●邱均平　李　江

链接分析与引文分析的比较*

摘　要　网络影响因子沿用了期刊影响因子对引文分析的基本思路,但作为链接分析的指标,用于网络环境中的质量评价是不可靠的。可以根据 Pagerank 算法提出用于论文质量评价的 Pagerank 算法;可以根据引文衰减系数提出"链接衰减系数"和"平均链接时距"用于研究网页的老化规律。理想的链接分析工具应当是一种专用搜索引擎。图 2。参考文献 16。

关键词　链接　引文　链接分析　引文分析
分类号　G354.2

ABSTRACT　Network impact factor is used by adopting the basic idea of journal impact factor concerning citation analysis. However, it is not reliable to be used as an indicator for the analysis of link analysis to be applied in the quality evaluation in the networked environment. The authors propose to use Pagerank algorithm to make a Pagerank algorithm for the evaluation of paper quality, and to analyze the aging of web pages by "citation decay index" and "average link time". An ideal link analysis tool should be a special search engine. 2 figs. 16 refs.

KEY WORDS　Link.　Citation.　Link analysis.　Citation analysis.
CLASS NUMBER　G354.2

链接分析,即链接分析法,或称网络链接分析或超链分析,是以链接解析工具、统计分析软件等为工具,用统计学、拓扑学、情报学的方法对链接数量、类型、链接集中与离散规律、共链现象等的分析,用于 Web 网络中的信息挖掘及质量评价的一种方法。

链接分析研究蓬勃开展的同时,因沿用引文分析的理论方法而受到质疑,如网络影响因子这一指标是否合理、商业搜索引擎作为链接分析工具得出的数据是否可靠、用链接分析这一方法评选核心网站是否可行等。概括起来,链接分析研究中有以下含糊之处亟待明确:(1)链接分析沿用了引文分析的哪些理论和方法?这些理论和方法应用于网络环境的可行性如何?(2)链接分析还可以借鉴引文分析中的哪些优势?链接分析又具备哪些优势值得引文分析借鉴?(3)理想的链接分析工具是什么样的?

我们将在链接分析和引文分析的多角度比较中探明这些问题。

1　链接与引文的比较

引文分析中的"引文"通常指文献结尾处的参考文献,不含脚注、间注、夹注等,现有的引文分析工具也不提供脚注、间注、夹注的查询。链接分析中的"链接"通常指超链接,即一种文件指针,简单的声明文档或信息之间的关系[1]。引文和链接有相似之处:引文和链接群体形成的客观结构都是网状结构,都可用结点和有向箭头来刻画;引文和链接都含有主观动机,都是主体带有目的的行为。也正是在这两点上,两者又有很多各自特点,不可盲目借鉴。

1.1　链接与引文的网状结构比较

"引用"将施引文献与受引文献连结在一起,数量较多时便形成一种网状结构。如果用结点代表文献,用有向箭头代表文献之间的引用关系,描述在时间上,就可以绘出"引用时序网络图"如图 1。其中存在大量的同引结构和耦合结构[2](同引指两篇及两篇以上文献共同被后来的一篇文献或多篇文献所引用,耦合指两篇文献共同引用了一篇或多篇文献)。

"链接"将施链和被链网页连接在一起,数量较大时,也形成一种网状结构,按同样的方法,可以绘出"Web 图"[3],如图 2。其中同样存在类似于同引和耦合的结构——共链("共入链"类似同引,"共出链"类似耦合)。

* 本文系国家自然科学基金项目(编号 70673071)的研究成果之一。

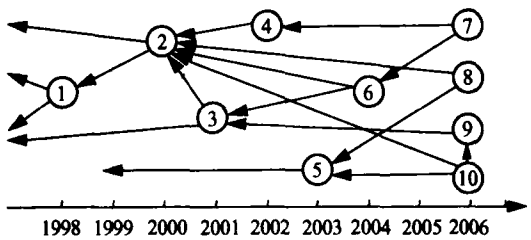

图1　引文时序网络图

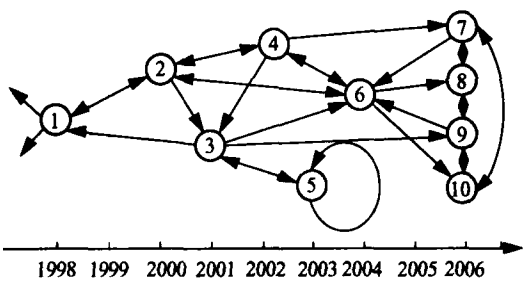

图2　Web 图

比较引文时序网络图和 Web 图,可以发现链接和引文的结构有 5 个特点。

(1)引文是静态的,链接是动态的。图 1 中不可在任何已有结点上增加新的施引的单向箭头,即文献一经出版,它的参考文献就一成不变。但图 2 中可以在任意结点上增加入链的单向(或双向)箭头,即网页的链接可以随时增、删、改。

