不同搜索引擎在网络影响因子分析中的比较研究

吴茵茵

(华南农业大学 图书馆，广东 广州 510642)

摘 要：网络影响因子是网络计量学研究中的一个重要分支，搜索引擎在网络影响因子的研究中起着重要的作用。本文利用三种搜索引擎对中国 10 所大学的总网络影响因子进行了分析，并对这几种搜索引擎进行了对比性研究。

关键词：网络计量学；搜索引擎；网络影响因子

中图分类号：G350　　**文献标识码**：A　　**文章编号**：1007-7634 (2005) 03-0431-05

Comparison Research of Different Search Engines in the Analysis of Web Impact Factors

WU Yin - yin

(*Library of South China Agriculture University*, *Guangzhou* 510642, *China*)

Abstract: Analysis of web impact factor is an important embranchment of webometrics, and search engine play a critical role in the research of web impact factor. This paper uses three search engines to analyze and compare the web impact factors of ten China mainland universities.

Key words: webometrics; search engines; web impact factor

1 引　　言

网络计量学是随着计算机网络技术的发展以及信息资源的数字化与网络化的背景下诞生的。网络计量学的概念见于 T.C.Almind 与 Peter Ingwersen 在 1997 年发表在的论文《万维网上的信息计量分析：网络计量学方法门径》，英文为 Webometrics。文中指出：信息计量方法完全可以应用于万维网，只不过是将万维网看作引文网络，传统的引文由网页来代替[1]。同年，网络电子期刊 Cybermetrics 在西班牙马德里创刊，该刊是科学计量学、信息计量学和文献计量学的国际电子期刊。自此，网络计量学的研究蓬勃发展起来，逐渐从传统的信息计量学分离出来而形成了一门新兴的独立学科。

作为网络计量学研究中的一个重要分支，网络影响因子的概念是从文献计量学中的影响因子的概念引申而来。在文献计量学中，影响因子就是在两年内某期刊上发表的所有文章被引用的总次数与这两年该期刊上所发表的文章总数的比值。如果把链接网页总数看作是文章的引用次数，把网页总数看作是文章总数，那么就得到了 Ingwersen[2] 于 1998 年提出的网络影响因子的定义：网络影响因子可以表示为某一时刻链接到网络上某一特定网站或区域的网页数与这一网站或区域本身所包含的网页数的比值。Mike Thelwall 在其 2000 年的发表的研究论文中则提出网络影响因子是链接到某网站或特定区域的网页数与该网站或区域的大小之比[3]。该网站或区域的大小并不一定局限与该网站的网页数，也可以用其他指标来进行衡量，如对学术机构的网站进

收稿日期：2004 - 04 - 23

作者简介：吴茵茵 (1979 -)，女，广西湛江人，助理馆员，华南农业大学图书馆信息部.

行研究时也可以是该机构的研究人员数、该机构的研究经费或者全日制学生的数目。就像传统的影响因子可以对于期刊、科学家、研究机构等进行评价一样，网络影响因子也可以用于对网站进行评价。根据链接网页性质的不同，可以分为总链接数、外部链接数、内部链接数，相应的网络影响因子也就据此分为总网络影响因子、外部网络影响因子和内部网络影响因子。

