

网络信息半衰期测度的探索性研究 *

段宇锋

摘 要 借鉴文献计量学“半衰期”的概念,提出测度网络信息效用衰减的两个指标 HL_{out} 和 HL_{in} ,并以中国和 OECD 确定的 24 个发达国家的网络信息为样本进行实证研究。研究数据表明, HL_{out} 的测定方法和结果更理想,该指标对网络信息的质量和效用评价具有一定的参考价值。图 1。表 2。参考文献 6。

关键词 网络信息 半衰期 网络链接

分类号 G354.2

ABSTRACT First, the author puts forward HL_{out} and HL_{in} to measure the decay of information utility by using the concept of “half-life” in bibliometrics. Then, an empirical study is made. Sampling the network information from websites of 24 developed countries and China, the data show that the result and the method of HL_{out} are more ideal, and they have reference values for evaluating the quality and utility of network information. 1 fig. 2 tabs. 6 refs.

KEY WORDS Network information. Half-life. Hyperlink.

CLASS NUMBER G354.2

信息的有效价值(或称为效用)随着时间的流逝而衰减是普遍存在的现象,文献计量学称之为“老化”,以半衰期、普赖斯指数、剩余有益性指标度量。其中,半衰期这一指标的影响最为广泛。在互联网领域,网络信息的效用衰减也是普遍存在的现象。虽然网络信息之间的链接关系与科学文献之间的引证关系存在诸多差异,但是,引文分析的方法和思路对网络计量研究仍具有借鉴价值^[1-2]。因此,本文以“半衰期”为切入点,对网络信息效用衰减的测度进行探索性研究。

1 测度指标

图 1 简要概括了网络信息依靠链接建立的联系。网页是网络信息存在的基本形式。以网页 S 为研究对象,它具有指向外部资源(网页集 1)的链接(出链,Outlinks);同时,它也被外部资

源(网页集 2)所链接(入链,Inlinks);此外,网页 S 还有指向其自身内部的链接(自链,Selflinks)。但是,Selflinks 大多属于结构性链接,不具备价值评价作用,因此在计量时应排除。

理论上,只要不是故意制造“垃圾链接”,无论出于任何目的,出链是对其指向资源价值的肯定,即使是批判价值。这是网络信息半衰期测度的第一个视角,我们将其标记为 HL_{out} ,其设计思路类似于文献共时半衰期。

换一个角度,只有有价值的信息才可能获得来自于其他信息的链接,因此,入链是信息价值和效用的体现。通常,经典信息的效用持续时间较一般信息长,入度(Indegree)也较大,WebQuery、PageRank 和 HITS 都基于这个假设。1997 年,Jeromy Carriere 和 Rick Kazman 开发出基于节点被链接数排序的 WebQuery 系统,使网络链接分析成为提高搜索引擎检索质量的重要方法^[3]。但是,这个系统的缺陷也显而易见。

* 本文为国家社科基金青年项目“基于网络信息计量的网络信息资源配置与调控机制研究”(05CTQ004)的阶段性研究成果。

譬如,为了提高某个节点的位置,可以创建大量的重复文档和链接指向该节点,网站之间的链接交流也会对其产生显著影响。围绕这一问题曾展开了许多研究,主要在于解决如何给不同的链接赋予不同的权重。其中比较重要的成果包括 Sergey Brin 和 Lawrence Page 开发的 PageRank 算法,以及 Kleinberg 开发的 HITS 算法,它们

分别被应用到 Google 和 IBM CLEVER 搜索引擎^[4-5]。既然入链体现了信息的效用,信息效用的衰减过程自然也就通过入链数随时间而变化的情况反映出来。这是我们测定网络信息效用的第二个视角,将依此设计的半衰期指标记为 H_{Lin} ,它类似于文献历时半衰期。

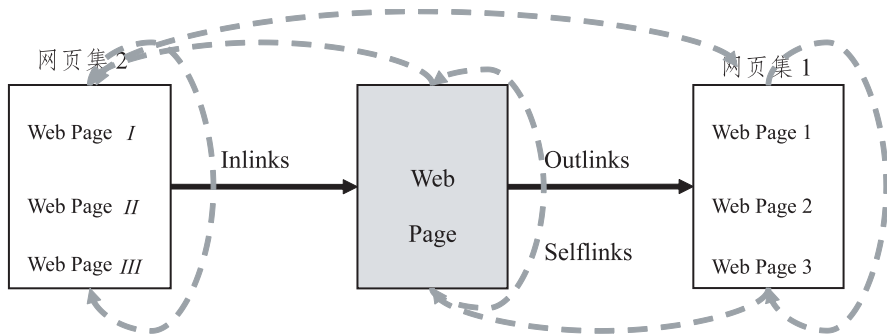


图 1 网络信息链接示意图

2 HL_{out} 的测定

2.1 研究方法及其结果

(1) 从 WebCPS(本研究开发的网络数据分析工具)的数据库中以随机方式抽取 10 万个属于中国 IPv4 地址段的 URL 及该文档的最后修改时间,分别记为 URL_{Sample} 和 T_{Sample} 。

