网页的链接情况作了分析,发现网址中的最高层域名服从 Lotka 定律,甚至对这些网址的引用也符合 Lotka 定律。与期刊中的引用关系一样,在对网站的引用中也存在自引,自引率大约为 30%,略低于一般期刊的自引比例。

Smith^[61]首创将引文分析方法应用于网络期刊。他选择 22 个澳大利亚科学家经常参阅的网络期刊,利用 AltaVista 统计从其它网站指向这些网络期刊的链接,但统计结果没有显示与 JCR 计算的影响因子之间有显著的联系。他认为指向网络期刊的超链接与引文有本质的区别,因为超链接的目标是整个期刊,而引文却是针对个别文章。在后来的研究中他又调查了哪些链接与传统引文相似,通过对链出页和链入页及链接动机的特点进行分类,表明总体上 20% 的链接可看作学术链,类似于传统引文的

功能^[62]。

将文献计量学中的“文献耦合”、“同被引”等概念应用到网络文献的链接分析中,Rousseau 将其称之为 Co-Sitation。Abraham 和 Larson 被认为是这方面研究的开创者,他们把互联网当作未来的分布式数字图书馆,将基于印刷型图书馆知识结构的分析工具和方法,例如同被引理论,应用于互联网环境的信息资源。Larson 利用搜索引擎收集到有关地球科学文献的同引数据,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^[63]。Zuccala^[64]以数学领域为例,对网络共链分析和作者共引分析从样本选取、数据搜索策略、构建矩阵、图示化和聚类、对结果的解释等角度详细比较,认为二者可称为网络计量学与文献计量学中的姐妹技术,但两者存在许多不同,有待深入研究。

5.2 基于网络论文的 W - W 引文

基于网络论文的 W - W 引文,主要研究网络上发表的学术论文的引用情况,如大量的开放获取和预印本论文及机构知识库中的文献。目前使用的主要索引工具有 Google 和 CiteSeer 等,2005 年汤姆森科技信息集团也推出“网络引文索引,Web Citation Index(TM)”。

网络为我们提供了新的引文信息,可以测评成果的不同方面。Kousha^[55]提出了 URL 引文这一网络引文形式,定义为网络文献的 URL 在另一网页文本中被提到(不论是否为超链接)。通过 Google 搜索,他分析了 2000 年 15 种图书情报网络期刊 182 篇论文中的 3045 个 URL 引文,这些引文 43% 有着与传统引文相同的正式学术动机,18% 为非正式学术动机。这些 URL 引文,82% 是英文,88% 可获取全文,58% 是非 HTML 文档;正式学术交流目的的 URL 引文主要源于非 HTML 文档中的文本类型。在后续研究中^[59],他选择 2001 年的 108 种开放获取期刊中的 1080 篇论文,URL/Web 引文(定义为在一个网页中提到论文的网址或题名或两者都有)的统计使用 Google 和 Google Scholar,并与传统 WoS 引文对比,发现三种引文数据显著相关。

与 Google 得出的引文相比, Google Scholar 引文与 WoS 引文相关度更高;在社会科学领域, Google Scholar 比 WoS 中有更多的引文数; Google Scholar 与 WoS 引文的比较,存在显著的学科差别;虽然网络引文有很好的发展前景,但目前用它来替代 WoS 还存在许多问题。2007 年 Kousha^[56] 又对 64 种自然科学领域开放获取期刊的 1577 个网络引文进行深入的分析,结果显示:仅有 25% 的引文为学术目的,其中网络文档引用占 23%,其它非正式学术引用占 2%;许多引用仅是一般的或具体的导航(45%)和自我宣传(22%);另外,网络引文类型在学科间也存在差异。

