

三种网络分析方法对比实证研究

岳增慧^{1,2} 方 曙¹

¹中国科学院国家科学图书馆成都分馆 成都 610041 ²中国科学院大学 北京 100049

〔摘要〕阐述共链、URL 共引以及网络标题共现这三种网络分析方法的概 念、特性、优势与局限性、数据来源与收集方法,并以全球 30 所高校及其网站为例,比较三者的相关性 与有效性。结果显示,URL 共引以及网络标题共现分析法是补充和替代共链分析 法的有效网络信息计量方法。

〔关键词〕共链 URL 共引 网络标题共现 比较 实证

〔分类号〕G250.7

Comparative Empirical Study of Three Network Analysis Methods

Yue Zenghui^{1,2} Fang Shu¹

¹Chengdu Library of Chinese Academy of Sciences, Chengdu 610041

²University of the Chinese Academy of Science, Beijing 100049

〔Abstract〕The paper expounds the concepts, characteristics, advantages and limitations, data sources, methods of collecting data of the analytical methods of co-link, co-URL citation and web co-title mention. Then it takes 30 selected university libraries and their websites as an example, and discusses the correlation and effectiveness of the three methods. Results from the case study suggest that the analytical methods of co-URL citation and web co-title mention can be appropriate replacements for co-link analysis.

〔Keywords〕co-link co-URL citation Web co-title mention comparison empirical study

1 引言

自 1997 年 Almind 和 Ingwersen 提出“网络计量学 (webometrics)”一词以来^[1],该领域的研究已取得显著进展。目前,最普遍的研究对象是网络链接,其主要研究方法之一——链接分析,已经由理论逐步发展成为了一项科学计量学乃至社会科学的实用工具^[2]。通过入链分析、互链分析及共链分析,可以获得大量的知识信息。链接分析主要利用网络搜索引擎获取原始数据(如 Alta Vista、All the Web、Google、Yahoo! 等),但近年来,执行数据收集的商业搜索引擎的可用性成为了链接分析研究的限制性因素^[3]。目前,Yahoo! 成为了唯一可以满足该需求的搜索引擎。但是,由于其被归并入微软的 Bing(由于过度的使用而撤消了其绝大部分的链接检索功能),Yahoo! 于 2012 年也将终止该项服务^[4]。这对该领域的发展来说是一个极大的威胁。鉴于此,我们需要挖掘其他类型的网络资源,替代和补充传统的超链接元素,扩展数据源并开发新的方法,以防止该领域研究的停滞,促进其发展。

网络共链分析的思想来源于文献共引分析。一般认为存在共链关系的网站,在主题上或多或少都具有一定的相似性或相关性,利用网站共被链入的次数(共链频次),可以测度网站内容的相关度或相似度。共链频次越高,相关度越大^[5]。基于此,衍生出了两个新的网络计量学概念——URL 共引以及网络标题共现。二者反映了网站或机构之间的间接联系,可以作为超链接的替代或补充,获取隐性网络信息资源和反映知识结构。

早在 1998 年,B. Cronin 等人便提出了 Web Mentions 的概念^[6]。随后,其便被广泛地应用于期刊文献的分析研究中。2010 年,L. Vaughan 等利用网络标题共现法测度了电信业两大阵营 WiMax 和 Long Term Evolution 中公司的关联性,并将其与共链分析结果进行了对比,证实了网络标题共现法的应用价值^[3]。URL citations 的概念提出较晚,K. Kousha 等人主要将其用于期刊文献集的评价^[7-8]。2006 年,D. Stuart 等利用 URL 引用数研究了英国西米德兰兹郡汽车产业不同机构间的三螺旋关系^[9]。2011 年,M. Thelwall 等比较了为学术机构收集网络引用数据的三种方法^[10]。2012

收稿日期:2012-05-29

修回日期:2012-08-13

本文起止页码:58-61,102

本文责任编辑:高丹

年, M. Thelwall 等又利用 URL 引用数以及机构标题涉及数构建并分析了链接与共链网络图^[2]。我国学者对该主题的研究相对较少, 较为典型的是张洋等人利用网络共链和共词分析法对国际银行业竞争态势进行研究, 并提出二者属互补关系, 应结合使用^[11]。