(2)引文只能是单向的,链接可以是双向的。引文网络图是有时间方向的,只能是后期的文献引用前期的文献,而 Web 图没有时间方向,后期和前期的网页可以互链。

(3)引文中无自引,链接中可以自链。引文分析中的"自引"主体只能是作者、期刊、学科、机构等。文献的引文中不可能包含文献自身,而网页的链接中可以包含该网页自身。

(4)引文网络图中,箭头带有固定时间,即文献 A 引用文献 B 是有固定时间的。这个固定时间正好是文献 A 的发表时间,链接结构图中则没有。

(5)引文呈现出主题集中,链接则呈现出主题发散。引文的正式性和文献出版的质量控制使得引文基本属于同一科学领域或关系紧密的领域,而网页的发布具有随意性以及链接的随意性等使得链接的内容很可能互不相关。

1.2　链接与引文的动机比较

科学文献的引用和被引用说明了对科学知识和情报内容的继承与利用,但引用动机比较复杂。按照 Garfield 的研究,文献引用的动机大致有"对开拓者表示尊重"、"对有关著作给予荣誉"、"核对其所用的方法及仪器"等 15 种情况[4]。1986 年,Brooks 根据前人的研究,将引文的动机分为 7 类[5]:新颖性(Currency)、负面证据(Negative Credit)、操作型信息(Operational information)、说服(Persuasiveness)、正面评价(Positive Credit)、提醒(Reader alert)、社会认同(Social consensus)等。1994 年,Baird Oppenheim 在 Garfield15 种引文动机的基础上提出了"受作者的师长影响而引用"、"在不慎重的情况下引用"等 17 种引文动机[6]。无论是 Garfield 的 15 种引文动机,还是 Brooks 的 7 种引文动机,还是 Baird Oppenheim 的 17 种引文动机,引文都是有用的,即便是负面证据,也是值得作者关注的问题,才会被作者不惜笔墨地批评、反驳并引用。

2000 年,Hak Joon Kim 将网络中学术论文的链接动机归纳为 3 类(Scholarly, social, Technological)共 19 种[7]。网络中的链接绝不止学术论文链接一种,事实上,其他类型页面的链接比学术论文的链接动机更复杂。袁毅针对整个网络上的各种类型链接,提出了 9 种链接动机[8]:推荐链接、相关链接、引用链接、扩展链接、评价链接、关系链接、服务链接、通讯链接、结构链接。推荐链接、相关链接、引用链接通常可以用于对网站质量的评价,排除商业因素,肯定的评价链接也具有质量评价功能,服务链接也具有一定的评价功能,这些链接属于实质性链接,而其他链接类型都违背了链接分析假设前提。

引文动机基本符合引文分析假设前提,这也是引文分析得以成功发展的原因。相比之下,链接分析中,大量的链接是滥竽充数的,势必影响链接分析结果的有效性,这也是为何链接分析的可靠性受到质疑。运用链接分析这一方法时,应从链接数中剔除非实质性的链接,这也是当前链接分析需要解决的难题之一。

2　链接分析与引文分析的比较

引文分析研究的发展过程中,人们先后提出若干概念:Grace 等人的核心期刊表(1927 年),Eugene Garfield 的"Citation Indexes"论文(1955 年),Brown 对引文分析领域的拓展(1956 年),Kessler, M. M 的"文献耦合"(1963 年),Eugene Garfield 的 SCI 印刷版(1964 年),Small 提出"同被引技术"(Co-citation)

(1973年),SCI网络版(1997年)等等。当前的研究主要集中在:方法适用性研究(引文分析的弊端等);网络引文分析(Web Citation Analysis);应用研究(引文应用于各类质量评价,专利引文分析,同引、耦合用于聚类分析,大学评价等)[9]。

链接分析的发展过程中有这些概念:McKiernan提出sitation(1996年),Larson的共链分析(1996年),Almind和Ingwersen的"Webmetrics"(1997年),Peter Ingwersen的"网络影响因子"(1998年),Sergey Brin和Lawrence Page提出"Pagerank算法"、J. Kleinberg提出"HITS算法"(1998年)等等。当前的研究主要集中在:链接分布规律研究(包含链接类型分布、链接数量分布等);网络影响因子研究;网络链接分析工具研究;沿用引文分析和方法的可靠性研究;链接分析应用研究(包括在网络信息检索中的应用、在网络社区发现中的应用、在Web拓扑结构建模中的应用、在信息挖掘中的应用——资源发现、竞争情报获取等)[10]。