Zhao^[4] 使用 CiteSeer 和 SCI,采用作者同被引分析法,研究 XML 领域的文献,结果表明使用网络作为数据源,相比 SCI 具有优势(引文数量更大、文献类型更多、引文包括的信息更多、分析使用更方便),同时也有不足,分析时最好能结合两种数据源。Zhao^[65] 进一步比较了基于互联网和基于期刊的学术传播模式,发现网上出版物与期刊文献代表学术传播过程中的不同阶段,其中期刊文献比网上出版物要滞后几年。这说明采用网上出版物进行引文分析具有揭示科研最前沿状况的优势。Brody^[66] 针对引文从使用到被分析具有几年的时滞,而网络文献的使用情况能方便地记录与统计的特点,使用电子预印本系统——arXiv.org 为工具,得出前期的网络文献使用统计可以预测文献引用情况的结论。Goodrum^[57] 等人探讨了学术传播系统向新模式的发展,讨论了研究这些新模式的重要性和使用全自动引文索引进行这类研究的可能性。他们以 CiteSeer 和 WoS 为工具,就高频被引文献在文献类型和出版日期的分布,对网络 and 传统出版物进行比较,发现会议论文在高频被引网络文献中的含量远远高于期刊中的含量,网络文献更多地引用近期文献;从两种数据源中提取的高频被引文献有很大的重合。另外, Di Cesare^[67] 以“人口老化”这一主题为例,使用 Google Scholar 为工具,对被引文献中的灰色文献与传统论文进行了深入对比。

网络引文数据具有的动态性和不可靠性,

给引文分析带来了困难。大部分网络文献并没有具体的出版日期。对学术评价的一般情况和总体结构分析时,网络引文十分方便,但在研究科学交流的进化和历时研究时则存在困难。由于网络文献类型的多样性,而一般网络搜索引擎或索引库都没有像 SCI 那样进行较严格的人工索引,所以数据的精确性还存在问题。网络文献缺乏有效的控制与规范对引文分析也是一种挑战。

6 网络引文研究的展望

引文分析是图书情报学十分关键的研究内容,由于其在科学评价、信息检索和知识发现等方面的应用,因此也成为图书情报学影响其它学科领域的重要内容。网络环境下,真正与传统引文相对应的不是网络链接,而是网络引文,网络引文分析是文献计量学中引文分析在网络环境下的延续与发展,成为当前研究的重要方向和趋势。网络引文的相关研究目前处于初步探讨阶段,以开创性和探索性研究为主,目前急需深入的分析与完善。另外随着网络应用的普及发展,网络引文还会有新的问题需要研究。

(1) 理论体系的构建。作为一个随着网络的普及和发展而诞生的新兴研究领域,与传统引文相比,网络引文的引用动机更加多样,引文研究的假设前提和机理更加复杂,因此需要深入分析网络引文的基础理论。从目前情况看,急需解决网络引文研究缺乏整体性和系统性的现状,清晰地界定与网络链接研究的重叠、交叉和难以定性的地方。4 种类型的网络引文分析有各自的研究重点和研究方法,需要整合,并借鉴传统引文分析和网络链接研究的理论,构建完整的网络引文分析理论体系。另外,科学研究领域划分的依据是其独特的研究对象,因而必须明确网络引文研究的对象、最基本的概念和研究范畴,明晰网络引文研究的背景与意义、性质、类型和目的,这是其方法 and 应用研究的基础。

(2) 研究方法的完善。目前的网络引文研究基本上是利用现有的商业搜索引擎或索引数

数据库,在具体的网络引文研究实践中,没有统一规范的分析方法、手段和工具。因此需要建立一套包括确定样本、获取数据、筛选、分析与处理、结果表示等过程的完善的方法体系。网络引文研究所使用的方法主要继承引文分析方法,以及其他科学研究普遍采用的一般方法,如统计分析、内容分析和域名分析法等,还没有形成领域独特的研究方法。尽管研究方法的探索和形成是一个长期的过程,但是方法上的突破往往会使整个研究领域产生质的飞跃,必须清楚地认识到其紧迫性和在整个研究中的重要性,结合传统引文和链接分析的方法,建立起属于网络引文分析所特有的研究方法和规则体系。

