CAN GOOGLE'S PAGERANK BE USED TO FIND THE MOST IMPORTANT ACADEMIC WEB PAGES?

Mike Thelwall[1]
m.thelwall@wlv.ac.uk

School of Computing and Information Technology, University of Wolverhampton,
35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK
Phone: + 44 1902 321470; Fax: + 44 1902 321478

Google's PageRank is an influential algorithm that uses a model of Web use that is dominated by its link structure in order to rank pages by their estimated value to the Web community. This paper reports on the outcome of applying the algorithm to the Web sites of three national university systems in order to test whether it is capable of identifying the most important Web pages. The results are also compared to simple inlink counts. It was discovered that the highest inlinked pages do not always have the highest PageRank, indicating that the two metrics are genuinely different, even for the top pages. More significantly, however, internal links dominated external links for the high ranks in either method and superficial reasons accounted for high scores in both cases. It is concluded that PageRank is not useful for identifying the top pages in a site and that it must be combined with a powerful text matching techniques in order to get the quality of information retrieval results provided by Google.

INTRODUCTION

Google's PageRank algorithm (Brin & Page, 1998) for ranking Web pages is an Information Retrieval (IR) algorithm that is relatively well-known to the general public because of its use in the Google Toolbar and in the company's marketing approach, "The heart of our software is PageRank™" (Google, 2002). It is also arguably the most influential and successful of the past five years, on the back of the search engine's number one status for online searching according to some measurements (Sullivan, 2002). Despite this, there do not appear to have been many studies focussing on the questions of how effective it is or under which conditions it is effective. PageRank is based upon the assumption that good quality pages are more likely to be linked to than poor quality ones and therefore that mining information about the link structure of the Web could be more effective at identifying the best pages matching search engine queries than a simple text-matching algorithm. In fact it goes one step further and incorporates the quality of the linking page in its iterative algorithm, described in detail below. In this context two natural questions to ask from a bibliometric perspective are whether the pages that are most highly linked to are significantly different from those that have the highest PageRank, and whether either method is capable of identifying the highest quality or most useful pages in a site. The questions have additional pertinence because of the growing body of informetric research that is based upon link counts (e.g. Larsen, 1996; Rousseau, 1997; Ingwersen, 1998; Smith, 1999; Leydesdorff & Curran, 2000; Thelwall, 2001a,b,f, 2002a; Smith & Thelwall, 2002). Potentially, such investigations may benefit from

[1] Journal of Documentation 58(6), 2002, to appear.

switching to PageRank, or another iterative rating system, in order to take into account in some way the quality of the inlinks rather than just their numbers.

PAGERANK AND OTHER WEB INFORMATION RETRIEVAL ALGORITHMS

On a mathematical level the PageRank algorithm finds the principal eigenvector of a matrix created from the link structure of the system. More descriptively, the matrix encodes the model of a surfer visiting Web pages in succession. At each page the surfer jumps to a completely random page with probability 0.85 and follows a random link chosen from the current page with probability 0.15. If the surfer is allowed to proceed in this fashion from any starting point for a very long time then the PageRank of any page is defined to be the probability that it they are viewing it after any given jump. The ranking system generated favours pages that are the target of many links, since they are more likely to be jumped to. It also weights more highly links from more important source pages since these sources are more likely to be jumped to and, therefore, more likely to originate a new jump. The rationale for the use of links is that they provide additional information about pages that can be used to help decide how important the page is, rather than what its content is about (Brin & Page, 1998). A more mathematical equivalent definition of the Page Rank algorithm can be found in Ng *et al*. (2001) in addition to the original Brin and Page article.

