基于竞赛图的大学网站链接分析研究*

杨波，王雪
（南京农业大学信息科学技术学院，南京 210095）

摘要：用于机构评价的链接分析是网络计量学研究的热点问题，而评价方法的科学性是保证评价结果可信的重要前提。针对当前WIF研究中的不足，本文提出了一种使用竞赛图进行机构评价的新方法，并通过对中国大陆综合实力排名前100位的大学网站的实证研究，验证了该方法的有效性和可靠性。最后，分析了可能对评价结果产生影响的各种因素。

关键词：链接分析；机构评价；竞赛图
中图法分类号：G256
DOI：10.3772/j.issn.1673-2286.2016.1.007

1 引言

从搜索引擎诞生之日起，链接分析就作为一种主要的Web超文本信息分析技术被广泛采用。如同文献计量学中的引文分析一样[1]，无论是PageRank还是Hits，都是以发现高质量的信息资源为目标的。随着超文本信息的急速增长，以及开放获取、机构仓储等网络出版形式的流行，对网络资源的评价和获取成为计量学研究的热点问题。美国学者Larson早在1996年就对此进行了开拓性的研究[2]。从网络计量学作为一个重要研究分支的提出到现在，链接分析始终是一种最主要的定量分析方法。和搜索引擎不同的是，网络计量学的链接分析更多倾向于资源评价，类似于引文分析的机构和期刊影响力评价。就机构评价而言，建立在科学的链接分析之上的网络计量学评价具有很多先天优势，它代表的是最广泛的意见，因此更注重综合的影响力评价。

网络计量学最初的研究主要集中在利用网络影响因子进行期刊评价[3]。借鉴于传统的期刊影响因子的概念，网络影响因子（Web Impact Factor）是利用期刊网站被其它网站的引用次数和网站的总网页数量之间的比率来衡量一个期刊在网络上的影响力，从而达到期刊评价的目的。这种方法用来评价期刊有一定的缺陷，它的评价功能过多依赖于信息技术的普及程度，因此在稳定性上受到学科因素的影响太大。比如，对工程领域的期刊评价存在一定的显著性[4]，而对农业科学的期刊评价效果却不是非常理想[5]。而就大学评价而言，这种影响却要小得多，甚至可以忽略，因为几乎所有的大学都有足够的技术力量来建立和维护自己的网站。本质上，WIF体现了大学网站在WWW上的流行度（Popularity），也体现了大学在网络空间的受欢迎程度。WIF的评价功能建立在各个大学在网络空间中受欢迎的程度和大学的整体实力有显著关系的假设上。

由于现有的衡量大学网站影响力的评价方法在结果的有效性和方法的完善性等方面存在一定的不足，无论是在软件资源、硬件资源，还是在人力资源和生源上，各个大学之间都存在直接或者间接的竞争关系。大学只有在努力提高自己的科研能力和管理水平的前提下，才能获得良好的声誉，从而在上述的资源竞争中获得优势。竞赛图中的弧类似于高校之间的竞争关系，大学作为图中的顶点，“击败”的大学越多，越体现出自身的综合实力和在同行中的认可度。正是在这一理论假设之下，本研究拟采用竞赛图的原理来探索利用网站之间的网络链接关系进行大学评价的新方法。本文拟引入竞赛图的原理，以大陆综合实力排名前100位的大

* 本研究得到国家社会科学基金“基于社区发现的学术Web主题显著度研究”（编号：13CTQ031）和教育部人文社科基金“基于社区发现的Web主题图构建技术研究”（编号：11YJC70030）资助。