(3)应用实践的深入。目前对于实证分析仅限于少量样本的尝试性分析,而且得出的结论差异很大,缺乏大规模的样本分析和具有普遍性的有意义的结论,当然这与研究方法和手段的不成熟有关。而且用于获取数据的搜索引擎或数据库也存在收录范围和数据稳定性问题,结果的重复性展示也必须解决。网络引文的实证研究应从现有的研究学科推广至整个科学界,并进行多语言和多地区的研究;应该在认识网络引文独特性特征的基础上,克服传统引文研究的局限,积极探索新的应用领域。

参考文献:

- [1] Garfield E. Citation indexing-Its theory and application in science, technology and humanities [M]. Philadelphia: ISI Press, 1983: 12-56.
- [2] Nicolaisen J. Citation analysis [J]. Annual review of Information Science and Technology, 2007 (13): 609-641.
- [3] Borgman C L. Scholarly communication and bibliometrics [J]. Annual Review of Information Science and Technology, 2002 (36): 3-72.
- [4] Zhao D. Citation analysis using scientific publications on the Web as data source: A case study in the XML research area [J]. Scientometrics, 2002, 54 (3): 449-472.
- [5] 杨思洛,仇壮丽. 网络引文研究现状及展望 [J]. 图书情报工作, 2009 (10): 41-44.
- [6] Vaughan L, Shaw D. Web citation data for impact assessment: A comparison of four science disciplines [J]. Journal of the American Society for Information Science and Technology, 2005, 56 (10): 1075-1087.
- [7] Evans J A. Electronic publication and the narrowing of science and scholarship [J]. Science, 2008 (5887): 395-399.
- [8] Larivière V, Gingras Y, Archambault É. The decline in the concentration of citations, 1900-2007 [J]. Journal of the American society for information science and technology, 2009, 60 (4): 858-862.
- [9] Larivière V, Archambault É, Gingras Y. Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900-2004) [J]. Journal of the American Society for Information Science and Technology, 2008, 59 (11): 288-296.
- [10] Egghe L. A model showing the increase in time of the average and median reference age and the decrease in time of the Price Index [J]. Scientometrics, 2010, 82 (2): 243-248.
- [11] Meho L, Rogers Y. Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science [J]. Journal of the American society for information science and technology, 2008, 59 (11): 1711-1726.
- [12] Roth D L. The emergence of competitors to the Science Citation Index and Web of Science [J]. Current Science, 2005, 89 (9): 1531-1536.
- [13] Bar-Ilan J. Some measures for comparing citation databases [J]. Journal of Informetrics, 2007 (1): 26-34.
- [14] Lopez-Illescas C. Coverage and citation impact of ontological journals in the Web of Science and Scopus [J]. Journal of informetrics, 2008, 2 (4): 304-316.
- [15] Gavel Y, Iselid L. Web of Science and Scopus: A journal title overlap study [J]. Online Information Review, 2008, 32 (1): 8-21.
- [16] Lokman I. Assessing the scholarly impact of information studies: A tale of two citation databases-Scopus and Web of Science [J]. Journal of the American society for information science and technology, 2009 (12): 2499-2508.
- [17] Costas R. Variations in content and format of ISI databases in their different versions: The case of the Science Citation Index in CD-ROM and the Web of Science [J]. Scientometrics, 2007, 72

(2): 167-183.