本文阐述了共链、URL 共引以及网络标题共现这三种网络分析法的概念、特性、优势与局限性、数据来源与收集方法, 并以全球 30 所高校及其网站为例, 比较了三者的相关性与其有效性。

2 三种方法的概念与特征

2.1 概念

共链包括共入链与共出链。所谓共入链, 就是指两个网站同时被其他网站链接; 所谓共出链, 是指两个网站共同链接向其他网站, 亦称耦合。一般情况下所讨论的共链指的是共入链。网络共链分析提供的是分析对象之间由链接关系形成的相互关系的亲疏程度, 通过一组网站之间的共链关系可以形成共链网络, 网络节点之间的远近可以反映它们的亲疏关系, 进而揭示网站间的关系模式, 并利用聚类分析、多维尺度分析等统计分析方法, 将这种群落关系表示出来, 以实现网络结构可视化及网络群体间相互关系的挖掘^[12]。

传统定义中的“链接”, 一般指的是“超链接”, 但是也可以使用 URL 引用“链接”或者网络标题“链接”来代替。由此衍生出 URL 共引以及网络标题共现分析的概念。

所谓 URL 共引, 是指两个网站的 URL 同时被其他网站引用/包含。所谓网络标题共现, 是指两个网站/机构的标题名称同时出现在其他网站的页面中。与共链分析法相对应, URL 共引分析以及网络标题共现分析法分别是指利用 URL 共引关系以及标题共现关系对分析对象进行相似性测度的方法, 进而揭示网站/机构之间的关系模式, 实现隐性知识信息的挖掘与获取。

2.2 特征

根据共链、URL 共引以及网络标题共现的定义, 三者主要具有以下特征, 如表 1 所示:

表 1 共链、URL 共引与网络标题共现的特征

类型	第三方网页中的元素	指代的实体	是否具有导航性
共链	超链接	网站	是
URL 共引	URL(不管是否具有超链接)	网站	是
网络标题共现	文本(网站机构名称)	机构	否

网站 A 与 B 存在共链是指第三方网站 C 的页面

中同时含有指向网站 A 与 B 的链接; 而网站 A 与 B 存在 URL 共引是指第三方网站 C 的页面中同时含有网站 A 与 B 的 URL 或域名, 但不需要是指向网站 A 与 B 的链接; 网站/机构 A 与 B 存在网络标题共现是指第三方网站 C 的页面中同时含有网站/机构 A 与 B 的文本名称。共链与 URL 共引在结构上是相同的, 二者都嵌入网页中并且都指向其他的网页; 网络标题共现相对普通, 可以是一个网络机构或非网络机构的名称。

网络标题共现与 URL 共引以及共链的一个重要的区别在于, 前者暗指机构, 而后者指代的实体为机构的网站。对于网络机构(如 sina)来说, 它们的区别可能较小, 但是对于非网络机构(如大学)来说, 这种区别是显著的。

另外, 共链与 URL 共引都具有较为明确的网站导航性; 然而, 网络标题共现并没有明显的导航特性。因此, 它们的创建动机并不是等价的, 在不同的环境下, 存在差异性^[2]。

2.3 优势与局限性

三者的优势与局限性如表 2 所示:

表 2 共链、URL 共引与网络标题共现分析的优势与局限性

类型	是否受搜索引擎可用性限制	是否具有语言敏感性	是否具有查询多义性	是否受搜索引擎自身因素影响
共链	是	否	否	是
URL 共引	否	否	否	是
网络标题共现	否	是	是	是

共链分析会受到搜索引擎链接检索功能可用性的限制, 因此, 只能在有限的搜索引擎中使用。然而, URL 共引以及网络标题共现分析不会受到此类限制, 可以作为传统超链接的替代和网络资源的补充。

网络标题共现分析具有明显的语言敏感性, 标题所使用语言的不同(如“Tsinghua University”与“清华大学”)对于查询结果会产生很大的影响; 相反, 由于链接或 URL 使用全球统一的 html 语言进行编码, 因此, 共链分析与 URL 共引分析并没有语言敏感性^[3]。