学为样本，在对大学的链接网络进行转换的基础上，形成新的评价方法。这种方法充分考虑了样本之间的相互关联关系，可以以比较小的代价达到相对理想的评价效果。

2 相关研究

大量利用网络上的链接关系进行机构评价（大学或院系）的研究结果显示，网络链接数或者WIF和机构的研究力量或综合实力之间是有显著的相关关系的。早在1998年，丹麦学者Ingwersen就对机构网站的WIF测度进行了尝试性的研究[3]，虽然结果不是很理想，但为以后类似的研究提供了新的研究思路。Mike Thelwall利用搜索引擎分别统计了指向英国各个大学的四种不同的链接来源（.edu、.ac.uk、.uk和WWW），结果表明这些大学的网站的WIF和大学的研究实力是显著相关的[6]。对美国大学的心理学、化学和历史三个学科院系网站链接分析的研究结果发现，心理学和化学学科的院系网站的入链数和研究实力之间存在显著的相关关系，而从事历史学研究的院系由于比较少使用到Web，没有发现明显的相关规律[7]。Mike Thelwall在研究中国大陆和台湾两个地区的大学入链数和研究实力之间的关系时发现，台湾地区的大学研究实力排名和入链数有着比较明显的相关关系[8]。邱均平等人对入链数和大陆大学综合实力之间的关系进行了开拓性的研究，他们也发现网站的入链数和大学的综合实力和研究实力之间有显著的相关关系[9]。

总结以上研究，目前WIF的测度方式主要有以下两种：

（1）被链接次数（inlink）和总网页数之比。这种方式采用了期刊WIF的计算原理，也是用于大学评价的WIF计算的最初公式。这种被广泛采用的计算公式忽略了Web信息资源一个很重要的特点，那就是易于创建。从技术上讲，一个大学网站的总网页数量可以在很短时间内急剧增长，而被链接数的增长却比较困难（不考虑人为故意）。这样就导致以这个公式计算的WIF的评价结果有失公允。

（2）直接采用被链接次数（inlink）。由于以上存在的问题，研究者们开始逐渐探索利用被链接次数直接作为测度指标的方法。从多次对不同样本的实验结果来看，效果要明显优于传统的WIF测度方式[4,10]。

除了上述方法外，也有研究将大学或者学院的在编教学人员数量等作为考虑因素参与到WIF的测度中[7,11]。严格来说，直接采用入链数来衡量大学网站的影响力的方法不属于WIF的经典定义范畴。上面的两种途径在不同程度上都存在缺陷，第一种受到总网页数的影响比较大，而第二种方式忽略了被评价样本之间的关系，没有考虑到样本之间自评的功能。

3 研究方法

链接分析结论的可靠性主要由三个方面来保证：（1）样本选择的合理性；（2）数据采集方法的可靠性；（3）数据分析方法的科学性和多样性。前两个方面是基础，主要解决链接分析研究需要的基础性的数据保障问题。而在噪音信息影响非常严重的情况下，如何在合理的样本范围内，采用科学的方法获取可靠的链接数据，也是需要研究的问题。研究方法部分将对样本选取的依据、数据采集方法和数据分析方法作进一步说明。

3.1 样本选取

本研究选取了文献[12]中排名前100位的大学作为初始样本集合，由于域名问题和网站脚本解析等技术原因，部分样本被剔除，最后确定的样本容量是93。在这93个样本网站中，对由于学校合并等因素产生的多域名现象的大学网站，采用了合并计算的方法，以保证数据的完整性。

3.2 数据采集

现有国内外以计量分析为目的的链接分析研究主要采用的数据获取方式有三种：（1）使用大型商业搜索引擎，如AltaVista、Google等；（2）第三方商业网络爬行软件与自主开发相结合的方式，如Blinker[13]、Offline Explorer+webStat[14]；（3）自主开发链接抓取工具，如CheckWeb[15]、Thelwall[16]等开发的网络爬虫。

由于以上数据采集方式在有效性、数据采集的合理性以及采集策略的灵活性方面都存在一些问题，本研究将采用自行开发的链接分析专用数据采集系统——LinkDiscoverer[10]。可以根据各个网站的具体情况为LinkDiscoverer指定采集策略，包括

设置链接深度、过滤规则、超时控制、DNS缓存等，LinkDiscoverer根据采集策略并行采集样本网站的链接数据。在采集过程中，采集系统会对链接进行在线分类，并做初步统计。