One other key link based IR procedure is Kleinberg's (1999) topic distillation algorithm, which is primarily for topic-specific searching. This uses links to decide how important pages are for a specific topic, rather than in general. It works by starting with a query and identifying relevant pages through text semantics, then using the link structure within this collection to allocate pages iteratively (a) an authority score by summing the weights of the incoming link pages and (b) a hub score by summing the weights of the outgoing link target pages. PageRank has been shown to be intrinsically the more stable of the two, however, with the Kleinberg algorithm being sensitive to small changes in link structures (Ng *et al*., 2001). There is no known study that scientifically demonstrates that either is effective in a clearly defined sense of identifying the best information on the Web, but the success of Google is still a powerful argument for the importance of PageRank. It may well be the case that the various search engine companies perform extensive testing and know the answers to these questions but do not make them available for commercial reasons. The scientific TREC competition results were not promising, however (Hawking *et al*., 2000), although this could have been due to an untypical test corpus or the variant of PageRank used (Thelwall, 2002b). According to Gao *et al*. (2001), "[The recent] research of web retrieval has focused on link-based ranking methods. However, none had achieved better results than content-based methods in TREC experiments". Other research into link-related algorithms serves to confirm the importance of this area (Haveliwala, 1999; Broder *et al*., 2000; Lifantsev, 2000; Rafiei & Mendelzon, 2000; Richardson & Domingos, 2001). Bharat and Mihaila (2001) for example develop a new algorithm and demonstrate through user evaluations that its performance is comparable with PageRank. Unlike PageRank, however, the other Web IR algorithms integrate the text analysis with link analysis, making them unsuitable for tasks such as finding the 'best' overall pages.

THE RESEARCH QUESTIONS

This paper reports on a study to apply PageRank to databases of the link structures of the Web sites of UK, Australian and New Zealand universities. This algorithm is chosen for its arguable pre-eminence in addition to its suitability for the task of finding the overall best pages on a site. The three nations selected are chosen for the free availability of link data for them and because they represent in international terms relatively early Internet adopters and extensive Web users. The 10 highest ranked pages for each university will be analysed as well as the 100 highest for each national system. Reasons for differences between PageRank and inlink counts will be uncovered from an investigation into the inlink structure of the pages in question. This is essentially an investigative and qualitative bibliometric approach (e.g. Gläser & Laudel, 2001; Goodrum *et al*, 2001) rather than one of formal scientific hypothesis testing. The theoretical context is the hypothesis that the top ranked pages will either contain high quality content or will be gateways to other useful pages. The two specific questions addressed are as follows.

- Are the pages given the highest rank by PageRank clearly the most useful or highest quality in the system analysed, or can their high positions be the result of unrelated factors?
- Is PageRank more successful than simple inlink counts at identifying the top pages?

METHODOLOGY

The link structure of the national university systems was obtained from a publicly available database (cybermetrics.wlv.ac.uk/database) described in detail in Thelwall (2001d) and obtained by an information science Web crawler (Thelwall, 2001c). This covers the proportion of each Web site that could be found by iteratively following links from the home page, excluding copies of pages from other sources (mirror sites) when identified. Mirror sites are a particular problem because it is not known to what extent Google's spider crawls them. For example there are numerous copies of Sun's Java documentation on UK university Web sites and ideally Google would ignore these and only crawl the original on the Sun Microsystems website. Any additional copies in Google's database would clearly be wasting space. Nevertheless, identifying and eliminating duplicate pages is a technically challenging job, despite published research on speeding up the process (Heydon & Najork, 1999) because of the sheer size of the Web. The databases used will include some mirror sites that have been missed due to human error, which is possibly similar to the situation for Google. The names of the 156 universities crawled can be obtained from the originating database site, via the domain names files.

The link database consists of a separate file for each institution containing the link structure of its website. The most challenging part of the research was in writing a program to encode the URLs into numbers for the PageRank algorithm. This was difficult because of the memory taken by the URLs and the number of string comparison operations that were required to ensure that each URL was given a unique number. One combined link structure file was constructed for each national system and used to build a matrix of its link structure, and one separate link structure file was also created for each institution. This was then loaded into a new program coded to execute the PageRank algorithm, and ranks obtained from it. The procedure followed was essentially the same as the non-blocked version described by Haveliwala (1999) for small computers, except that no pages were eliminated from the system due to a lack of links. Instead, a correcting factor was incorporated to adjust for the affect of pages in the system without links. Although in the largest case the full link structure matrix would have been too big to store as an array, with 4×10^{14} entries, it could in fact be stored with only 2×10^{7} entries as a sparse matrix, recording the location of the non-zero entries, with unrecorded locations being implicitly zero. The PageRank list was combined with the URL key file and sorted to produce two top 10 lists for each institution, one for PageRanks and one for inlinks. Similar top 100 lists were produced for each whole system. Table 1 summarises basic information about the databases. The UK database is just over 10% of the size of the original Brin & Page (1998) corpus.