2 搜索引擎在网络影响因子研究中的现状

作为网络计量学研究的最基本的研究工具，搜索引擎在网络影响因子的研究当中同样有着不可替代的作用。目前搜索引擎的数量很多，较为著名的有 AltaVista、Fast、Google、Yahoo、Go、Infoseek、Excite 等。考虑到网络计量学研究的需要，多数学者一般选用 AltaVista 和 Fast 两种搜索引擎来进行研究。这是由于 AltaVista 和 Fast 能够提供多种类型的限制检索，如主机名限制、超链接限制、域名限制、文件类型限制、新闻组限制、主题限制等。此外，AltaVista 还提供布尔逻辑检索、截词检索、字段限制检索、日期限制检索、范围限制检索、动态分类检索、指定语种检索、位置检索等多种检索功能。目前，国外学者多采用 AltaVista 搜索引擎来进行研究。例如 Ingwersen 在 1997 年选择了 7 个国家、4 个顶级域名和 6 个学术机构的网站，利用 AltaVista 的高级检索功能得到了网络影响因子的数值[4]。Alastair G.Smith 在 1998 年 10 月选取了澳大利亚的大学网站和电子期刊网站，也是利用 AltaVista 搜索引擎，对大学或者研究机构与电子期刊的网络影响因子进行了评价。此外，Alastair G.Smith 还与 Mike Thelwall 利用自己设计的爬行器和商业搜索引擎 AltaVista 对英国、澳大利亚和新西兰大学之间的相互连接部分进行了统计[5]。Owen Thomas 和 Peter Willett 在 1999 年对英国大学图书情报学系网站的网络影响因子做了分析，所使用的工具同样是 AltaVista 搜索引擎[6]。国内有些研究者还采用了 Fast 搜索引擎对影响因子进行了分析，如杨涛利用 Fast 搜索引擎，对中国大陆 20 所大学网站的总网络影响因子、外部影响因子、总科研影响因子、科研网络影响因子和教育网影响因子进行了统计分析[7]。Google 搜索引擎虽然高级搜索功能有限而限制了它在网络影响因子方面的应用，但是由于其强大的网络信息搜索能力，对于计算总网络因子的研究来说也是一种有力的工具。

虽然搜索引擎在网络影响因子的评价当中发挥了重要的作用，但是由于搜索引擎功能上的差异，其网络影响因子的实用价值并未得到证明。尤为重要的是，目前的研究当中多采用单一的搜索引擎，而不同搜索引擎在网络影响因于研究中的差异性尚未引起足够的重视。本文拟采用最常用的 AltaVista 和 Fast 以及功能强大的 Google 索引擎，对中国大陆 10 所大学网站的总网络影响因子进行分析，来比较搜索引擎在中文信息处理中的差异性，并以此来说明搜索引擎在网络计量学研究中的应用。

3 搜索引擎的基本原理

搜索引擎是搜索数据的程序，在 Web 环境下它搜索由“机器人”聚集的 HTML 文件的数据库。目前采用的是自动跟踪标引软件，一般称之为“机器人”（Robot）、“蜘蛛”（Spider）、“Web 爬行器”（Webcrawer）、“漫游者”（Webwanderer）或者“蠕虫”（Worm）等[8]。

搜索引擎的原理，可以看做三步：从互联网上抓取网页→建立索引数据库→在索引数据库中搜索排序。首先，利用能够从互联网上自动收集网页的机器人程序，自动访问互联网并沿着任何网页中的所有 URL 爬到其它网页，重复这过程，并把爬过的所有网页收集回来。其次，由分析索引系统程序对收集回来的网页进行分析，提取相关网页信息，根据一定的相关度算法进行大量复杂计算，得到每一个网页针对页面内容中及超链中每一个关键词的相关度（或重要性），相关度越高，排名越靠前，并用这些相关信息建立网页索引数据库。然后，当用户输人关键词搜索后，由搜索系统程序从网页索引数据库中找到符合该关键词的所有相关网页。最后，由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户[9]。

4 不同搜索引擎在网络影响因子分析中的对比

（1）数据来源与搜索方法。10 所中国大学来自于 2002 年中国大学排行榜的前 10 名，依次为清华大学、北京大学、浙江大学、复旦大学、南京大学、华中科技大学、武汉大学、西安交通大学、吉

林大学和上海交通大学。上述数据来自于武书连和他所在的中国管理科学研究院科学学研究所发布中国的大学排名。

(2) 网络影响因子计算方法与数据检索方法。由于 Google 搜索引擎无法分析外部和内部网页链接总数，为了统一比较的方便，本次试验当中只研究总网络影响因子，即某一时刻链接到网络上某一特定网站或区域的网页总数与这一网站或区域本身所包含的网页总数的比值。