3.3 基于竞赛图的数据分析

由数据采集工具返回的结果形成的初步分析数据是一个链接网络，这个网络是一个有向带权非简单图。竞赛图的基本输入要求是该链接网络必须是一个有向二值简单图，因此需要对初始的链接网络进行去弧操作和二值转换。本文所采用的方法是在竞赛图中的“击败”思想原则下，利用样本之间的相互链接数作为判断依据，对链接网络进行去弧操作。具体做法是，如果样本A链接样本B的次数多于样本B链接样本A的次数，则去掉A到B的弧（B击败A），反之去掉B到A的弧；如果遇到两者相互链接数相同的情况，则比较两个样本在链接网络中的度，如果A的度数大于B，则去掉B到A的弧，反之去掉A到B的弧；如果两个样本的相互链接数和度均相等，则判断两个样本各自被其他所有样本链接的总频次，去掉总频次小的一方到另外一方的弧。在去弧的同时，将剩余弧的权重设置为1。

图论中，完全图的每条边确定一个方向后所得的有向图称为竞赛图。在比赛中，竞赛图作为一种解决循环赛的名次排列问题的模型，以选手为顶点，若选手u胜了选手v，则由选手u向选手v连一条弧，最后形成一个有i个顶点的竞赛图$\vec{K}_i$。图1中表示的是有4个顶点的竞赛图的一种表示，这样对选手进行排名的问题就可以通过竞争图中每个顶点的出度大小的排序来解决[17]。

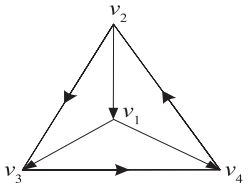

图1 竞赛图图例

图1可以用一个邻接矩阵$A=(a_{ij})_{v\times v}$来表示，如果存在从顶点到i的有向边，那么$a_{ij}=1$，否则为0。邻接矩阵A中第i行的和等于该竞争图中第i个顶点的出度，即第i个选手的得分，则该竞赛图的一阶得分向量可以表示为：$S^{(1)}=Ae$（e为v维单位向量）。如果选手的一阶得分相等，则需要计算二阶得分：$S^{(2)}=A^2\cdot e$。矩阵A^2第i行元素之和表示被选手v_i间接（两步）打败的选手数量之和。如果二阶得分不能确定排名，需要计算各顶点三阶得分。计算之前需要将A^2中所有非零元素改写成1，得到矩阵$\tilde{A}^2$，第i行第j列元素表示是否存在从v_i到v_j长度为2的有向路径。三阶得分的计算公式如下：$S^{(3)}=\tilde{A}^2\cdot A\cdot e$，依此类推，四阶得分$S^{(4)}=\tilde{A}^3\cdot A\cdot e$。直到计算出$v$阶得分向量时仍有未确定名次部分，可作为并列排名[17]。

4 实验

本研究利用了2008年5月20日至6月15日之间对93个样本网站采集的历史数据，对异常爆发链接过滤后，共获得链接7,209,197条，在此基础上形成了原始的链接矩阵。该矩阵经过去孤和二值转换后，输入竞赛图接口，得出最后的N阶得分（本研究在计算第三阶得分后得到不重复排名）。下面将对分批计算出的N阶得分与大学的综合实力排名之间的关系作详细说明。

4.1 N阶得分与大学排名的相关度分析

表1中列出的是本研究所选取的93所大学的综合实力排名和由竞赛图计算的N阶得分获得的排名之间的Spearman相关系数（在0.01的显著性水平下）。根据竞赛图的原理，一阶得分排名表示选手直接打败别的选手的数量，二阶得分排名表示两步打败的选手的数量，依次类推。随着N的不断增加，在竞争图中需要考虑的选手越多，针对每个选手的排名数据也越发散。由相关系数可以看出，综合实力排名和各阶得分的相关度逐渐降低，并且一阶得分和二阶得分的相关度明显要高于三阶得分。相关度的降低说明数据越发散（N越大），直接利用该得分数据进行大学排名的可靠性越低。

