用于链接关系检索的搜索引擎的比较研究

吕俊生　　杨金凤

(中国科学院资源环境科学信息中心　兰州　730000)　(甘肃省科学技术信息研究所　兰州　730000)

摘　要　对用于网络链接关系检索的搜索引擎,从检索功能、检索表达式、基本性能等方面进行了系统的调研分析,并对几个搜索引擎进行了比较研究,提出了用于链接分析的搜索引擎的选择方案。

关键词　搜索引擎　链接分析　网络计量学　比较研究

加菲尔德创立的引文索引揭示了由文献群体及其相互引证关系形成的引文网络。通过对引文网络中各种要素及其相互关系的研究,描绘出文献的增长、老化以及其它分布规律,还可以揭示一个学科中重要文献之间的关系,进而反映学科结构的发展演变。引文方法已经成为科研管理中一个重要的定量评价方法。

与引文网络相似,在互联网中无数的站点(或网页)及其相互指引的链接关系也形成了一个网络,而 WWW 和搜索引擎的发明则可将这种隐含的链接关系揭示出来。在互联网中,站点(或网页)A 链接了站点(网页)B,通过这一链接引导人们进入 B 站点(网页),在网络中实际就是站点(网页)A 引用了站点(网页)B,这反映了站点(网页)A 和站点 B 之间存在着内在的联系。

网络链接关系需要借助于搜索引擎来予以揭示,但并不是所有的搜索引擎都具备这一功能,即使具备这一功能,也还存在着较大的差异。本文着重对几个重要的能检索链接关系的搜索引擎进行评价,以发现各自的特点和优势。

1　主要的能用于链接关系检索的搜索引擎概述

许多著名的搜索引擎都能检索某一网页的被链接情况,可直接用于检索链接关系。目前国内的一些著名的搜索引擎(如新浪、搜狐、21 世纪)采用的都是百度搜索技术,百度技术不能对链接情况进行检索,网易则采用的是 Google 技术。本文对搜索引擎的评价主要针对国外。

国外著名的搜索引擎大都支持多语种检索,能对中文网站的被链接情况进行检索。著名的搜索引擎专业评价网站 Searchenginewatch 重点推荐了 5 家能检索链接情况的网站: Alta Vista, Fast Search, Google, Hotbot(Inktomi 技术支持), Northlight。Northlight 是收费的搜索引擎,Alta Vista 目前在我国无法连通,其它 3 家可用于检索链接关系的检索(详情见表 1 和表 2)。

4 个搜索引擎中,除了 Alta Vista 目前在我国无法使用外,其它 3 个都可免费使用。

表 1　四个可检索链接关系的免费搜索引擎基本情况表

搜索引擎	网址	创建年代
Fast Search	www. alltheweb. com	1999
Google	www. google. com	1999
Alta Vista	www. altavista. com	1995
Inktomi	http://www. hotbot. com/ http://search. msn. com/	1996

表2　几个搜索引擎的主要门户站点:

搜索引擎	门户站点	网址
AltaVista	AltaVista	www. altavista. com
Northern Light	Northern Light	http://www. northernlight. com/
Google	iWon	http://home. iwon. com/index_ gen. html
	AOL Search	http://www. aol. com/
	Google	www. google. com
Fast	LYCOS	www. lycos. com/default. asp
	alltheweb	www. altavista. com
Inktomi	HotBot	http://www. hotbot. com/
	MSN Search	http://search. msn. com/

2　基本功能指标状况[1]

表 3　四个搜索引擎的基本功能比较

搜索引擎	支持语种	简体中文检索/显示	区分站内链接
Fast Search	49	能/能	能
Google	35	能/能	不能
Alta Vista	25	不能/能	能
Inktomi	38	不能/不能	能

从以上 3 个指标比较来看,Fast Search 支持的语种最多,能同时进行简体中文检索与显示,有区分站内和站外链接关系检索的功能,其三项综合功能最优。

3　链接关系检索式比较[2]