(2) 访问 URL_{Sample} ,提取文档中的所有 Outlink 信息,分别记为 $URL_{Outlink1}$ 、 $URL_{Outlink1}$ …… $URL_{Outlinkn}$ (统称为 $URL_{Outlinks}$)。访问 $URL_{Outlinks}$,记录对应文档的最后修改时间 $T_{Outlink1}$ 、 $T_{Outlink2}$ …… $T_{Outlinkn}$ 。若 URL_{Sample} 或 $URL_{Outlinks}$ 无响应、 URL_{Sample} 代表的文档不含 Outlink 或 $URL_{Outlinks}$ 全都不属于我国 IPv4 地址空间,则直接访问下一个样本。若 $URL_{Outlinks}$ 中的某个不属于我国 IPv4 地址空间或者访问无响应,则处理 $URL_{Outlinks}$ 集合中的下一个。

(3) 计算并记录 $T_{Sample} - T_{Outlink1}$ 、 $T_{Sample} - T_{Outlink2}$ …… $T_{Sample} - T_{Outlinkn}$ 的值。

(4) 重复以上步骤,直至处理完所有样本。

(5) 排除小于 0 的值后排序,计算中值,该

值即为中国网络信息的 HL_{out} 。

重复上述过程,获取 OECD 确定的 24 个发达国家的 HL_{out} (见表 1)。

表 1 网络信息的 HL_{out}

国别	HL_{out} (天)	国别	HL_{out} (天)
中国	9.28	新西兰	212.59
瑞典	45.73	瑞士	218.10
芬兰	101.08	葡萄牙	219.55
英国	132.58	意大利	222.25
德国	140.99	西班牙	238.16
日本	141.10	韩国	239.78
美国	152.50	澳大利亚	280.08
法国	166.67	希腊	309.44
荷兰	172.53	挪威	315.73
冰岛	177.88	卢森堡	320.78
加拿大	182.33	丹麦	363.67
奥地利	196.48	比利时	530.61
爱尔兰	205.85	—	—

2.2 讨论

(1)本研究提取 Outlink 的标准是文档所含指向其所在主机以外的 URL。采用这一约束条件可以实现两个目的:一是排除无意义的结构性链接;二是排除因采用自顶向下的网站设计方法造成 $T_{Sample} - T_{Outlinks} < 0$ 的可能性。

(2) URL_{Sample} 和 $URL_{Outlinks}$ 指向的文档在发布后都可能不断地被修改。如果文档 $URL_{Outlinks}$ 的修改发生在文档 URL_{Sample} 之后,则文档 URL_{Sample} 中的 Outlink 实际上不再具有效用计量价值,所以,应排除 $T_{Sample} - T_{Outlinks} < 0$ 的数据。

(3) HL_{out} 的值越大,表明网络信息的效用衰减速度越慢,效用持续时间越长。中国网络信息的半衰期只有 9.28,发达国家的平均水平是 220.27,表明我国网络信息中效用历久不衰者比发达国家少。

3 HL_{in} 的测定

3.1 研究方法及结果

(1)从 WebCPS 的数据库中以随机方式抽取属于中国 IPv4 地址段的 URL 及该文档的最后修改时间。

(2)将该 URL 提交 Yahoo! Site Explorer,获得该 URL 的 Inlinks 信息,分别记为 $URL_{Inlink1}$ 、 $URL_{Inlink2} \cdots \cdots URL_{Inlinkn}$ (统称为 $URL_{Inlinks}$)。若该 URL 无对应的 Inlink 信息,则直接提交下一个样本 URL。

(3)访问 $URL_{Inlinks}$,相应文档的最后修改时间记为 $T_{Inlink1}$ 、 $T_{Inlink2} \cdots \cdots T_{Inlinkn}$ 。若 $URL_{Inlinks}$ 中的某个无响应,则处理 $URL_{Inlinks}$ 集合中的下一个。

(4)计算并记录 $T_{Inlink1} - T_{Sample}$ 、 $T_{Inlink2} - T_{Sample} \cdots \cdots T_{Inlinkn} - T_{Sample}$ 的值。

(5)重复以上步骤,直至处理完所有样本。

(6)取大于 0 的数据排序,计算中值,该值即为 HL_{in} 。

重复上述过程,获取 OECD 确定的 24 个发达国家的 HL_{in} (见表 2)。

表 2 网络信息的 HL_{in}

国别	HL_{in} (天)	国别	HL_{in} (天)
卢森堡	4.40	意大利	31.94
荷兰	4.82	美国	35.06
挪威	5.16	芬兰	37.13
瑞典	6.13	希腊	44.72
英国	6.88	葡萄牙	46.29
新西兰	6.93	澳大利亚	47.99
丹麦	8.49	西班牙	49.03
瑞士	11.82	冰岛	54.97
德国	16.27	日本	58.18
比利时	21.32	韩国	63.68
奥地利	25.73	爱尔兰	65.07
中国	26.89	加拿大	83.05
法国	29.06	—	—