表1 N阶得分与大学排名相关系数

	一阶得分	二阶得分	三阶得分
综合实力排名	0.598	0.563	0.343

4.2 竞赛图综合得分与大学排名的相关度分析

虽然一阶得分和大学综合排名的相关度最高，但一阶得分的排名中有相当一部分样本的得分相同。为了

进一步确定得分相同的样本的排序，需要对这部分样本的二阶得分进行比较，如果二阶段得分相同，需要参考三阶得分。依次类推直到计算出v（v为样本空间）阶得分后，如果仍然有未确定排名，再确定为并列排名。本研究中，计算出三阶得分后，所有样本的排名被唯一确定。在综合各阶得分后，形成竞赛图的综合得分，此排名和大学综合实力排名的相关系数为0.597，如表2所示。

表2 综合得分（Tscore）与大学排名（URank）相关系数

Spearman's rho	URank	Correlation Coefficient	1.000	.597(**)
		Sig. (2-tailed)	.	.000
		N	93	93
	Tscore	Correlation Coefficient	.597(**)	1.000
		Sig. (2-tailed)	.000	.
		N	93	93

** Correlation is significant at the 0.01 level (2-tailed).

虽然综合得分排名与大学综合实力排名的相关系数和一阶得分排名相当，但一方面综合得分排名考虑了多阶得分，从理论上更加科学合理。另一方面，一阶得分排名中大量的并列排名的问题得到了克服，使得排名结果更加符合实用情况。

5 结论

从文献[9]中对入链数和WIF分别和大学综合实力的相关度来看（在某些情况下去除了合并后的某些大学和一些数据异常的大学），平均相关系数略低于本文。文献[6]对英国大学的WIFs和大学排名之间的相关系数大部分都在0.4以下。由于上述研究在样本、数据来源或者参照系上和本研究不同，相关度的高低不能完全代表方法的优劣，但至少在一定程度上说明本研究所采用的数据采集方法、链接权重计算方法和排名算法是合理有效的。运用竞赛图的原理，从大学网站之间的链接关系中挖掘代表大学综合实力的规律的方法是可行、有效的。

虽然本研究通过实证研究验证了竞赛图在大学网站链接分析方面的适用性，但多方面因素可能会对最后的评价结果产生一定的影响：

（1）排名参照系。虽然参照系的客观性对利用网络链接进行探索性机构评价研究很关键，但所有的大学排名都存在一定的争议性，所以本文以此作为对竞争图适用性评估的客观性还需要进一步验证。

（2）数据采集技术。大学合并产生的多域名问题是数据采集中比较棘手的问题，虽然从技术上问题能得到一部分解决，但彻底解决难度很大。此外，链接解析技术和链接质量评估等问题也会对结果产生很大影响。

（3）链接图转换算法。本研究需要将原始有向带权非简单图转换成一个二值简单图，是否存在比二值转换更为有效的算法还有待研究。

参考文献

[1] Garfield E. Citation Indexes for Retrieval and Research Evaluation [EB/OL]. [2009-10-07]. http://www.garfield.library.upenn.edu/papers/ciretreseval-capri.html.

[2] Larson R R. Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace [C]// Proceedings of the 59th annual meeting of the ASIS. NJ: Information Today. Medford: ASIS, 1996: 71-78.

[3] Ingwersen P. The Calculation of Web Impact Factors [J]. Journal of documentation, 1998, 54(2): 236-243.

[4] An L, Qiu J. Research on the relationships between Chinese journal impact factors and external web link counts and web impact factors [J]. Journal of Academic Librarianship, 2004, 30(3): 199-204.

[5] 屈卫群,杨波,阎素兰.农业期刊的期刊影响因子和网络影响因子比较研究[J].中国科技期刊研究,2005,16(5):658-661.