事实上,从"citation"到"sitation",从"期刊影响因子"到"网络影响因子",从"文献的同引与耦合"到"网页的同引与耦合",都表明链接分析带上了引文分析的烙印。然而,网络与期刊文献毕竟有太多不同,我们从假设前提、测度指标、工具3个深层次的角度将链接分析与引文分析作比较。

2.1 链接分析与引文分析的假设前提比较

对引文的分析和对链接的分析都基于一定的假设前提,不同假设基础上的测度指标适用范围不同,应用也不同。

引文分析主要有以下5条假设前提[11]:(A1)参考文献体现了作者使用过的全部最重要的文献;(A2)参考文献都是优质文献;(A3)文献被使用情况体现文献的价值;(A4)被引文献与引用文献内容相关,两文献同引或耦合,则两文献内容相关;(A5)所有引文都是等价的。

M. R. Henzinger认为链接分析的思想基于两条假设[12]:(B1)从网页A指向网页B的超链是网页A对网页B的推荐或认可;(B2)如果一条超链接将网页A和网页B链接起来,则网页A和网页B可能有共同的主题。陈定权综合了6位外国学者的研究[13],将上面的两个基本假设引申为以下假设:(B3)一个网页被多次引用,即很多网页有指向它的链接,则这个网页很重要;(B4)一个网页尽管没有被多次引用,但被一个重要网页引用,则这个网页很重要;(B5)一个网页的重要性被均匀分布并传递到它所引用的网页;(B6)如果网页p和q同被引,则它们可能是相关的,同被引强度越大,相关度越大;(B7)如果网页p和q耦合,则它们可能是相关的,耦合强度越大,相关度越大。

(B4)和(B5)以Pagerank算法为背景,认为不同的链接是不等价的。与此相对应的另一种假设就是沿用了引文分析假设(A5),即所有的链接都是等价,这一假设的一个典型实例便是"网络影响因子",在链接分析实践中,这两种假设是并存的,这也是链接分析方法不成熟的一个典型表现。事实上,各种各样的引文和链接的动机都将导致引文分析和链接分析的假设不成立。在此,我们以所有的假设都成立为基础对二者作比较。

(1)引文和链接都表示认可,且次数越多,认可程度越高;引文和链接中的同引和耦合都表示内容相关,且同引和耦合程度越大,内容相关程度越大。

(2)引文分析假设"参考文献体现了作者使用过的全部最重要的文献",因为如果"有引用,无引文(参考文献)",则引文分析的结果失真。但链接分析的假设中没有这一项,因为很难有"有推荐(或认可),无链接"(如果将所有的链接都改为文字表述,则各网页之间无法构成网络)。这一点正好印证了链接与引文的机理不同,单从这一点上看,链接分析的结果比引文分析的结果更客观。

(3)引文分析假设所有引文都等价,这显然是不合理的。不同质量的页面所给的入链显然是不等价的,质量高的页面的"推荐"显然分量更重、价值更高。基于此,我们认为陈定权的假设(B4)和(B5)的描述是相对更合理。

比较完了引文分析与链接分析假设,根据以上3点,我们对引文分析假设做一个大胆的修正,即将假设(A5)的内容改为:"一篇文献(或一种期刊)尽管没有被多次引用,但被一篇重要的文献(或一种重要的期刊)引用,则这篇文献(或这种期刊)很重要;一篇文献的重要性被均匀分布并传递到它所引用的文献"。这样一来,传统的期刊影响因子便不再适用了,我们构想以Paperank算法(类似于Pagerank算法)取而代之,这种算法不仅可以更有效地评价期刊质量,也可更有效地评价论文质量和学者的学术影响。

2.2 链接分析与引文分析的测度指标比较

根据引文分析指标的应用,我们将它分为以下几

类:(1)引文数量与分布规律测度指标:引文数、平均引用数、自引数与自引率、被引用数与引用数的比值;(2)期刊质量测度指标:被引用数、影响因子、即年指标;(3)论文质量与著者学术水平测度指标:被引用数;(4)文献老化规律测度指标:衰减系数。