Table 1. *Information about the databases used*

Country	Australia	New Zealand	UK
University Web sites included	38	8	110
Crawl dates	10/2001-1/2002	1-2/2002	6-7/2001
Total pages	3,804,612	341,667	6,920,448
Total links	20,054,017	2,119,677	32,516,604

The first analysis was a simple calculation to see whether PageRank was more effective at identifying useful pages than inlink counts. The two assumptions made are that (a) the most useful pages are institutional home pages and that (b) these are normally the root pages of their own domain names. Based upon these assumptions, the test applied was to see which of the two mechanisms ranked this type of URL most highly. As an aside, home page finding is a recognised IR task, for which links have been found useful (Xi & Fox, 2001).

The second analysis is a large combined experiment to determine whether PageRank or inlink counts reveal the most important pages on a site, and whether one appears to be better than the other. The investigation is conducted by using tables of the top pages from both methods and evaluating these qualitatively. Separate results are reported for individual universities and for national university systems. These are potentially significantly different entities under the hypothesis that links between institutions carry a higher information value than those within a single site (Kleinberg, 1999; Thelwall, 2001a).

RESULTS AND DISCUSSION

Home page ranking

As can be seen in Table 2, there is no significant difference between the success of raw inlink counts and PageRank in the rank of the university home page, based only upon the link structure of the university Web site on its own. In almost all of the tied cases the home pages were number one in both lists. In only five cases the home pages were not in the top ten of either list. No statistical test is needed to see that the differences are negligible, but standard hypothesis tests for proportions would show this.

Table 2. *A comparison of the ranking of university home pages produced by PageRank and by simple inlink counts operating on each university Web site on its own*

System	Home pages ranking higher with PageRank	Home pages ranking higher with inlink counts	Home pages ranking the same with both methods	Home pages not in the top 10 in either list
Australia	1	3	33	1
New Zealand	2	1	5	0

UK	12	10	84	4

Individual universities

The top ten lists of individual universities were examined for patterns. In many lists there was a group of closely related URLs that came from a large subsite with a navigation bar linking to the main pages. Often these were the main official pages of the site, as is the case for La Trobe University, shown in Table 3. The dominance of the main pages in this case is caused by the existence of a standard links bar at the top of all official pages. A big difference can be seen between the results from this site and those of Wolverhampton (Table 4), where the main pages were not indexed due to their use of Active Server Pages queries. The official links bar for other pages uses a server side map that is also not indexable, although the URL of the map can be seen ranked third in the table. This is a clear case of design decisions dominating the top results of the PageRank calculation for individual universities.

Table 3. *The ten highest ranked pages for La Trobe, using PageRank on internal links only, with PageRanks linearly scaled to make the largest equal to 1*

Count	PageRank	Page
8952	1	www.latrobe.edu.au
10058	0.513953	www.latrobe.edu.au/international
9910	0.506899	www.latrobe.edu.au/search
9862	0.505285	www.latrobe.edu.au/contact
9966	0.505202	www.latrobe.edu.au/about
9858	0.504549	www.latrobe.edu.au/sitemap
9909	0.50315	www.latrobe.edu.au/teaching
9903	0.502879	www.latrobe.edu.au/research
9903	0.502845	www.latrobe.edu.au/faculties
9902	0.502827	www.latrobe.edu.au/campuses

Table 4. *The ten highest ranked pages in Wolverhampton, using PageRank on internal links only, PageRanks scaled*

Count	PageRank	Page
3171	1	www.wlv.ac.uk/disclaimer/official.html
3037	0.7812036	www.wlv.ac.uk
2802	0.6703956	www.wlv.ac.uk/resources/uni.nav.bar.map
1898	0.5537226	www.scit.wlv.ac.uk/appdocs/php
4474	0.4495484	www.scit.wlv.ac.uk/~cm1914/cp2027/docs/api/overview-summary.html
4475	0.4286861	www.scit.wlv.ac.uk/~cm1914/cp2027/docs/api
4469	0.4277823	www.scit.wlv.ac.uk/~cm1914/cp2027/docs/api/deprecated-list.html
4468	0.427765	www.scit.wlv.ac.uk/~cm1914/cp2027/docs/api/index-files/index-1.html
4468	0.4277458	www.scit.wlv.ac.uk/~cm1914/cp2027/docs/api/help-doc.html
811	0.374287	www.wlv.ac.uk/disclaimer/personal.html