表 4　几个搜索引擎的检索式比较

名称	链接检索标识符	举例	站内排除标识符
AltaVista	Link:	link: cheminfo. gov. cn	- url:
Northern Light	Link:	link: cheminfo. gov. cn	- url:
Google	Link:	link: cheminfo. gov. cn	无此功能
Fast	link. all:	link. all: cheminfo. gov. cn	- site:
Inktomi	linkdomain:	linkdomain: cheminfo. gov. cn	- domain:

基金项目:受中国科学院资源环境科学信息中心主任基金项目资助(项目名称:网上化学化工经济信息资源的评价与服务)。

作者简介:吕俊生,男,1957 年生,副研究馆员,从事信息资源建设与信息服务研究;杨金凤,女,1964 年生,副研究馆员,从事科技查新工作。

表 4 反映了 5 个搜索引擎链接关系检索表达式, 其中只有 Google 不能排除站内链接。

4 链接数量检索结果比较

表 5 搜索到的链接数量比较(国外)

搜索引擎	网址	检索对象(国外)	站内链接数	站外链接数	链接总数
Fast Search	www.alltheweb.com	cosmos.com.mx	67331	1775	68491
Google	www.google.com	cosmos.com.mx			185
Alta Vista	www.altavista.com	cosmos.com.mx			69436
Inktomi	http://www.hotbot.com/	cosmos.com.mx		2677	3230
	http://search.msn.com/	cosmos.com.mx		2677	3230

表 6 搜索到的链接数量比较(中国)

搜索引擎	网址	检索目标	站内链接数	站外链接数	链接总数
Fast Search	www.alltheweb.com	cheminfo.gov.cn	227	18924	19157
Google	www.google.com	cheminfo.gov.cn			132
Alta Vista	www.altavista.com	cheminfo.gov.cn			183
Inktomi	http://www.hotbot.com/	cheminfo.gov.cn		1046	1060
	http://www.search.msn.com/	cheminfo.gov.cn		1046	1060

表 5、表 6 反映了无论是对国外网站还是国内网站的链接检索, 几个搜索引擎的检索结果差异极大。其中 Fast Search 的检索结果最多, Inktomi 和 Alta Vista 次之, Google 最少。

著名的链接评价网站(http://www.linkpopularitycheck.com) 检索到的一组数据, 反映了几个搜索引擎数据检索结果的差异[3]。

表 7 MSN

Site	Links Found	Popularity compared to chemnet.com.cn(%)	Details
chemnet.com.cn	6012	100%	Details
www.cheminfo.gov.cn	983	16.3	Details
www.chem.com.cn	932	15.5	Details
www.hgzx.com.cn	338	5.62	Details

表 8 AltaVista

Site	Links Found	Popularity compared to chemnet.com.cn(%)	Details
chemnet.com.cn	5013	100	Details
www.chem.com.cn	880	17.5	Details
www.cheminfo.gov.cn	183	3.65	Details
www.hgzx.com.cn	9	0.17	Details

表 9 AllTheWeb

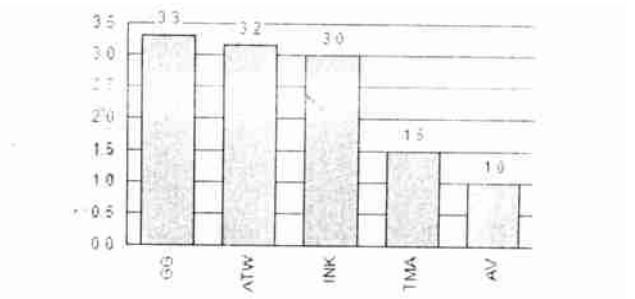
Site	Links Found	Popularity compared to chemnet.com.cn(%)	Details
chemnet.com.cn	8605	100%	Details
www.chem.com.cn	799	9.28%	Details
www.cheminfo.gov.cn	48	0.55%	Details
www.hgzx.com.cn	4	0.04%	Details

2004/2/26 数据

其中, AllTheWeb 对 4 个检索对象数据检索结果的差异最大(100%, 9.28%, 0.55%, 0.04%); AltaVista 次之(100%, 17.5%, 3.65%, 0.17%); MSN 最小(100%, 16.3%, 15.5%, 5.62%), 只是 3 个搜索引擎检索结果对 4 个检索目标大小的排序是一致的。另外, 3 个搜索引擎对 4 个检索对象的检索结果在数量上的差异明显。可见, 采用的搜索工具不同, 其统计分析结果也会有较大的差异。

5 几个搜索引擎的性能比较

5.1 几个搜索引擎的索引量比较[4]



KEY: GG= Google, ATW= AllTheWeb, INK= Inktomi, TMA= Teoma, AV= AltaVista. See the Major Search Engines page for links to these services.