对于不同的搜索引擎，有不同的搜索语法。下面以清华大学为例来说明三种搜索引擎的搜索语法，如表 1 所示。

表 1　三种搜索引擎的数据收集方法

搜索引擎	搜索项目	搜索语法
AltaVista	网页总数	domain: www. tsinghua. edu. cn
	链接总数	link: www. tsinghua. edu. cn
Fast	网页总数	Word Filters: Must include www. tsinghua. edu. cn in the host
	链接总数	Word Filters: Must include www. tsinghua. edu. cn in the link to URL
Google	网页总数	site: www. tsinghua. edu. cn
	链接总数	link: www. tsinghua. edu. cn

(3) 搜索结果。表 2 是本次搜索得到的结果。本次试验的是 2004 年 4 月 4 日，所有的检索任务是在上午 10 - 12 时完成的。

表 2　10 所大学网站的各种链接数和总影响因子数值

搜索引擎	学校	网页总数	链接总数	影响因子	学校	网页总数	链接总数	影响因子
AltaVista	清华大学	167	2257	13.51	华中科技大学	101	108	1.07
	北京大学	1154	1255	1.08	武汉大学	265	242	0.91
	浙江大学	377	386	1.02	西安交通大学	309	219	0.71
	复旦大学	329	341	1.04	吉林大学	114	108	0.95
	南京大学	437	496	1.14	上海交通大学	151	167	1.11
Fast	清华大学	745	394	0.53	华中科技大学	61	36	0.59
	北京大学	820	494	0.60	武汉大学	131	93	0.71
	浙江大学	213	132	0.62	西安交通大学	120	99	0.83
	复旦大学	203	125	0.62	吉林大学	136	63	0.46
	南京大学	257	146	0.57	上海交通大学	73	53	0.73
Google	清华大学	28400	3010	0.11	华中科技大学	2360	449	0.19
	北京大学	46800	3240	0.07	武汉大学	903	481	0.53
	浙江大学	23600	882	0.04	西安交通大学	1790	778	0.43
	复旦大学	3660	1190	0.33	吉林大学	7860	488	0.06
	南京大学	14800	557	0.05	上海交通大学	866	776	0.90

在表 2 中，我们可以看出三种搜索引擎所得到的网页总数、链接总数和总网络因子数值有着明显的差别。需要特别指出的是，AltaVista 搜索引擎所检索到的清华大学的网页总数数目明显过少，链接总数与网页总数的差异过大，这造成了它的影响因子远远大于其他学校。造成这种现象的原因尚不清楚，是否是 AltaVista 搜索引擎的 BUG 尚无法确定。从网页总数的数量来看，Google 得到的网页总数最多，其次为 AltaVista，网页总数最少的为 Fast。就链接总数而言，同样是 Google、AltaVista 和 Fast 由多到少的顺序。从学校的分布来看，清华大学和北京大学的网页总数和链接总数均为最高（AltaVista 收录的清华大学的网页总数非常少是例外）。从数值来看，AltaVista 所收录的网页总数在 1000 以下的大学占 90%，而利用 Fast 检索到的 10 所大学的网页总数均在 1000 以下。这一方面说明了 AltaVista 所收集到的中文信息的数量要大于 Fast，相比之下要更适合于中文网站的网络信息计量学研究。另一方面两种搜索引擎所检索到的网页最多不超过 3000，说明了 AltaVista 和 Fast 两种搜索引擎尽管在英文网站的检索方面功能十分强大，但是在收录中文网站方面还做的很不够。这就导致了在利用这两种搜索引擎研究中文网络信息的时候无法得到足够完整的信息，从而说明网络计量学研究当中语言的特殊性对于结果有着明显的影响。此外，还有一种可能是网站建设当中大量应用的 Flash 技术或者是网站上的蠕虫病毒增加了搜索引擎蜘蛛深入爬行的难度，使得被收录的网页总数很少。相比之下，Google 搜索引擎所收录的网页数目要大的多，所收录的网页总数在 1000 以下的大学仅有 20%，从而说明利用 Google 来进行中文信息的检索是合适的。但是。由于 Google 搜索引擎不支持一些高级搜索，比如无法得到站内链接数目和站外链接数目，使得无法利用 Google 分析外部和内部网络影响因子，所以在网络影响因子的分析过程中应用性就大打折扣。