In addition to instances of domination by official pages, other cases were also found where sets of computer documentation or other types of large subsite had high interlinking patterns. This can be seen in Table 4 and is also illustrated for the case of the Royal Melbourne Institute of Technology (RMIT), as shown in Table 5. It can be seen that the main pages on a large PHP: Hypertext Preprocessor (PHP, a recursive acronym) Web page server-side scripting language site have attracted a large number of links, in actual fact from the standard navigation links found on all other pages of this large set of documentation. This is a case where a combination of the sheer size of the resource and its inclusion of a standard set of navigational links have combined to give its key pages a huge inlink count. It appears to be for use only in one student course and is a copy of documentation produced elsewhere, so from an external Web user's point of view it would not be considered as important content on the RMIT site. The fifth page in Table 6 is the home page of the RMIT Research and Development Section, which hosts a large subsite with a link to their home page on each page. This is an example of a similar phenomenon: the internal size of a subsite determining the rank of its home page.

Table 5. *The ten most linked to pages in RMIT, counting only internal links, PageRanks scaled*

Count	PageRank	Page
12485	1	www.rmit.edu.au
6829	0.5016145	www.rmit.edu.au/webmaster/disclaimer.html
3286	0.4388598	www.viscom.rmit.edu.au/robin/talks.htm
2075	0.0274677	kroid.mds.rmit.edu.au/cs843/ref/php/downloads.php
2075	0.0274677	kroid.mds.rmit.edu.au/cs843/ref/php/docs.php
2075	0.0274677	kroid.mds.rmit.edu.au/cs843/ref/php/faq.php
2075	0.0274677	kroid.mds.rmit.edu.au/cs843/ref/php/support.php
2075	0.0274677	kroid.mds.rmit.edu.au/cs843/ref/php/bugs.php
2075	0.0274677	kroid.mds.rmit.edu.au/cs843/ref/php/links.php
2075	0.0274677	kroid.mds.rmit.edu.au/cs843/ref/php/copyright.php

In terms of the difference between high inlink count pages and those with a high PageRank, a comparison of Table 5 and Table 6 shows that there can be real differences. Many of the RMIT pages in Table 5 have a relatively low PageRank as a result of each linking page containing a large number of other links, which dissipates the effect of each individual link through sharing 'rank' between all targets of a page. Some pages in Table 6 have about half as many inlinks but more than double the PageRank because the pages that link to them have fewer overall links.

Table 6. *The ten highest ranked pages in RMIT, using PageRank on internal links only, PageRanks scaled*

Count	PageRank	Page
12485	1	www.rmit.edu.au
6829	0.5016145	www.rmit.edu.au/webmaster/disclaimer.html
3286	0.4388598	www.viscom.rmit.edu.au/robin/talks.htm
1248	0.1176734	www.homepages.eu.rmit.edu.au/bondy/ saskiabondyfamtreesite/persons.html
666	0.0819186	www.rmit.edu.au/departments/rd
1097	0.0796801	bonza.rmit.edu.au
1084	0.0786278	bonza.rmit.edu.au/search.html
1084	0.0786278	bonza.rmit.edu.au/essays
1083	0.0785995	bonza.rmit.edu.au/links
1083	0.0785995	bonza.rmit.edu.au/contact.html

The number one page in Table 4 and the number two page in Table 5 demonstrate another feature of both PageRank and inlink counts: the high score that pages can have which possess a legal function in regard to Web content. At the University of Wolverhampton, the page with the highest PageRank is the legal disclaimer that all official pages are supposed to contain a link to. The RMIT disclaimer page clearly also enjoys a similar status. This is a problem from an IR or bibliometric perspective, as the page does not contain information of unusually high value. There are also several highly ranked copyright pages in other university lists (see tables 8 and 9).

National systems

Tables 7 to 9 give the top 10 pages from each national system, after applying PageRank to their combined link structure files. The top 100 pages were produced in each case but the rest are not shown for reasons of space. Although the university home pages in each list are natural inclusions, none of the other pages could be regarded as containing unusually useful information, rather they owe their position to a relatively ephemeral cause such as the ones discussed above. The UK's top page is a case in point. It attracts only internal links from its own site and is linked to from a huge collection of pages, each containing a description of one of the modules taught at the University of Staffordshire, all of which contain only one link. Ironically, the link appears to be an automatically inserted typo (the home page has an additional underscore: www.staffs.ac.uk/schools/art_and_design) and the link in question is non-functioning because there is no content between the start and end of the anchor tag. The lack of sharing with other links, however, is the main factor that has lead to a high PageRank.