除了语言敏感性外, 网络标题共现还存在一个问题, 即“一义多词”以及“一词多义”现象。所谓“一义多词”是指同一个实体具有几个不同的标题名称, 如果仅使用其中一种或几种标题名称进行查询的话, 会造成查询结果的遗漏。该问题可以通过“多重检索”来解决, 即对每一个标题名称单独进行检索, 然后将所有结果进行去重、合并, 得到最终结果。所谓“一词多义”, 一般出现在使用首字母缩写代表标题进行共现查询的时候, 是指同一个标题名称代表几个不同的含义,

这会造造成查询结果的虚假扩大(掺入非准确结果的噪声)。该问题可以通过在原标题中加入附加文本信息(限定信息) 来解决, 以保证所查询的是正确的实体。共链以及 URL 共引元素具有明显的指向性和唯一性, 因此, 并不存在该限制。

但是, 三者都依靠搜索引擎进行原始数据的获取, 因此都会受到搜索引擎覆盖率、检索变化以及检索异常^[9]等搜索引擎自身因素的影响。

3 数据来源与收集方法

3.1 研究对象选择

本文以高校及其网站为研究对象, 以先进性、地域代表性作为样本的选取原则。根据《美国新闻及世界报道》(U. S. News & World Report, 简称 U. S. News) 公布的 2011 年世界大学排行榜^[13], 从排名在前 60 位的高校中选取了分布在全球不同国家/地区的 30 所高校及其网站作为研究对象。

3.2 数据收集方法

目前进行三者数据查询所使用的搜索引擎一般为 Yahoo! 与 Bing^[2-3, 10, 14]。由于 Yahoo! 是目前唯一可以执行共链检索的搜索引擎, 因此, 本文选其作为数据收集的工具, 所用检索式如表 3 所示(以清华大学和哈佛大学网站为例) :

表 3 检索式

检索指标	检索式
共链	(http://www.tsinghua.edu.cn -site:tsinghua.edu.cn) AND (http://www.harvard.edu -site:harvard.edu)
URL 共引	"www.tsinghua.edu.cn" "www.harvard.edu" -site:tsinghua.edu.cn -site:harvard.edu
网络标题共现	"Tsinghua University" "Harvard University" -site:tsinghua.edu.cn -site:harvard.edu

注: 该表中网络标题共现检索文本中以英语为例。

考虑到搜索引擎的数据更新, 为了保持数据的一致性, 笔者在 2012 年 4 月 19-22 日集中采集了数据。

4 实证分析

4.1 相关性

为了验证共链网络、URL 共引网络以及标题共现网络之间是否具有相关关系, 笔者利用 Ucinet 软件对上述矩阵两两进行二次指派程序(Quadratic Assignment Procedure, QAP) 分析。QAP 分析可以用来研究两种关系矩阵之间是否相关^[15], 所得结果如表 4、表 5

所示:

表 4 共链、URL 共引与网络标题共现矩阵的 QAP 分析结果(QAP Correlations)

类型	共链	URL 共引	网络标题共现
共链	1.000	0.249	0.477
URL 共引	0.249	1.000	0.125
网络标题共现	0.477	0.125	1.000

表 5 共链、URL 共引与网络标题共现矩阵的 QAP 分析结果(QAP P-Values)

类型	共链	URL 共引	网络标题共现
共链	0.000	0.005	0.000
URL 共引	0.005	0.000	0.076
网络标题共现	0.000	0.076	0.000

其中, QAP Correlations 是指网络之间实际的, 即观察到的相关性系数^[16]; QAP P-Values 是指显著性水平, 当 $P < 0.05$ 时, 说明两个网络存在某种必然的联系^[17]。

