中国科技资源导刊 ISSN 1674-1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2008年5月 第40卷第3期 26-31 ISSN1674-1544 Vol.40 No.3 26-31,May, 2008

国内外链接分类研究综述

董 珏 李 江
(南京大学信息管理系,江苏南京 210093)

摘 要:入链数、网络影响因子(WIF)、Pagerank 算法等链接指标是国内外链接分析研究的热点,主要用于网页/网站的重要性评价。这些链接指标都有一个共同的假设前提——链接代表认可、推荐。然而多种统计表明,符合这一假设前提的链接不超过链接总数的 30%,这便引起了国内外对链接分类的广泛研究。本文根据近年国内外的相关文献,综述了链接分类的研究现状,并提出研究中的不足之处与发展方向。

关键词:链接分析;链接类型;网络信息计量学
中图分类号: G353.1 **文献标识码**: A DOI: 10.3772/j.issn.1674-1544.2008.03.005

Study on the Reliability of Link Analysis—Study of the Link Types

Dong Jue , Li Jiang
(Department of Information Management, Nanjing University, Nanjing, 210093)

Abstract: Link indicators, such as inlink count、WIF、Pagerank algorithm, which have been uesed to evaluate the web page and site's influence, is an important question in Link analysis study. These indicators are all based on the hypothesis that web links represent authorization or recommendation. But many reports showed the number of the links accord with the hypothesis coverd only 30% of the total web links. This paper based on the related articles both home and abroad, summarized the current researches on link classification, pointing out the shortages, exploring the orientation.

Keywords: link analysis, link types, webometrics

Mc Kiernan(1996)首先提出 Sitation 这一新术语,用以研究网页之间的引用关系。Almind、Ingwersen、Roussea 等学者的研究成功地建立了网络信息计量学和传统文献计量学之间的联系[1]。Harter 和 Ford(2000) 分析电子期刊的被链接情况发现网络链接中的自我链接对于评价学术价值是没有意义的,站外链接对文献的引用大部分也是间接引用,引文分析和链接分析的结果不具有相关性。Mike Thelwall[2] 的研究也证实链接次数并不能反映其学术价值。在链接分析时选用不同的工具也会造成结果的极大差异。此外,国内学者也开始对链接分析的可行性进行研究,有针对性地对链接类型进行分析。总体而言,国内外从事链接分析研究的学者不得不面对这样一个问题:链

第一作者简介:董珏(1984-),女,湖南溆浦人,研究生,研究方向是网络计量学。

收稿日期:2008 年 4 月 12 日。

接数、站外链接数、网络影响因子、PageRank 算法等链接指标是否能体现出“认可、推荐”的价值？因为多种统计表明，符合“认可、推荐”的链接不超过链接总数的 30%[3]。链接分类研究的意义在于：从链接总体中剔除不符合“认可、推荐”的链接，以提高入链接、网络影响因子、PageRank 算法等链接指标用于网页/网络重要性评价的效率。本文根据近年国内外的相关文献，综述了链接分类的研究现状，提出研究中的不足之处，并分析了研究方向。

1 国外链接分类研究

最早从 Woodruff (1995)开始，在网络链接的计数研究中将链接分为自我链接和外部链接两类进行计量分析。Robert C. Vreeland 根据链接靶向的差异，将与网站相关的链接分成该网站指向其他网站的链接 (luminosity) 和其他网站指向该网站的链接(visibility)[4]；Peter Ingwerse 将指向特定网站和网页的链接 (inlinks 或 inboundlinks) 分为来自外部的链接(external citations)和来自内部的链接(self - citations)[5]；Trigg R. 将网站指向自身内部的链接分为普通链接和评论性链接。之后出现了大量的链接计数在网络影响因子中作为重要指标的实证研究。有学者提出网络上的链接有可能只是一种标记甚至没有任何功能，因此链接的动机要比电子期刊引用动机复杂得多。网络不是一个质量可控的参考资源，信息可以被任何人在网上进行发布。单从简单的入链、链接总数上的统计已经不能满足对网络文献影响力的评价需要。