Table 7. *Australian top 10 pages, using PageRank on internal links only, PageRanks scaled, and Google's toolbar value also shown*

Count	PageRank	Toolbar	Page
23304	1	6 (moved)	www.unimelb.edu.au/pwebstats/pwebstats.html
32940	0.3007683	8	www.unimelb.edu.au
44129	0.2149047	7	www.monash.edu.au
22502	0.212978	8	www.unimelb.edu.au/disclaimer
10827	0.1948821	7	www.csse.monash.edu.au/disclaimers/user.html
18157	0.1839759	(not available)	www.gu.edu.au/cgi-bin/textflip.cgi
28977	0.1825789	8	www.uq.edu.au
7525	0.1820394	5	www.educ.utas.edu.au
18989	0.1717772	7	www.unisa.edu.au
34341	0.1686863	8	www.unsw.edu.au

Table 8. *New Zealand top 10 pages, using PageRank on internal links only, PageRanks scaled, and Google's toolbar value also shown*

Count	PageRank	Toolbar	Page
16990	1	3	www.otago.ac.nz/sas/common/images/copyrite.htm
7561	0.2361207	7	www.otago.ac.nz
6611	0.2183746	7	www.vuw.ac.nz
7953	0.2012788	7	www.massey.ac.nz/disclaim.htm
9723	0.189337	7	www.massey.ac.nz

2808	0.185448	5	webview.massey.ac.nz
2807	0.1854373	5	webview.massey.ac.nz/help/help.htm
5254	0.1850909	7	www.canterbury.ac.nz
2185	0.1621233	6	nix.tmk.auckland.ac.nz/SAL
3079	0.1488084	7	www.auckland.ac.nz

Table 9. *UK top 10 pages, using PageRank on internal links only, PageRanks scaled, and Google's toolbar value also shown*

Count	PageRank	Toolbar	Page
3843	1	4 (not available)	www.staffs.ac.uk/schools/art_anddesign
22691	0.9747133	8	www.st-and.ac.uk
17160	0.9735287	0	www.cc.ic.ac.uk/college/onlinedocs/sasonlinedocv8/sasdoc/sashtml/common/images/copyrite.htm
12841	0.9626068	4	bicss.mdx.ac.uk/css/public
26276	0.9160221	7	www.napier.ac.uk
3477	0.9152157	unranked	www.aom.bham.ac.uk/handbook/courses/glossary.htm
3464	0.9122335	3	www.ao.bham.ac.uk/handbook/courses/glossary.htm
27982	0.8607687	7	www.ulst.ac.uk
16851	0.8305826	8	www.leeds.ac.uk
18761	0.7950215	7	www-maths.mcs.st-and.ac.uk

The top Australian page is of a type not mentioned before, a Web statistics software home page, and this particular example is from the former site of Martin Gleeson's pwebstats program that attracts large numbers of links from server statistics pages generated by the software. There are similarly highly inlinked pages in the UK (Thelwall, 2002b). As can also be seen, there are help and glossary pages in the New Zealand and UK lists respectively. These may be useful in the context of the pages that link to them but probably much less so for the wider Web user. Also present are two departmental home pages, both as a result of credit links on large collections of pages. In the case of St Andrews, for example, the links come predominantly from the pages of an online history of maths archive. The ninth ranked New Zealand page is from a mirror copy of he Scientific Applications on Linux site, getting its links from within its own site. The URL is case-sensitive, hence the mixed case version shown in Table 8.

The pages referenced here were all loaded into a Web browser on the 19th of February, 2002 with Google's toolbar (toolbar.google.com) installed so that the PageRank feature could be used. This gives a number between 0 and 10 for each loaded page. Clearly these are not PageRanks as could be obtained directly from Brin & Page's algorithm, since the unmodified values would all be less than one, but the use of this word by Google to describe the displayed figures gives some cause to believe that they are related in a monotonic way (i.e. larger toolbar values come from larger PageRank values). The results of this exercise lead to the discovery that in some cases the toolbar PageRank figure given was applied to the domain and automatically reduced by one for each directory in the path of the URL, so that longer URLs tended to have lower 'PageRanks', irrespective of high inlinking as seen in tables 7 – 9. This was the case for the zero ranked page in the UK list, for example. It is surmised, then, that either the PageRank algorithm has been modified for the current version of Google, or the toolbar uses an approximate non-monotonic version of it in certain situations (i.e. it reverses the relative ranks of some pages).