参照文献段宇锋的研究[14],根据同样的分类方法,我们将链接分析测度指标分为以下几类:(1)链接数量特征测度指标:总链接数、出链数;(2)链接分布特征测度指标:链接密度、页面平均链接数;(3)网站影响力测度指标:入链数、网络影响因子;(4)网页重要性测度指标:Pagerank 算法、HITS 算法。

通过链接分析与引文分析测度指标的比较,可以看出,链接数量和分布规律的测度指标和网站影响力测度指标都类似于引文分析中的测度指标,如:自链数类似于自引数,出链数类似于引文量,页面平均链接数类似于平均引用数,入链数类似于被引用数,网络影响因子类似于期刊影响因子等。

事实上,链接分析中的有些指标对引文分析指标的沿用是不合理的,像入链数、网络影响因子等,这些指标用于评价网站影响力是不可靠的,原因为:首先,引文是静态的,而链接是动态的,可以随时对链接增、删、改,入链数和网络影响因子作为质量评价指标,必然导致极大的互链驱动。其次,引文链接的动机不同,链接中大量的非实质性链接在统计中仍不能有效剔除。第三,假设前提不同,入链数和网络影响因子符合引文分析中的假设(A5),却违背了陈定权的假设(B4)、(B5)。

被引次数这一指标用于文献质量评价缺乏说服力,Pagerank 算法、HITS 算法等网页重要性测度指标在网络信息检索中的成功应用启发我们借鉴其优点以替代被引次数。

引文分析指标中用衰减系数来描述文献的老化规律,而链接分析中暂无这一指标,通过对链接结构的分析,我们先提出两个前提:第一,在网络信息组织中,链接行为出现时,系统自动给链接标上时间,如同引文中的引用时间;第二,链接可以随意增、删、改,我们假设所有存在的链接都是有意义的,即无意义的链接全都会被删除掉。基于这两个前提,我们提出"链接衰减系数"这一概念,计算公式为:

链接衰减系数 = 近 m 个月的链接数/所有链接数

其中 m 可为整数 3、6、12 等,具体数值可根据链接时间分布特征确定。同样,在这两个前提的基础上,可以提出另一个用于描述文献老化规律的指标"平均链接时距",即所有链接时间与统计时间差距的平均值,计算公式为:

$$\bar{D} = D_0 - \frac{1}{n}\sum_{i=1}^{n} d_i$$

其中,$\bar{D}$ 表示"平均链接时距",D_0 表示"统计时间",d_i 表示"第 i 个链接的链接时间",单位均为年,精确到小数点后第二位,如统计时间为 2006 年 10 月 26 日,可表述为"2006.82",第一个链接的时间为 2005 年 10 月 26 日,可表述为"2005.82",这样便可得出单位为年的平均链接时距值。这一指标可用于描述网页的老化规律,如,以"sina 博客"这一主题为例,搜集所有 sina 博客,统计出链接日期平均值,即可算出链接衰减系数,这一指标值意味着 sina 博客的活跃年限,如果再统计出"sohu 博客"、"yahoo 博客"的链接衰减系数,便可作出横向比较,分析出哪一个门户网站的博客最为活跃。

2.3　链接分析与引文分析的工具比较

国外的引文分析工具主要有:SCI、SSCI、A&HCI(提供 Citation Index, Source Index, Corporate Index, Permuterm Subject Index 4 种索引);JCR(Journal Citation Report,提供 Journal Rankings, Source Data Listing, Journal Half-Life Listing, Subject Category Listing, Citing and Cited Journal Listing, Journal Title Abbreviations 6 种评估指标)。国内的引文分析工具主要有:CSCD(Chinese Science Citation Database,统计分析功能有:科学论著被引频次、著者被引频次、著者发文量、机构发文量、地区或国家发文量);CSSCI(China Social Science Citation Index,提供多种定量数据:被引频次、影响因子、即年指标、期刊影响广度、地域分布、半衰期等)。

初期的链接分析研究大部分都以搜索引擎(如 Google, Alltheweb, Altavista 等)为工具,因为它们提供了诸如 link、domain、host 这样的检索指令,可用来统计链接分析部分指标值,一些研究确实也取得了阶段性的重要成果,但对于聚类等复杂的链接现象仍缺乏有效的工具。搜索引擎对网络信息的覆盖率、检索结果的准确性、搜索引擎的不一致性等问题对网络链接研究也直接造成了严重影响。商业搜索引擎毕竟并非专门为链接分析而设计。除了商业搜索引擎,还有 Check Web、Webstat、Citeseer 和 Link-Agent 等软件用作链接分析工具,这些软件或者为商业目的制作,或者为少数学者自己的研究量身定做,通常用于链接分析的功能简单、不齐全,这类工具至今仍是小范围使