由表 4、表 5 的计算结果可知: ①共链网络与 URL 共引网络的相关系数为 0.249, 显著性水平为 $0.005 < 0.05$, 说明二者在统计学意义上存在一定的相关性, 且相关程度为 0.249; ②共链网络与标题共现网络的相关系数为 0.477, 显著性水平为 $0.000 < 0.05$, 说明二者在统计学意义上具有明显的相关性, 且相关程度为 0.477; ③URL 共引网络与标题共现网络的相关系数为 0.125, 显著性水平为 $0.076 > 0.05$, 说明二者在统计学意义上不存在明显的相关性, 且相关程度仅为 0.125。造成该现象的原因可能有两个: ①样本性质的限制。本例中选取的研究对象仅为高校及其网站, 其他类型的机构及其网站并未涉及。②样本规模的限制。本例中的样本量较小, 可能不能很好地反映整体网络的实际状况。③URL 共引与标题共现的本质机理并不相同, 决定了二者之间并不存在显著的相关关系。

4.2 有效性

评估某种分析方法有效性的途径主要有两种: 一种是将分析结果呈现给专家, 由专家直接判断其价值与效用, 该方法的弊端是容易掺杂入个人的主观意识; 另一种方法是将其结果与权威的外部测度源进行对比, 以确定其有效性大小。

《美国新闻及世界报道》(U. S. News) 公布的世界大学排行榜, 其排名融合了 academic peer review(学术评分)、employer review(雇主评分)、student to faculty(师生比例)、international faculty(国际学院评分)、international students(国际学生评分)^[13]等指标, 是当今世界权威的

大学排行榜,可作为其他测度方法有效性判断的一个重要的外部来源。点度中心度是指与某节点存在直接联系的其他节点的数目,是衡量节点重要性大小的一项指标;其值越大,说明该节点与网络中其他节点的联系越密切,在网络中的点度核心程度越高。为评估共链、URL 共引以及网络标题共现分析法的有效性,笔者将各高校及其网站的点度中心度与其在大学排行榜中的排名进行相关性分析,所得结果如表 6、表 7 所示:

表 6 30 所高校的大学排行榜排名与共链、URL 共引以及网络标题共现点度中心度排名

中文名称	原排名	新排名	共链点度中心度排名	URL 共引点度中心度排名	网络标题共现点度中心度排名
哈佛大学	1	1	11	5	1
剑桥大学	2	2	9	6	7
耶鲁大学	3	3	2	16	2
伦敦大学学院	4	4	4	3	12
牛津大学	6	5	12	7	8
芝加哥大学	7	6	10	8	9
杜克大学	14	7	3	9	5
康奈尔大学	15	8	8	13	4
斯坦福大学	16	9	1	20	3
澳大利亚国立大学	17	10	13	14	13
麦吉尔大学	18	11	7	2	10
苏黎世联邦理工学院	21	12	18	18	25
东京大学	22	13	14	15	14
香港大学	24	14	16	10	28
曼彻斯特大学	26	15	15	26	26
卡耐基梅隆大学	27	16	6	4	11
巴黎高等师范学校	28	17	27	29	30
多伦多大学	29	18	5	1	6
新加坡国立大学	30	19	23	17	15
香港科技大学	35	20	22	22	27
巴黎综合理工学院	36	21	30	30	16
悉尼大学	38	22	20	12	24
不列颠哥伦比亚大学	40	23	17	11	18
洛桑联邦高等理工学院	42	24	19	25	23
大阪大学	43	25	21	23	22
首尔国立大学	48	26	29	27	21
阿姆斯特丹大学	49	27	28	28	20
清华大学	50	28	25	19	17
哥本哈根大学	51	29	26	24	19
海德堡大学	57	30	24	21	29

注“新排名”是指按照 USNews 原排名的顺序依次赋予样本高校及其网站的排名次序,其取值区间为[1, 30]。

表 7 30 所高校新排名与共链、URL 共引以及网络标题共现点度中心度排名的相关性

类型	新排名	共链	URL 共引	网络标题共现
新排名	1.000	0.783	0.613	0.683
共链	0.783	1.000	0.749	0.744
URL 共引	0.613	0.749	1.000	0.597
网络标题共现	0.683	0.744	0.597	1.000