CONCLUSIONS

It is very clear from the data that the top ranked pages, either with PageRank or with raw inlink counts, are there as a result of navigational architecture policy decisions primarily rather than on their own individual merit. Perhaps the clearest example of this is a comparison of the Ulster University home page with that of Wolverhampton. The former attracts the highest inlink count of all UK pages whereas the latter attracts relatively few as a result of global site design decisions. It must be concluded from this that PageRank and inlink counts are not reliable methods for ascertaining the most valuable resources on a large university Web site or national system of university sites. Contrasting these findings with those of Thelwall (2002b) it can be seen that the root cause of the problems is the inclusion of internal links. Inlink counts based upon external links only yield much "better" results, although still not perfect. This is a real problem for the PageRank algorithm because it depends on internal links to function, for example redistributing link votes from the home page of a site to links on its other pages, as would be needed for PageRank to propagate from an important multipage site, such as the Humbul Humanities Hub (www.humbul.ac.uk).

Comparing the UK top 100 results with those for external links only (Thelwall, 2002b) it can be seen that the two are fundamentally different. Both contain many university home pages but the latter does not contain any of the other pages typically found on the standard internal navigation bar. Perhaps the most damning evidence is that the single page that is probably the most widely used resource on a UK Web site, the UK clickable map, does not appear at all in either UK top 100 reported here, despite having 891 external links from other UK universities.

In the context of the results presented here it is hard to believe that plain PageRank is effective as an IR algorithm, even when combined with simple text matching. The fundamental problem is the allocation of equal weight to internal links as external ones and the loophole that this gives to allow navigational policy decisions to swamp the relatively small number of links created for non-navigational reasons. The algorithm may be more effective on a global scale, where there are relatively more external links but the difference will not be of the order of magnitude needed to make a real difference for university Web sites. It may be, however, that huge sites such as Yahoo! do improve the results by bestowing higher PageRanks on the better Web pages, but this would affect only a few pages on each university site. The original Google patent (Page, 1998) does mention myriad potential customisations of the algorithm, including treating inter-site links differently, and so it is likely that the version in use at the time of testing was different from the original. Another point that should be mentioned is that this analysis has been confined to the top ten or 100 pages of each set and therefore ignores the overwhelming majority of pages. Nevertheless, it can be seen that the same kinds of arguments can also apply for these: pages that are part of a highly interlinking navigational structure will rank much better than others that are the target of only one or two external links, even though such links are probably a much better indicator of high quality. It could be argued that a high degree of interlinking is a good indicator of quality at least in site design, but in this case the PageRank is still totally dependent on the absolute number of pages involved and the extent to which links are also present to other resources.

Despite the number-intensive mathematical algorithms used to produce the data presented here, this has been essentially a qualitative study. This is not seen as a weakness in the context of the very clear-cut nature of the results obtained. Indeed the qualitative approach, focussing on investigating the cause of the problems has enabled the gaining of insights into the reasons why the ranking methods have been unable to produce convincing lists of the top pages in the sites covered.

In conclusion, PageRank is not an effective method for identifying the "best" Web pages in a university system because of its domination by internal links, an argument that would still apply even if all mirror sites had been removed from the data. The astonishing accuracy of Google (Thelwall, 2002c) must be due to its complementary use of a very effective text-based matching algorithm, which must itself be incorporated into the final ranks. A promising future direction for bibliometric research is to develop a variant of PageRank that can harness the potential of its system for transferring rank iteratively through links in a way that would not be dominated by internal site links.

ACKNOWLEDGEMENTS

I would like to thank the referees for their very helpful comments.

REFERENCES

Bharat, K & Mihaila, G. A. (2001). When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics, In: *Tenth International World Wide Web Conference*. Available at http://www.www10.org/cdrom/papers/474/index.html

Brin, S. & Page, L. (1998). The Anatomy of a large scale hypertextual web search engine, *Computer Networks and ISDN Systems*, 30(1-7), 107-117. Available at http://citeseer.nj.nec.com/brin98anatomy.html

Broder, A., Kumar, R., Maghoull, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33(1-6), 309-320.