图 1 几个搜索引擎的索引量比较

比较表 4、表 5 和图 1, 搜索引擎索引量的大小与链接关系检索结果在数量上不成正比。Google 的索引量最大, 但搜索到的数量最小。这种差异与矛盾现象对链接分析评价数据获取的客观性提出了质疑, 也为搜索引擎性能的改进提出了要求。

5.2 死链接情况比较[5]

表 10 死链接情况比较

Search Engine	Dead(%)	400 errors only(%)
AltaVista	13.7	9.3
Excite	8.7	5.7
Northern Light	5.7	2.0
Google!	4.3	3.3
HotBot	2.3	2.0
Fast	2.3	1.8
MSN Inktomi	1.7	1.0
Anzwers	1.3	0.7

Greg R. Notess 从 1999 年 1 月 5 日至 2000 年 2 月 21 日, 对上述搜索引擎在一年的时间里进行了 6 次检测, 发现各搜索引擎死链接率大大下降。表 10 是最后一次检测的结果, 反映了死链接的排序情况: MSN Inktomi, Fast, Google, AltaVista。

5.3 搜索结果的新颖情况比较[6]

表 11 搜索结果的新颖情况比较

Search Engine	Newest Page Found	RoughAverage	OldestPage Found
MSN (Ink.)	1 day	4 weeks	51 days
HotBot (Ink.)	1 day	4 weeks	51 days
Google	2 days	1 month	165 days
AlltheWeb	1 days	1 month	599 days*
AltaVista	0 days	3 months	108 days
Gigablast	45 days	7 months	381 days
Teoma	41 days	2.5 months	81 days
WiseNut	133 days	6 months	183 days

注: On one search of the six, AlltheWeb continues to find several extremely old records (Sept. 25, 2001, Feb. 25, 2002, June 29, 2002, Dec. 24, 2002, and Jan. 2, 2003). If these were not included, AlltheWeb's oldest would only be 103 days old.

表 11 是 Greg R. Notess 于 2003 年 5 月 17 日调查统计的结果, 反映了几个搜索引擎检索结果的新颖情况, 除 AltaVista 平均周期较长外, 其他平均在 1 个月之内, 差异不大。

5.4　唯一命中情况比较[6]

表 12　唯一命中情况比较

Search Engine	Unique Hits	Unique(%)	Times
AllTheWeb	7	18	2
AltaVista	0	0	0
Direct Hit	0	0	0
Google	41	43	4
HotBot	0	0	0
iWon (Inktomi)	0	0	0
MSN (Inktomi)	2	7	2
NLResearch	4	15	1
Teoma	3	13	2
WiseNut	14	30	3

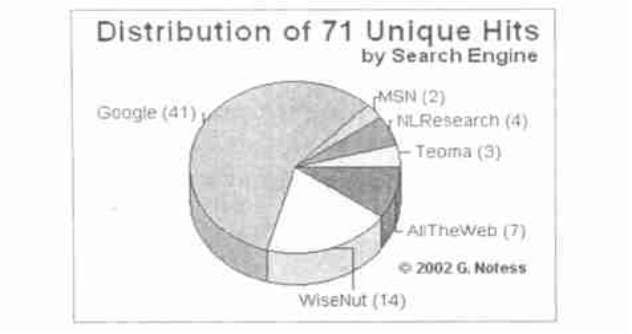


图 2　几个搜索引擎唯一结果占有量

表 12 和图 2 是 Greg R. Notess 于 2002 年 3 月 6 日做的一个唯一命中调研结果, 反映了几个搜索引擎唯一结果的占有量, 占有量越高, 其检索效率越高。其排序为: Google, AllTheWeb, MSN(其他与链接关系检索无关的略)。