在总网络影响因子方面，整体上来说在数值上并无明显性的区别（AltaVista得到的清华大学网络影响因子过大例外）。由于直接比较数值的大小无法反映数据的细微变化趋势，因此为了更为明显的比较三种搜索引擎所得到的网络影响因子，笔者对数据分别进行了归一化处理，使得10所大学的总网络影响因子的变化趋势清晰的表现出来，具体如图1所示。为了显示的方便，大学的名称均采取了简称处理。

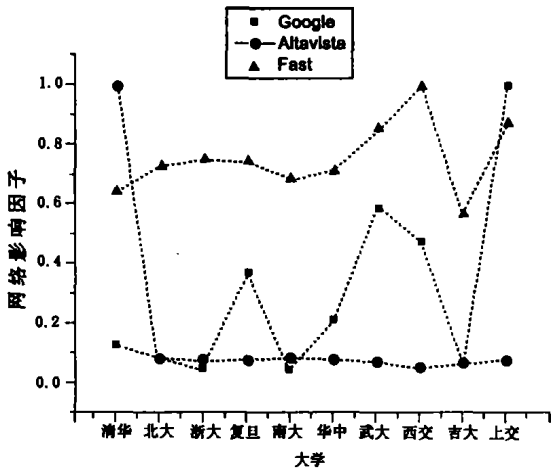

图1 基于三种搜索引擎的10所大学的归一化网络影响因子

从图1中可以看出，三种搜索引擎所得到的网络影响因子并无相同的变化趋势。对AltaVista来说，其变化趋势为一个L型曲线，除了清华大学的影响因子一枝独秀之外，其他的都相差无几。对Fast来说，其变化趋势分为两段，第一段基本上表现为一条变化较小的直线，从华中科技大学开始则表现为一个N型曲线，其波动性非常明显。对Google来说，其变化趋势则接近于WV曲线，其变化波动非常剧烈。三种搜索引擎中，网络影响因子最高的分别为清华大学、西安交通大学和上海交通大学，最低的分别为西安交通大学、吉林大学和南京大学（对Google来说，浙江大学和南京大学的总网络影响因子并列最低）。作为中国公认的实力最强的清华大学和北京大学在总网络影响因子的计算中并没有表现出其强势的地位，尽管其网页总数和链接总数是最多的。

三种搜索引擎在网络影响因子的变化趋势上各不相同，使得我们无法确定哪一种搜索引擎所得出的网络影响因子更具有价值。因为从理论上来讲，尽管具体数值会存在一定的差异，但是只要变化趋势是相同的，就说明所得到的网络影响因子的曲线是有着实际意义的，那就是反映了各大学网站真实的影响力。然而排除掉大学排名的因素，这三种搜索引擎所得到的总网络影响因子的变化趋势各不相同，说明在目前的情况下总网络影响因子的还不能用于对大学的评价。这与杨涛在研究20所中国大学网络影响因子时所得出的结论是一致的。由于Google搜索引擎的限制，本次试验没有进行外部影响因子和内部网络影响因子的分析。但是根据国内外类似的研究，现在的条件下没有任何一个影响因子的作用远远超过其他影响因子，其网络影响因子对于大学评价的可靠性和正确性仍值得商榷。

需要特别指出的是，网络影响因子的有效性和可靠性除了有着本身定义上的局限性之外，其结果还受到搜索引擎性能的影响。本次试验当中使用的都是商业搜索引擎，它们存在的主要问题是覆盖度和稳定性的问题。由于搜索引擎只能对网络的一部分进行标引，所以网络覆盖度非常有限。Lawrence和Giles估计任何搜索引擎的网络覆盖度都不会超过16%[10]。如果单纯计算网络影响因子为目的，就必须保证搜索引擎有比较高的覆盖度。而在搜索引擎的稳定性问题上，Rousseau的实践序列分析结果表明以往网络影响因子研究中经常使用的AltaVista，其检索结果每天的波动性很大，稳定性很差[11]。这是由于商业搜索引擎处于商业化目的的考虑，把检索的响应速度放在首位而牺牲了学术研究所需要的长期稳定性、可重复性和高覆盖度[12]。国外的一些学者们设计了专门的网络爬行器，并已经用于网络影响因子的分析，取得了较好的效果[13]。相比之下，网络爬行器有着针对性强和搜索范围广等优点。在以后的工作当中，可以将搜索引擎的研究结果和专门的网络爬行器得到的结果进行对照，来比较两者在网络影响因子分析当中的优劣性。