- [18] Whitley K M. Analysis of sciFinder scholar and Web of science citation searches [J]. Journal of the American Society for Information Science and Technology, 2002,53(14):1210-1215.
- [19] Kousha K, Thelwall M. Google book search: Citation analysis for social science and the humanities [J]. Journal of the American society for information science and technology, 2009(8):1537-1549.
- [20] 杨思洛. 网络引文的可获得性体系之构建[J]. 中国图书馆学报, 2009(3):52-58.
- [21] Zhang Y. The impact of Internet based electronic resources on formal scholarly communication in the area of Library and Information [J]. Journal of Information Science, 1998,24(4):241-254.
- [22] Zhang Y. Scholarly use of Internet-based electronic resources [J]. Journal of the American Society for Information Science and Technology, 2001,52(8):628-654.
- [23] Lawrence. Persistence of web references in scientific research [J]. IEEE Computer, 2001, 34(2):26-31.
- [24] Rumsey M. Runaway train: Problems of permanence, accessibility, and stability in the use of web sources in Law Review citations [J]. Law Library Journal, 2002,94(1):27-39.
- [25] David L. The use of the World Wide Web by medical journals in 2003 and 2005: an observational study [J]. Pediatrics, 2007,119(1):53-60.
- [26] Herring S D. Use of electronic resources in scholarly electronic journals: A citation analysis [J]. College and Research Libraries, 2002,63(4):334-340.
- [27] Kushkowski. Web citation by graduate students: A comparison of print and electronic theses [J]. Portal: Libraries and the Academy, 2005(2):259-276.
- [28] Dellavalle R P, et al. Information science: Going, going, gone: Lost Internet references [J]. Science, 2003,302(5646):787-788.
- [29] Oermann M H. Web citations in the nursing literature: how accurate are they [J]. Journal of Professional Nursing, 2008,24(6):347-351.
- [30] Veena R. Web citation behavior in scholarly electronic journals in the field of library and information science [J/OL]. Webology, 2008,5(2). [2009-10-20]. <http://www.webology.ir/2008/v5n2/a57.html>.
- [31] Brown C. The role of Web-based information in the scholarly communication of chemists: Citation and content analyses of American Chemical Society Journals [J]. Journal of the American Society for Information Science and Technology, 2007, 58(13):2025-2065.
- [32] Davis P M. The effect of the web on undergraduate citation behavior: 1996-1999 [J]. Journal of the American Society for Information Science and Technology, 2001(4):309-314.
- [33] Davis P M. The effect of the Web on undergraduate citation behavior: A 2000 update [J]. College & Research Libraries, 2002,63(1):53-60.
- [34] Davis P M. Effect of the Web on undergraduate citation behavior: Guiding student scholarship in the networked age [J]. Portal: Libraries and the Academy, 2003(1):41-51.
- [35] Robinson A. Has the Web really contributed to a decline in the scholarly quality of undergraduate citation behavior? A Canadian case study [J]. Canadian Journal of Information and Library Science, 2003(2):141.
- [36] Deborah J. Worries with the Web: A look at student use of Web resources [J]. College & Research Libraries, 2001(1):11-23.
- [37] Harter. Electronic journals and scholarly communication: a citation and reference study [J/OL]. Information Research, 1996,2(1). [2009-10-20] <http://InformationR.net/ir/2-1/paper9a.html>.
- [38] Spinellis D. The decay and failures of web references [J]. Communications of the ACM, 2003, 46(1):71-77.
- [39] Hester E J, et al. Internet citations in oncology journals: A vanishing resource? [J]. Journal of the National Cancer Institute, 2004,96(12):969-971.
- [40] Wren J D. 404 not found: The stability and persistence of URLs published in MEDLINE. Bioinformatics, 2004,20(5):668-672.
- [41] McCown F, et al. the availability and persistence of web references in D-Lib Magazine. [2009-10-20]. <http://arxiv.org/abs/cs.OH/0511077>.
- [42] Dimitrova D V. Consider the source: Predictors of online citation permanence in communication journals [J]. Portal: Libraries and the Academy, 2006(3):269-283.
- [43] Aronsky D, Madani S, Carnevale R J, Duda S, Feyder M T. The prevalence and inaccessibility of Internet references in the biomedical literature at the time of publication [J]. Journal of the Ameri-