由表 7 可知,30 所高校及其网站的共链、URL 共引以及网络标题共现点度中心度的排名与 USNews 公布的世界大学排名次序呈现出明显的相关性,其相关

程度分别为 0.783、0.613 和 0.683。说明三者都是进行网络信息计量的有效方法,都能够有效地揭示和映射网络信息资源及其特征。

共链与 URL 共引以及网络标题共现点度中心度的排名相关系数分别为 0.749 和 0.744。说明 URL 共引以及网络标题共现法,作为新的网络计量学方法,在揭示网络信息特征等方面与共链分析法存在很强的相关性,呈现出相似的功能,因此,可以作为共链分析法的替代或补充,用于网络计量分析研究。

URL 共引与网络标题共现点度中心度的排名相关系数为 0.597,说明二者在揭示网络资源重要性方面具有一定的相关性,与相似度。

5 结 论

本文通过对共链、URL 共引以及网络标题共现分析法进行比较,得到了以下结论:

- 共链、URL 共引以及网络标题共现分析法在概念、特性、数据收集方法等方面既相互联系又存在着一定差异。
- 共链网络与标题共现网络具有明显的相关性;共链网络与 URL 共引网络存在一定的相关性;URL 共引网络与标题共现网络在统计意义上不存在明显的相关性。
- URL 共引以及网络标题共现分析法是补充和替代共链分析法的有效的网络信息计量方法;二者可以作为新的数据源,代替传统的超链接数据,进行网络计量研究。

以上关于共链、URL 共引、网络标题共现分析法的相关性及有效性的结论都是建立在 30 所高校及其网站基础上的,其可推广性还需要进一步的讨论。因此,下一步的研究需要注意以下几点:扩展研究领域(如商业机构或政府机构);扩大样本规模,避免样本统计随机性过大而给分析结果带来影响;进行定性分析,补充和完善定量分析结果。

参考文献:

[1] Almind T C, Ingwerson P. Informetric analysis on the World Wide Web: Methodological approaches to “webometrics” [J]. Journal of Documentation, 1997, 53(4): 404-426.

[2] Thelwall M, Sud P, Wilkinson D. Link and co-inlink network diagrams with URL citations or title mentions [J]. Journal of the American Society for Information Science and Technology, 2012, 63(4): 805-816.

(下转第 102 页)

- [4] Van Wijk R ,Jansen J P ,Lyles M A. Inter and intra - organizational knowledge transfer: A meta - analytic review and assessment of its antecedents and consequences [J]. Journal of Management Studies 2008 45 (4) : 830 - 857.
- [5] 姜磊 季民河. 城市化区域创新集群与空间知识溢出——基于空间计量经济学模型的实证[J]. 软科学 2011(12): 86 - 90.
- [6] 陈金丹 胡汉辉. 产业集群网络上的知识转移分析[J]. 科学与科学技术管理 2010 100 (2) : 100 - 104.
- [7] 赵顺龙 张艳君. 技术联盟中知识转移网络模型的构建与分析: 基于复杂网络理论 [J]. 科技进步与对策 2010 27(15) : 123 - 126.
- [8] 吴玉鸣. 官产学 R&D 合作、知识溢出与区域专利创新产出[J]. 科学学研究 2009 27(10) : 1486 - 1494.
- [9] 蔡伟毅, 张俊远. 全球化中的知识溢出与技术进步分析[J]. 江淮论坛, 2009(6) : 47 - 53.
- [10] 简兆权 郝雪云. 区域创新系统内部知识转移障碍及对策分析——以泛珠江三角洲区域为例 [J]. 科学学与科学技术管理 2010(5) : 166 - 170.
- [11] 赵大丽 高伟 李艳丽. 知识转移方式对区域创新能力的影响研究[J]. 科技进步与对策 2011 28(16) : 32 - 37.
- [12] Levin D Z , Cross R. The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer [J]. Management Science , 2004(1) : 1477 - 1490.
- [13] 陈武. 知识传播机理的物理学视角探讨: 从知识势差的角度来解释知识流动 [J]. 科技和产业 2011(1) : 110 - 113.
- [14] 王金营. 人力资本在技术创新技术扩散中的作用研究 [J]. 科技管理研究 , 2000(1) : 12 - 14.
- [15] 张海涛 吴艳玲. 基于知识势能的高校科研团队内部的知识转移机理 [J]. 图书情报工作 2010(20) : 110 - 114.
- [16] 张庆普, 李志超. 企业隐性知识流动与转化研究 [J]. 中国软科学 , 2003(1) : 88 - 92.
- [17] 左美云. 知识转移与企业信息化 [M]. 北京: 科学出版社, 2006.
- [18] 闫凡荀, 武忠. 知识转移的系统机理探析 [J]. 情报杂志 , 2008(8) : 134 - 136.
- [19] 周仕平, 郭东强. 中小型虚拟企业间知识转移影响因素分析 [J]. 湖北经济学报 2010 8 (4) : 99 - 104.
- [20] 李建军. 硅谷模式及其产学研创新体制 [D]. 北京: 中国人民大学 2000.
- [21] 王炳富, 刘芳. 创新集群内知识转移的模式与特点研究 [J]. 科技进步与对策 2010 27(16) : 135 - 138.
- [22] 李伟超. 图书馆联盟知识转移与共享文化建设 [J]. 情报理论与实践 2011 34(5) : 24 - 26.
- [23] 耿蕊. 中国动漫产业集群发展研究 [D]. 武汉: 武汉大学 2010.
- (作者简介) 杨刚, 男, 1968 年生, 教授, 博士, 发表论文 30 余篇; 王磊, 女, 1987 年生, 硕士研究生, 发表论文 2 篇; 宫丽莹, 女, 1989 年生, 硕士研究生, 发表论文 1 篇; 张佳硕, 女, 1988 年生, 硕士研究生, 发表论文 1 篇。