1.1 链接行为动机研究

在网络链接动机的研究中，Hak Joon Kim[6] 将链接行为的动机分为 4 类：(1)学术。提供附加信息和背景信息；提供范例；证实和支持论点；对所阐述的问题以图像等直观的方式予以表明；对术语、概念、符号下定义和解释；提供历史背景；提供有关设想、概念和理论；与自身进行比较；提供数据和统计信息；表明目前的研究状况并确定目前研究所处的阶段；描述所使用的方法。(2)社会。公开和宣传被链接的信息；对作者和机构荣誉的归属；被链接信息的作者或机构在该领域的显赫声望；证实自己掌握该领域最重要和最新的研究成果；编辑或编辑政策鼓励使用超链接。(3)技术。为读者提供便捷的存取方式，因为被链接的电子信息资源能以链接的方式获得而使用链接。(4)价值增值。将其与传统引用动机比较，除了与传统的引文动机性质相似的学术和社会动机外，Kim 的结果表明有 39% 的链接至少是基于技术动机而使用的，这对资源引用的影响力评测价值不大。

另外，Chu[7] 还分析了指向学术网站的链接，发现链接的 50% 是指向资源或者目录信息，只有 27% 的链接动机是研究或者教学/学习。由于网络链接的广泛动因以及网络链接目标文献的复杂性，网络链接与引用的动机是不同的。Wilkinson 等人[8] 发现链接到大学网站的链接中只有不到 1% 的链接是正式的研究引用。Smith[9] 采用分类和随机抽样的方法，分析大学网站外部链接的来源网页和目标网页的类型，得到的结果表明，在研究选用的对象中，20% 的链接与引用相似。Thelwall[10] 研究证实了大学网站链接与大学地理距离之间的相关性，从侧面反映了地理距离对网站链接创建的潜在影响。而国内在利用间接相关分析研究潜在的链接动机方面还是空白。

Mike Thelwall 将电子期刊中的引用链接从广泛的网络链接中区别开来，提出电子期刊引用与普遍的网络资源链接是有区别的，可以根据上下文的情况来研究电子期刊中的引用行为，在调查链接的创建动机时，可以将注意力限定在一类特定的网页或网站上。

Scharnhorst, A.[11] 总结了几种与传统引文的作用近似的链接形式：(1) 以文档间的相互联系形式出现的超链接，Rousseau 将它归于“Sitation”，但 Mike 在此将这个术语描述为超链接或链接；(2) 在电子期刊中的超链接比一般链接更接近于“引用”这个功能，被定义为“文献超链接”；(3)在网页中出现的文献不一定是期刊论文，还有来自灰色文献的引用，这也是一种引用关系，但不同于传统引文，因为他们的源文献不是一个被参考的论文或会议论文，这种链接被定义为“网络引

用";(4)常规的引用可以从电子期刊论文和学术期刊或会议论文的电子版本中的参考文献部分找到。他指出 Kim 的调查发现了一些新的引用电子资源的动机,比如对于在线信息的可获得性以及引用数字内容的能力等,但很多其他的调查直接推论出链接类型的创建动机可以通过在源页面中诠释链接上下文以及通过访问链接目标页面的形式来建立。

1.2 链接上下文分析

学者们通过分析上下文内容(主要是链接源网页和目标网页的内容性质)对链接进行分类的研究,最早由 Cronin、Snyder、Rosenbaum 等人(1998)开始。Cronin 等调查了那些被高度引用的学术机构网页的上下文环境,将网页分为 11 类,发现这些学术机构的名字出现在不同的上下文关系中,如课程阅读清单、最近的资料通告、资源指南、个人或机构主页、邮件或目录等。后来的学者也针对不同的研究对象进行了研究(表 1)。

表 1 根据网页间内容关系(上下文关系)的分类研究举例[12,13]

学者	研究对象	分类
Borgman 等	OA 期刊论文(2002)	导航链、所有权链、社会链接和没有用的链接
AlastairG Smith	大学、研究机构或个人、电子期刊网站(2003)	非实质性研究、实质性研究(包括一般信息链接、正式研究引用、支持赞助商/鸣谢、关于链接创建者的自链接、相关网页、地理信息、广告、软件下载 9 类)
Bar Ilan	大学网站间互链接(2004) 学术机构网站(2005)	面向研究的、教育相关的、职业或工作相关的、行政管理的、一般信息的、个人的、社会的、技术的、导航的、表面的、其他和无法定义的
Jepson	商业引擎搜索学科主题(2004)	科学性质的、与科学相关、教学、低质量的、"噪音"
Heting Chu	学术机构入链	服务、主页、研究、教学