3.2 讨论

(1)在向 Yahoo! Site Explorer 提交的获取指定 URL 的 Inlinks 信息的指令中,我们将参数 omit_inlinks 的值置为“domain”,以排除来自与 URL 相同域的网页。其目的是排除结构性链接。

(2)取 $T_{Inlink} - T_{Sample}$ 的数据是因为只有了解样本,指向样本的链接才具有效用计量价值,所以, T_{Inlink} 只能晚于 T_{Sample} 。如果 $T_{Inlink} - T_{Sample} \leq 0$,说明样本修改后,含有指向它的链接的文档没有随之进行修正。当然,这包括两种情况:一是没必要修正;二是有必要修正,但没有修正。如属于前者,排除 $T_{Inlink} - T_{Sample} \leq 0$ 的数据会造成数据缺失。

(3)Yahoo! Site Explorer 对于每条检索指令返回的数据集最多包括 1000 条记录,这会丢失部分 Inlink 信息。换言之,我们无法控制和保证数据的完整性。

利用 SPSS 分析 HL_{out} 和 HL_{in} 的相关关系,结果 Pearson Correlation 数值为 -0.078 (Sig. (2-tailed) = 0.712),表明两者不具有统计意义上的

相关关系。对此,有待继续深入探讨。

情报学家劳恩指出:“文献增长得越快,文献的半衰期就越短。”^[6]这意味着对于互联网处于不同发展阶段的国家,网络信息半衰期可能有显著差异。根据全球互联网的发展状况,目前普遍认为中国的互联网正处于成长期,而发达国家基本进入了成熟期。因此,可以得到推论:中国网络信息的半衰期会比其他国家短。表1的结果与该推论相符,而表2中的数据未支持该推论。同时,考虑到计算 HL_{in} 的数据完整性相对较差,所以,我们认为 HL_{out} 的结果更符合实际情况,对网络信息的质量和效用评价具有一定的参考价值。

参考文献:

- [1] Stephen P. Harter, Charlotte E. Ford. Web-based analyses of E-journal impact: Approaches, Problems, and Issues[J]. Journal of the American Society for Information Science, 2000, 51(13): 1159-1176.
- [2] 邱均平, 李江. 链接分析与引文分析的比较(上接第8页)
- [4] 转引自:俞君立. 中国文献分类法百年发展与展望[M]. 武汉:武汉大学出版社,2002:15.
- [5] 黄俊贵. 明确图书馆本质 体现图书馆本质:从“图书馆”定义说开去[J]. 新世纪图书馆,2006(4):3-6,11.
- [6] 杜小宁. 德里达中国讲演录[M]. 北京:中央编译出版社,2003:50-51.
- [7] 范并思. 论重构图书馆学基础理论的体系[J]. 图书馆论坛,2007(6):43-48.
- [8] 傅荣贤. 论中国图书馆学学术思想输出的必要性和可能性[J]. 情报资料工作,2008(1):34

[J]. 中国图书馆学报, 2008(1):60-64.

- [3] Jeremy Carriere, Rick Kazman. WebQuery: Searching and visualizing the web through connectivity[OL]. [2008-07-15]. <http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html>.
- [4] Sergey Brin, Lawrence Page. The Anatomy of a large-scale hypertextual web search engine[OL]. [2008-07-15]. <http://infolab.stanford.edu/~backrub/google.html>.
- [5] Jon M. Kleinberg. Authoritative sources in a hyper-linked environment[OL]. [2008-07-15]. <http://www.cs.cornell.edu/home/kleinber/auth.pdf>.
- [6] Line, M.B. The 'half-life' of periodical literature: apparent and real obsolescence[J]. Journal of Documentation, 1970,26(1):46-54.

段宇锋 华东师范大学商学院信息学系副教授,博士。通讯地址:上海市闵行区吴泾镇永德路228弄41号401室。邮编200241。

(收稿日期:2008-10-26)

-37.

- [9] 傅荣贤. 具体文献在古代书目结构中的作用及其对现代知识管理的启示[J]. 图书情报工作, 2007(12):120-123.
- [10] [德]恩斯特·卡西尔. 符号神话文化[M]. 李小兵,译. 北京:东方出版社,1988:29.

李满花 黑龙江大学图书馆馆员。通讯地址:哈尔滨市南岗区学府路74号。邮编150080。

(收稿日期:2008-09-10;修回日期:2008-11-03)