6　检索功能分析

6.1　Fast Search　可检字段有全文、题名、URL、主机名、链接到某网页的所有网页(网站) 等 5 种; 运算符有 Must include, Must not include, Shoule include 三种; 匹配方式有全部匹配、部分匹配、精确短语匹配三种。

有地区限定方式、语言限定方式、媒体类型限定方式、文献类型限定方式、文档的扩展名限定方式、域名限制方式(com, gov, dell.com, etc., 如"include.edu.cn"表示将检索结果限制在中国教育科研网内, "Exclude results from"表示排除某一类型域名的结果)。

6.2　Inktomi　检索和检索结果显示不支持中文简体; 有域名和网址限制方式、语种限定方式、地区限定方式、时间设定方式、检索关键词限定方式。

6.3　Google　高级检索功能中的各种限定功能在链接检索中无意义, 不能区分站内检索。

7　总　评

Google: 不能区分站内和站外链接, 高级检索功能中的各种限定功能在链接检索中无意义, 虽然索引量大, 但检索到的链接数量少, 目前尚不适合于用于链接分析研究。

Inktomi: 不支持中文简体的检索, 检索结果不能显示中文简体, 检索结果数量高于 Google, 但大大低于 Fast; 由于链接检索输入的是域名, 检索结果虽不能显示中文简体, 但不妨碍检索结果的数量, 不影响链接分析研究; 各项高级限定检索功能可与链接检索式联合使用, 检准性好。可用于链接分析研究, 尤其是链接分析的比较研究。

Fast Search: 检索功能最全, 高级检索中的各项限制功能都可与链接检索式联合使用, 支持的语言最多, 同时支持中文简、繁体检索和显示, 检索结果的数量最大等。但其致命的弱点是检准性差, 比如检索 www.chem.com(美国化工产品供求及生产商数据库) 被链接的情况, 在检索结果中多数是 www.chem.com.cn (中国万维化工城) 的被链接, 又无法在检索策略中排除 www.chem.com.cn (中国万维化工城) 的结果出现; 但如果输入的是 www.chem.com.cn(中国万维化工城), 在结果中就不会出现 www.chem.com(美国化工产品供求及生产商数据库) 被链接的情况, 所以在检索有顶级域名(国别区分) 的网站时可保证检准率。因此在检索没有顶级域名的网站(网页) 必须小心使用。

AltaVista: 由于目前无法使用, 难以进行调研分析。

8　结　论

目前在国内进行链接分析研究可使用 Fast Search 和 Inktomi。Fast Search 的检索功能最强, 查全率高, 更接近于链接数量的全貌; 但对于无顶级域名的链接检索(尤其是美国的网站(网页), 要格外谨慎。Inktomi 的检索功能较强, 查准率高, 有利于比较研究, 但查全率不足。Google 由于其检索功能的局限, 目前尚无法用于链接分析之用。

参 考 文 献

1　http://www.alltheweb.com/ 2003/10/5/　http://www.google.com/intl/zh-CN/ 2003/10/5/　http://www.hotbot.com/ 2003/10/5/

2　http://www.searchenginewatch.com/ 2003/10/5/

3　http://www.linkpopularitycheck.com/cgi-local/popularity2.cgi?url=chemnet.com.cn&MSN=checked&AllTheWeb=checked&AltaVista=checked&url1=www.cheminfo.gov.cn&url2=www.hgzx.com.cn&url3=www.chem.com.cn&radio_frequency=never&name=&email=&url=%3C%21--url--%3E&url1=%3C%21--url1--%3E&url2=%3C%21--url2--%3E&url3=%3C%21--url3--%3E&display=true&submit=Check+My+Link+Popularity%21 2003/10/5/

4　2003/9/2 数据: http://searchenginewatch.com/reports/article.php/2156481

5　Dead Links Report by Greg R. Notess Data from Feb. 21, 2000 http://www.searchengineshowdown.com/stats/dead.shtml 2003/10/5/

6　http://www.searchengineshowdown.com/stats/freshness.shtml Data from May 17, 2003. by Greg R. 2003/10/5/

7　Unique Hits Report by Greg R. Notess http://www.searchengineshowdown.com/stats/unique.shtml 2003/10/5/

(责编: 愚王京)