其中,最具代表的是 Alastair G Smith(2003)[14],他将网络链接分为实质性和非实质性两类,将链接的来源网页与目标网页的内容特征进行分析,从而研究链接的目的,考证了真正直接指向信息资源的链接才能用来计算 sWIF。Alastair 研究的网站类型主要是大学、研究机构或个人、电子期刊网站,其将网页的内容分为 11 类,指出源网页的主要类型是名录和主题指南(48%),其次是参考书目/出版物列表以及一般信息来源、正式出版物(各占 9%);目标网页主要是链向机构主页(36%),而站点/页面入口性质的目标网页占了 67%,其次是涉及信息目录、名录/出版物列表、正式出版物内容的网页。统计出的链接目标仅有 10% 是出于正式的研究引用。类似研究还有 Bar Ilan(2004,2005)。

Bar Ilan(2004,2005)、Jepson 等(2004)、Harter 和 Ford(2000) 使用了内容分析和预先确定的分类表来检测链接动机。在 Harter 等人的研究中将 294 个链向期刊网站的链接分为 13 类,指出超过一半的链接是来自具有指向相近主题网络资源的网页,7.8% 是来自电子期刊论文和会议论文或介绍的链接,而只有这 7.8% 才是等同于引用的。Borgman 和 Furner(2002) 的研究指出只有社会链接是对引用有影响的。Thelwall 等(2003) 也通过对学术网站的主题分类发现需要有附加的研究来保证结论的有效性和可靠性。

这些研究虽然有效但也受到了挑战,Wilkinson 等 (2003) 的研究显示,不同的分类者得出的类型体系有很大的差异。以上诸多的实证分析没有反复核对过网络链接的分类,即他们预先订好的分类表不一定可靠。因此在后续的研究中学者们都采用了一个动态的分类体系,即在研究中不断地修正其分类。Vaughan 和 Shaw(2003)就将其在研究过程中发现的来自其他文献发布的引用和课堂阅读清单这两个子类加入其使用的对引用源的类型中(期刊、作者、服务、课堂、文献、会议和其他)。Thelwall 等也都采用了该类方法,将在研究中发现的子类加入到其预先订好的大类中去。

通过对链接上下文环境的研究划分出的链接关系在网络链接算法的应用中得到了体现,例如,Chakrabarti[15] 等就结合网页的超链接信息和文本信息对网页进行分类,提高了算法的精确度。在主要依靠 Web 链接结构信息分析的应用领域中,对 Web 链接结构信息的发掘必须在某种程

度上加入对特定网页信息内容的分析，以提高其分析的精确度。例如在分析中引入网页文档的标题信息和超链接上的标题信息等。

2 国内链接分类研究

根据链接的方向，链接可分为入链、出链、相互链接；按存在范围可分为站内链接、站外链接、系统间链接。在最初的链接分析研究中，这两种划分方法是最常用也是最原始的分类体系。例如，网络影响因子的计算就是采用了入链与链接总数的比值。然而，正如在第二部分的论述中提到的，这种粗略的分类不能合理地反映学术研究机构网站和电子期刊的影响力，从而对链接的动机和语义环境进行了类型划分，以找出最能反映资源间引用情况的链接，合理地测评网络影响力。

2.1 链接关系分析

国内学者结合传统引文动机的理论基础，通过一些实际调查提出了关于链接关系的分类体系，主要以邱均平和俞培果以及刘雁书和方平的研究为代表(表2)。后者主要是对站外链接关系进行的特征分析，而邱均平则是对内部链接和外部链接进行了全面划分。

在目前的研究中，国外的研究在很大程度上集中于对某一类学术研究机构或电子文献资源的链接引用分析，在这个研究范围内发现链接的动机研究与电子的或非电子的引用动机研究有着相似的理论基础，可以将链接计数作为影响力的指标，并且从具体的数据中得到体现。国内的学者则从一个较为广阔的网络环境出发，全面地提出网络链接的不同创建动机。从理论上说，二者都来源于传统的引文动机分析，并适当做了扩展。这些研究为在进行链接分析时提取最有易于学术价值测评的统计数据提供了理论依据。

2.2 链接功能属性分析

单从行为认知上对链接进行划分是不够的，网络链接具有与传统引文不同的技术手段和方法，深入地了解其结构属性，有利于开发出合理的链接分析工具，从而更科学地进行链接分析研