(上接第 61 页)

- [3] Vaughan L , You J. Word co-occurrences on webpages as a measure of the relatedness of organizations: A new webometrics concept [J]. Journal of Informetrics , 2010 , 4(4) : 483 - 491.
- [4] When will the change happen? How long will the transition take? [EB/OL]. [2012 - 05 - 19]. <http://help.yahoo.com/l/us/yahoo/search/alliance/alliance-2012.html>.
- [5] 王建芳 冷伏海. 共引分析理论与实践进展 [J]. 中国图书馆学报 2006 32(6) : 85 - 88.
- [6] Cronin B , Snyder H W , et al. Invoked on the web [J]. Journal of the American Society for Information Science , 1998 , 49(14) : 1319 - 1328.
- [7] Kousha K , Thelwall M. Motivations for URL citations to open access library and information science articles [J]. Scientometrics , 2006 , 68(3) : 501 - 517.
- [8] 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.
- [9] Stuart D , Thelwall M. Investigating triple helix relationships using URL citations: A case study of the UK West Midlands automobile industry [J]. Research Evaluation , 2006 , 15(2) : 97 - 106.
- [10] Thelwall M , Sud P. A comparison of methods for collecting Web citation data for academic organizations [J]. Journal of the American Society for Information Science and Technology , 2011 , 62 (8) : 1488 - 1497.
- [11] 张洋 余厚强. 基于网络共链分析和共词分析的国际银行业竞争态势研究 [J]. 情报杂志 2012 31(1) : 95 - 100.
- [12] 岳增慧 王娟 郑春厚. 基于共链分析的高校图书馆网络结构研究 [J]. 图书情报工作 2011 55(9) : 118 - 121.
- [13] 2011 年世界大学排名 [EB/OL]. [2012 - 05 - 19]. <http://www.university-list.net/rank1.htm>.
- [14] Thelwall M. A comparison of link and URL citation counting [J]. ASLIB Proceedings , 2011 , 63(4) : 419 - 425.
- [15] 苏娜 张志强 刘志辉. 基于链接分析的图林博客显著度分析 [J]. 情报资料工作 2010(1) : 98 - 102.
- [16] 刘军. 整体网分析讲义: UCINET 软件实用指南 [M]. 上海: 上海人民出版社 2009.
- [17] 李纲 王忠义. 人际情报网络与知识本体网络之间关系研究 [J]. 情报杂志 2011 30(6) : 36 - 39.
- (作者简介) 岳增慧, 女, 1985 年生, 博士研究生, 发表论文 4 篇。方曙, 男, 1957 年生, 研究员, 博士生导师, 发表论文 70 余篇。