- can Medical Informatics Association, 2007, 14 (2):232-234.
- [44] Eysenbach G, Trudel M. Going, going, still there: Using the WebCite service to permanently archive cited web pages [J]. J Med Internet Res, 2005,7(5):60-81.
- [45] Ducut E. An update on Uniform Resource Locator (URL) decay in MEDLINE abstracts [J]. BMC Med Inform Decis Mak, 2008,8(1):214-221.
- [46] Casserly F. Web citation availability a follow up study[J]. LRTS, 2007(1):42-53.
- [47] Gray K. Web 2.0 authorship: Issues of referencing and citation for academic integrity [J]. The Internet and Higher Education,2008(11):112-118.
- [48] Scopus in Detail: What does it cover? [OL]. [2009-10-20]. <http://info.scopus.com/detail/what/>.
- [49] Van Impe, Rousseau R. Web-to-Print Citations and the Humanities [J]. Informetrics, 2006(8):422-426.
- [50] Vaughan L, Shaw D. Bibliographic and Web Citations: What is the Difference? [J]. Journal of the American Society for Information Science and Technology, 2003,54(4):1313-1324.
- [51] Vaughan L. A new look at evidence of scholarly citation in citation indexes and from web sources [J]. Scientometrics, 2008,74(2):317-330.
- [52] Bauer K, Bakkalbasi N. An Examination of Citation Counts in a New Scholarly Communication Environment [J]. D-Lib Magazine, 2005,11(9):1-8.
- [53] Van T, Beigbeder M. Web co-citation: Discovering relatedness between scientific papers [J]. In Proceedings of AWIC, 2007(42):343-348.
- [54] Kousha K, et al. Using the Web for research evaluation: The Integrated Online Impact indicator [J]. Journal of Informetrics, 2010(4) 124-135.
- [55] Kousha K, Thelwall M. Motivations for URL citations to open access library and information science articles [J]. Scientometrics, 2006,68(3):501-517.
- [56] Kousha K, Thelwall M. How is science cited on the Web? A classification of google unique Web citations [J]. Journal of the American Society for Information Science and Technology, 2007, 58 (11):1631-1644.
- [57] Goodrum A A, McCain K W, Lawrence S. Scholarly publishing in the Internet age: A citation analysis of computer science literature [J]. Information Processing and Management, 2001 (37): 661-675.
- [58] Zhao D, Logan E. Citation analysis using scientific publications on the Web as data source: A case study in the XML research area [J]. Scientometrics, 2002,54(3):449-472.
- [59] Kousha K, Thelwall M. Google scholar citations and google web/URL citations: A multi-discipline exploratory analysis [J]. Journal of the American Society for Information Science and Technology, 2007,58(7):1055-1065.
- [60] Rousseau R. Sitations: An exploratory study [J]. Cybermetrics: International Journal of Scietometrics, Informatics and Bibliometrics, 1997,1(1):4-7.
- [61] Smith A G. A table of two Web spaces: Comparing sites using Web impact factors. Journal of Documentation,1999,55(9):577-592.
- [62] Smith A. G. Web links as analogues of citations [J]. Information Research, 2004,9(4):188.
- [63] Larson. Bibliometrics of the World Wide Web: an Exploratory Analysis of the Intellectual Structure of Cyberspace. [2009-10-20]. <https://sherlock.ischool.berkeley.edu/asis96/asis96.html>.
- [64] Zuccala A. Author cocitation analysis is to intellectual structure as Web Colink Analysis is to... [J]. Journal of the American Society for Information Science and Technology, 2006,57(11):1487-1502.
- [65] Zhao D. Web-based and print journal-based scholarly communication in the XML research field: a look at the intellectual structure [C]. In Proceedings of the American Society for Information Science and Technology 2004 annual Meeting: Managing and Enhancing Information: Cultures and Conflicts. 2004:72-83.
- [66] Brody T, Harnad S, Carr L. Earlier Web Usage Statistics as Predictors of Later Citation Impact [J]. Journal of the American Association for Information Science and Technology, 2006,57(8):1060-1072.
- [67] Di Cesare. The impact of Grey Literature in the web environment: A citation analysis using Google Scholar. [2009-10-20]. www.greynet.org/images/GL9,_page_49.pdf.

杨思洛 湘潭大学公共管理学院知识资源管理系讲师,武汉大学信息管理学院博士生。通讯地址:湘潭。邮编:411105。

(收稿日期:2009-11-23;修回日期:2009-12-27)