表2 国内代表性链接关系分类

研究学者	研究对象	类型	动机
邱均平等[16]	内部链接	网站结构链接	体现网站结构和层次关系
		信息关联链接	相当于参考文献和相关主题的链接
	站外链接	信息推荐链接	推荐相关内容网站，有的是商业目的
		信息来源链接	标明信息来源，表明知识产权、责任归属
		网络结构链接	根据访问目的选择网站，方便快速访问
刘雁书等[17]	站外链接	推荐链接	正反面引用
		合作链接	引用服务，主办单位，信息来源，内容相关
		相关链接	反映内容相关程度
		资源链接	链接被链网页的某种资源
		通讯链接	接到通讯服务
		广告链接	商业广告，服务相关，个人网站资助性广告
袁毅[18]	学术网站	推荐链接	肯定性链接
		相关链接	内容相关，利用关系
		引用链接	内容引用，反映高质量网站
		扩展链接	背景资料，注视、数据链接
		评价链接	肯定或否定评价
		关系链接	机构间纵向、横向、利益链接，用户链，背景链，合作链
		其他	服务链、通讯链、结构链

究。网络链接通常由节点、热标、链3部分构成[19]，结合这3部分的类型就能给链接进行详细的分析。

目前，可以通过合适的算法对链接的结构进行分析，实现网页的聚类、内容分析等，从而确定某一领域的核心网站以及主题内容。它是分析网络信息组织的重要手段。例如Joslyn，基于图形的超媒体系统将链接分为推理链和导航链。基于国外学者的研究，国内的学者进行了拓展(表3)。

另外，按链接的方式可将超文本系统中应用的链分为实链和虚链；根据链的端点分为双、多、单和无端链；根据链接关系分为自我链接、同被链接、链接耦合等。这些分类主要基于链接的结构分析，就引用分析而言，邱均平等的分类更贴近我们的分析需要。这种分类有利于我们了解网络信息的组织形式，确定链接的有效性，从而评价网页的质量，进行网页内容聚类分析，优化统计算法，是网络信息计量学的重要研究领域。

3 链接分类研究发展方向

文献[22]指出在链接创建时，与网页内容有关的动机统称为链接创建的直接动机，与网页内容无关的动机统称为链接创建的潜在动机。国外的链接动机分析采用定量与定性结合的方法计算出了不同链接动机的比例，而国内的研究则仅仅定性地分析出链接的类型，在实证研究方面做得很少。然而，从上述的分析比较可以看出，国外的链接分类主要从链接的引用行为目的进行划分，借鉴引文动机的分类体系构建链接的分类体系；而国内从更广泛的意义上对链接的目的或结构功能进行分类研究。当前，国内外的链接分类研究主要依赖于小样本的主观分类，而对于海量链接，依靠主观分类是不现实的，所以，下一步的研究方向应是开发一套自动的链接分类方法（“链接识别”[3]）。基于国内外链接分类研究的相关文献，针对链接分类研究中的不足之处，笔者认为链接分类的发展方向包括以下3个方面：

(1) 构建合理的分类体系。在引文分析的研究中，Chubin和Moitra等人就曾提出过一个具有相互排斥作用的分类结果。借鉴这个思想，我们也应对网络链接中有实质意义的链接进行合理分类。在上述研究结果中，大多数体系存在着严重的类别交叉重复，且各家自成一体，给研究造成了很大困难。我们赞成将链接分为实质性与非实质性链接两大类，并重点对非实质性链接进行重点研究。

(2) 动态分类过程。以上诸多的实证分析没有与网络链接的分类核对，不知他们预先制订好的分类表是否可靠。Wilkinson等(2003)的研究显示，不同的分类者得出的类型体系有很大的差异。因此，在后续的研究中学者们都采用了一个动态的分类体系，即在研究中不断地修正其分类。Vaughan和Shaw(2003)就将其在研究过程中发现的来自其他文献发布的引用和课堂阅读清单这两个子类加入其使用的对引用源的类型中(期刊、作者、服务、课堂、文献、会议和其他)。Kousha和Thelwall等也都采用了该类方法，将在研究中发现的子类加入到其预先订好的大类中去。

表3 国内根据功能属性对链接的分类

研究学者	类型	子类或说明
邱均平 黄晓斌[20]	导航链	目次、注释、实例、索引、扩展、相关、应用、等价、引用、评价、分解、版本、聚合
	执行链	通过热标跟应用程序相连，可激发一个操作
	类型链	允许用户描述两个节点的关系，可以定义链的类型
	推理链	Is－a链、Has－a链、蕴涵链
	自动链	与相似主题或满足条件的节点自动连接
张海涛 刘甲学 宋川[21]	基本结构链	基本链、交叉索引链、节点内注释链
	推理链	索引链、IS－a链、Has－a链、蕴含链、执行链、自动链接、类型链
	导航链	迁移链、放大链、动链、视图链