5 结 语

根据以上的研究，我们可以得出以下结论。

（1）各种搜索引擎所得到的网络影响因子变化曲线各不相同，决定了目前情况下根据网络影响因子来分析各大学网站的影响力是不可行的，以此来对各大学的综合实力进行排名尚缺乏足够的科学依据。

（2）国外搜索引擎对于中文网站信息的处理能力有待加强。虽然AltaVista和Fast经常被国外学者

用于网络影响因子的分析，但是本次试验所得到的数据表明它们对中文网站的信息处理能力上面还有较大的缺陷。搜集数据的完整性较 Google 而言还有很大的差别。后者虽然能够获得更为全面的信息，但是目前还尚未能够满足网络影响因子更为复杂的分析要求。

(3) 鉴于商业搜索引擎的缺陷，将搜索引擎和专门的网络爬行器所得到的信息进行对比和综合，对准确的分析网络影响影子可能更加有帮助。

(4) 研究方法有必要作进一步的改进。目前对链接网页的处理上，每一个网页的价值都是相等的，即权值相同。严格意义上来讲，这是不确切的。对于链接网页应该根据网页所在的域名或者不同的链接动机，应该赋予不同的权值。这样所得到的网络影响因子可能更有价值。

(5) 在利用搜索引擎对网络计量学进行研究时，必须充分的考虑到中文信息搜索和英文或者西欧语言搜索方面的差别。语言的特殊性决定了搜索能力的差别，从而在很大程度上影响了所获得信息的完整性，这一点在 AltaVista 搜索引擎检索到清华大学的网页总数过少得到了证明。因此，开发能够适应中文信息处理并具备高级检索功能的搜索引擎，对网络计量学的研究是非常必要的。

参考文献

[1] Almind, T. C. and Ingwersen, P. Informetric analyses on the World Wide Web: methodological approaches to "Webometrics" [J]. Journal of Documentation, 1997, 53 (4): 404 - 426.

[2] P. Ingwersen. The calculation of web impact factors [J]. Journal of Documentation, 1998, 54 (2): 236 - 243.

[3] Thelwall, M. Web impact factors and search engine coverage [J]. Journal of Documentation, 2000, 56 (2): 185 - 189.

[4] Smith A. G. A tale of two web spaces: comparing sites using web impact factors [J]. Journal of Documentation, 1999, 55 (5): 577 - 592.

[5] Smith, A. and Thelwall, M. Web impact factors and university research links. Proceedings of the 8th International Conference on Scientometrics & Infometrics [C]. Australia, 2001. 16 - 20.

[6] Thomas, O. and Willett P. Webometric analysis of departments of librarianship and information science [J]. Journal of Information Science, 2000, 26 (6): 421 - 428.

[7] 杨　涛. 网络信息计量学实证研究：对国内 20 所个大学网站的分析 [J]. 图书情报工作，2003，(9)：61 - 66.

[8] 林　芳. 10 个 WWW 搜索引擎的比较研究 [J]. 图书情报工作，1999，(5)：37 - 40.

[9] http: //www. sowang. com/yyy1. htm

[10] Lawrence S, Giles C L. Accessibility of Information on the Web [J]. Nature, 1999, (40): 107 - 109.

[11] Rousseau R. Daily Time Series of Common Single Word Searches in Alta Vista and Northernlight [J]. Cybernetrics, 1999, 2 (3): 105 - 110.

[12] 刘春茂，王　琳. 网络影响因素研究的动态分析 [J]. 情报理论与实践，2004，27 (1)：65 - 71.

[13] Thelwall, M. Results from a Web Impact Factor Crawler [J]. Journal of Documentation, 2001, 57(2): 177 - 191.

(责任编辑：刘凤勤)