用。总的来说，当前没有能让研究链接分析的学者们满意的链接分析工具。

比较链接分析工具和引文分析工具，我们认为，理想的链接分析工具应该是一种具备以下特征的专用搜索引擎：像引文分析工具一样收录某一特定领域的页面，并应尽量全面；具备SCI等工具的索引功能（如链接索引等），具备JCR的指标值统计功能（如站链接总数、入链数、出链数、链接密度、链接衰减系数、Google PR值等）；能够有效解决检索过程中的"不一致性"问题；能有效识别实质性链接和非实质性链接。

3 结论

（1）网络影响因子沿用了期刊影响因子对引文分析的基本思路，但作为链接分析的指标，用于网络环境中的质量评价是不可靠的。

链接分析对引文分析理论和方法的沿用包括：①研究对象：链接数量分析、链接类型分析、链接离散规律分析、链接的共链与耦合分析等；②测度指标：自链数、网络影响因子等；③研究方法：链接分析的"样本的选择——样本原始数据的获取——网络链接的解析——数据的统计分析"[15]类似于引文分析的"选取统计对象——统计引文数据——引文分析——作出结论"[16]。链接分析沿用引文分析理论在网络结构中的特征和规律分析、网页聚类等方面是可行的，以类似于引文分析的研究步骤也是可行的，但网络影响因子用于质量评价，其结果是不可靠的，原因在于：链接与引文的网状结构不同；链接的动机与引文的动机不同；链接分析与引文分析的假设前提不同。

（2）可以根据Pagerank算法提出用于论文质量评价的Paperank算法；可以根据引文衰减系数提出"链接衰减系数"及"平均链接时距"用于研究网页的老化规律。

链接分析虽沿用了引文分析中的部分理论，但并未束缚于其中，链接网络结构以其自身的特点孕育出了Pagerank算法等惊世之作。Pagerank算法用于Google搜索引擎取得了巨大的成功，作为链接分析中的闪光之处，其思路可用于引文分析。如果按这一思路提出Paperank算法，则可用于论文质量评价，而且如果这一指标用于电子期刊数据库的检索结果排序，取代其原有的"相关性排序"，将会给用户带来方便。"链接衰减系数"源于文献老化规律分析中的"衰减系数"，"平均链接时距"也是在这一概念的启发下提出，二者均可用于分析网页的老化规律。

（3）理想的链接分析工具应该是一种具备以下特征的专用搜索引擎：具备Google的网页覆盖率（只需元数据内容，无需全文）；具备SCI等工具的索引功能（如链接索引等）；具备JCR的指标值统计功能（如站链接总数、入链数、出链数、链接密度、链接衰减系数、Pagerank值等）；能够有效解决检索过程中的"一致性"问题；能有效识别链接类型。

参考文献

1,14,15 段宇锋. 网络链接分析与网站评价研究. 北京：北京图书馆出版社，2005
2 党亚茹. 引文网络系统的结构模型化. 图书情报工作，1996(4)
3,13 陈定权. 自动主题搜索的应用研究：[博士学位论文]. 北京：中国科学院文献情报中心，2003
4 Eugene Garfield. Can citation indexing be automated. 1965. [2006-10-26]. http://www. garfield. library. upenn. edu/essays/V1p084y1962～73. pdf
5 Terrence A Brooks. Evidence of Complex Citer Motivations. Journal of the American Society for Information Science (1986～1998). 1986,37(1)
6 Baird. Do citations matter. Journal of information science, 1994,20(1)
7 Hak Joon Kim. Motivations for hyperlinking in scholarly electronic articles: A qualitative study. Journal of the American Society for Information Science. 2000,51(10)
8 袁毅. 核心网站评选的理论与方法. 北京：北京图书馆出版社，2005
9 彭爱东. 专利引文分析在企业竞争情报中的应用. 情报理论与实践，2004(3)
10 Kousha, K. & Thelwall, M. (2005). Motivations for URL Citations to Open Access Library and Information Science Articles. Scientometrics, 2006, 68(3)
11 罗式胜. 文献计量学概论. 广州：中山大学出版社，1994
12 Monika R. Henzinger. Hyperlink analysis for the web. IEEE Internet computing, 2001, 5(1): 5～50
16 邱均平. 文献计量学. 北京：科学技术文献出版社，1988

邱均平 武汉大学信息管理学院教授，博士生导师。通讯地址：武汉。邮编430072。
李 江 武汉大学信息管理学院2005级情报学硕士。通讯地址同上。

（来稿时间：2007-01-05）