(3) 链接自动识别。计算机根据链接源页面和目标页面之间的关系，自动识别链接类型的过程。这主要依靠有效的链接分析工具来进行。除了有效地识别出内链、外链和死链外，还要能根据研究者的定义识别出实质性与非实质性链接。目前常用的方法是对 URL 进行特征分析，对实质性链接进行类型和内容分析。这也是网络信息计量学中的研究难点和重点之一。

4 结 语

目前，链接分析在很多方面都存在着缺陷。例如网络引用行为的复杂性造成的计量问题、评价指标的全面性以及分析工具和算法的改进等。笔者认为，任何一项研究首先应当对其研究对象进行透彻的分析，开发出合理的分析工具，改善研究方法，找到有力的理论支持。传统的引文分析在不断地寻求合理的引用行为的理论方法体系，以求能够合理地描述该行为并对其进行合理的量化和测度。现在的链接分析也面临着同样的问题。全面地分析链接行为和链接结构是做好链接分析的基础，同时也是网络信息计量学重点研究的问题。

参考文献

[1]Liwen Vaughan, Debora Shaw, Bibliographic and Web Citations: What Is the Difference? [J], Journal of the American Society for Information Science and Technology, 2003, 54(14): 1313 - 1322.

[2] 赵蓉英，段宇锋，邱均平．网络信息计量学研究 (I)——网络连接研究的现状及趋势 [J]．情报学报，2005, 24 (2).

[3]李江．链接分析工具研究[D]．武汉大学，2007.

[4] Vreeland Robert C. Law libraries in hyperspace: A citation analysis of world wide web sites[J]. Law Library Journal, 2000 (1).

[5] Ingwersen Peter. The calculation of web impact factor[J], Journal of Documentation, 1998(2).

[6] Kim, H. J. Motivations for hyperlinking in scholarly electronic articles: A qualitative study[J] . J. Am. Soc. Inf. Sci., 2000, 1(10): 887 - 899.

[7] Chu H. Reasons for sitation (hyperlinking): what do they imply for Webometric research? [C]. Paper presented at the International Conference on Scientometrics and Informetrics, 9th. 25 - 29 August 2003, Beijing.

[8] Wilkinson D., Harries G., Thelwall M. Price E. Causes of academic Web site interlinking: Evidence for the Web as a *novel source of information on informal scholarly communication* [J]. Journal of Information Science, 2003, 29(1): 59 - 66.

[9] Smith A. G. Web links as analogues of citations[J]. Information Research, 2004, 9(4).

[10] Thelwall M. Evidence for the Existence of Geographic Trends in University Website Interlinking[J] . Journal of Documentation, 2002, 58(5): 563 - 574.

[11]Scharnhorst, A., Thelwall, M.. Citation and hyperlink networks[J]. Current Science, 2005, 89(9): 1518 - 1523.

[12] 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.

[13]李江．链接分析工具研究[D]．武汉大学，2007.

[14] Alastair G Smith. Classifying links for substantive Web impact factors[EB/OL]. [2007 - 12 - 17]. http: // www. vuw. ac. nz/staff/alastair_smith/publns/issi_2003_classn. ppt.

[15] 李剑，金蓓弘．Web 链接结构信息研究综述 [J]．计算机科学，2003, 30(4).

[16] 俞培果，邱均平．Web 页面链接动机分析及链接侧度研究[J]．情报科学，2003, 3(21).

[17] 刘雁书，方平．利用链接关系评价网络信息的可行性研究[J]．情报学报，2002, 21(4).

[18]袁毅．链接分析用于学术网站评价存在的问题及解决办法[J]．情报学报，2005, 5(24).

[19]张海涛，刘甲学，宋川．超文本系统信息结构组成元素——链的分析[J]．情报科学，2002, 20(4).

[20]邱均平，黄晓斌．WWW 网页的链接分析及其意义[J]．中国图书馆学报，2002(6).

[21]张海涛，刘甲学，宋川．超文本系统信息结构组成元素——链的分析[J]．情报科学，2002, 20(4).

[22] 王丽伟．基于链接的网络计量指标与科学评价 [D]．吉林大学，2006.