[6] Thelwall M. A comparison of sources of links for academic Web Impact Factor calcinations [J]. Journal of Documentation, 2002, 58: 66-78.

[7] Rong T, Thelwall M. U.S. academic departmental Web-site interlinking in the United States Disciplinary differences [J]. Library and Information Science Research. 2003, 25(4): 437-458.

[8] Thelwall M, Tang R. Disciplinary and linguistic considerations for academic Web linking: an exploratory hyperlink mediated study with Mainland China and Taiwan [J]. Scientometrics, 2003, 58(1): 153-179.

[9] Qiu J, Chen J, Wang Z. An analysis of backlink counts and Web Impact Factors for Chinese university websites [J]. Scientometrics, 2004, 60(3): 463-473.

[10] Yang B, Qin J. Data Collection System for Link Analysis [C]// Digital Information Management, 2008. ICDIM 2008. Third International Conference on IEEE, 2008: 247-252.

[11] Smith A G, Thelwall M. Web Impact Factors for Australasian Universities [J]. Scientometrics, 2002, 54(3): 363-380.

[12] 武书连,吕嘉,郭石林.2008中国大学评价[J].科学学与科学技术管理,2008,29(1):42-51.

[13] Ortega J L, Aguillo I F, Cothey V, et al. Maps of the academic web in the European Higher Education Area - an exploration of visual web indicators[J]. Scientometrics, 2008, 74(2): 295-308.

[14] 段宇峰.网络链接分析与网站评价研究[M].北京:北京图书馆出版社,2005:129-140.

[15] Magnusson C. CheckWeb [EB/OL]. [2004-09-20]. http://www.algonet.se/~hubbabub/how-to/checkweben.htm.

[16] Thelwall M. A Web Crawler Design for Data Mining [J]. Journal of Information Science, 2001, 27(5): 319-325.

[17] 高随祥.图论与网络流理论[M].北京:高等教育出版社,2009:276-277.

作者简介

杨波，男，1981年生，博士，南京农业大学信息科学技术学院副教授，研究方向：信息计量学，E-mail：boyang@njau.edu.cn。
王雪，女，1990年生，硕士，南京农业大学信息科学技术学院研究生，研究方向：信息计量学。

Link Analysis on University Websites Based on Tournament Algorithm

YANG Bo, WANG Xue
(College of Information Science & Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: Link analysis for institution evaluation is a hot topic in webometric study and scientific evaluating methods, and algorithms are important premises of credible results. A new method for institution evaluation applying tournament is proposed in this study to solve the problem in previous studies on WIF. The validity and reliability of the method are verified in the experiment on a sample composed of the websites of Top 100 University in Mainland China. Finally, several aspects concerning to the performance of the method are analyzed.

Keywords: Link Analysis; Institution Evaluation; Tournament Algorithm

（收稿日期：2015-12-29）

书讯

《科技报告体系构建研究》

为推进我国科技报告制度建设，强化科技报告资源共享服务，贺德方研究员率领中国科学技术信息研究所科技报告研究团队，进行了国家社会科学基金重点项目“中国科技报告资源体系构建”（11ATQ006）研究，并对20多年来中国科学技术信息研究所相关研究和实践进行了归纳、凝练、整理和补充，最终形成了《科技报告体系构建研究》一书。

本书作为国家社会科学基金重点项目的主要研究成果，总结了科技报告产生发展的管理历程、凝练了科技报告制度的建设路径、制订了科技报告资源的整合方案，提出了科技报告体系的构建模式，归纳了科技报告实践的操作过程。本书对各级科技计划管理人员强化科技计划项目过程管理具有借鉴作用，对科研人员撰写高质量科技报告具有指导作用，对各类科研机构做好科技报告呈交、推进科技项目的规范管理和机构知识库建设具有参考价值，对图书信息机构做好科技报告深层次加工和收藏利用具有引导作用，也可供高校信息管理、科技政策与管理等专业研究生学习参考。

《科技报告体系构建研究》于2014年12月由科学技术文献出版社出版，定价78.00元。