Gao, J. Walker, S., Robertson, S., Cao, G., He, H., Zhang, M. & Nie, J-Y (2001). TREC-10 Web Track Experiments at MSRA 384-392. TREC 2001. Available: http://trec.nist.gov/pubs/trec10/t10_proceedings.html.

Gläser, J. & Laudel, G. (2001). Integrating scientometric indicators into sociological studies: methodical and methodological problems. *Scientometrics*, 52(3), 411-434.

Goodrum, A. A., McCain, K. W., Lawrence, S. & Giles, C. L. (2001). Scholarly publishing in the Internet age: a citation analysis of computer science literature. *Information Processing and Management*, 37(5), 661-676.

Google (2002). Google Technology. Available: http://www.google.com/technology/. Accessed 3 July, 2002.

Haveliwala, T. (1999). Efficient Computation of PageRank. *Stanford University Technical Report*. Available http://dbpubs.stanford.edu:8090/pub/1999-31

Hawking, D., Bailey, P. and Craswell, N. (2000). ACSys TREC-8 experiments. In: *Information Technology: Eighth Text REtrieval Conference (TREC-8)*, NIST, Gaithersburg, MD, USA, pp.307-315.

Heydon, A. & Najork, M. (1999). Mercator: A scalable, extensible Web crawler. *World Wide Web*, 2, 219-229.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment, *Journal of the ACM,* 46(5), 604-632.

Larson, R. R. (1996). Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace. ASIS 96. Available at: http://sherlock.berkeley.edu/asis96/asis96.html

Leydesdorff, L. & Curran, M., (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy, Cybermetrics, 4. Available at: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Lifantsev, M. (2000). Voting model for ranking Web pages. In Graham, P. & Maheswaran, M. (eds), *Proceedings of the International Conference on Internet Computing,* Las Vegas, Nevada, USA, CSREA Press, pp. 143-148.

Ng, A. Y., Zheng, A. X. & Jordan, M. I. (2001). Stable algorithms for link analysis. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001),* New York: ACM Press, pp. 258-266.

Page, B. (1998). United States Patent 6,285,999. Available: http://patft.uspto.gov/.

Rafiei, D. & Mendelzon, A. O. (2000). What is this page known for? Computing Web page reputations, *Computer Networks,* 33(1-6), 823-835.

Richardson, M. & Domingos P. (2001). The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. Poster at *Neural Information Processing Systems: Natural and Synthetic 2001.* Available at: http://www.cs.washington.edu/homes/mattr/doc/NIPS2001/qd-pagerank.pdf

Rousseau, R., (1997). Sitations, an exploratory study, *Cybermetrics*, 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Smith, A. G. (1999). A tale of two web spaces: comparing sites using Web Impact Factors. *Journal of Documentation*, 55(5), 577-592.

Smith, A. & Thelwall, M. (2002, to appear). Web Impact Factors for Australasian Universities, *Scientometrics*, 54(1-2).

Sullivan, D. (2002). Google Tops In "Search Hours" Ratings. Available: http://searchenginewatch.com/sereport/02/05-ratings.html. Accessed 3 July, 2002.

Thelwall, M. (2001a). Extracting macroscopic information from web links. Journal of the American Society for Information Science and Technology, 52 (13), 1157-1168.

Thelwall, M. (2001b, to appear). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation*.

Thelwall, M. (2001c). A web crawler design for data mining, *Journal of Information Science*, 27(5), 319-325.

Thelwall, M. (2001d). A publicly accessible database of UK university website links and a discussion of the need for human intervention in web crawling, University of Wolverhampton. Available: http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf.

Thelwall, M. (2001e). The top 100 linked pages on UK university Web sites: high inlink counts are not associated with quality scholarly content, University of Wolverhampton.

Thelwall, M. (2001f). Results from a Web Impact Factor crawler, *Journal of Documentation*, 57(2), 177-191.

Thelwall, M. (2002a). A comparison of sources of Links for academic Web Impact Factor calculations. *Journal of Documentation*, 58, 60-72.

Thelwall, M. (2002b). Subject gateway sites and search engine ranking, *Online Information Review*, 26(2), 124-138.

Thelwall, M. (2002c, to appear). In praise of Google: finding law journal Web sites, *Online Information Review,* 26(4).

Xi, W. & Fox, E.A. (2001). Machine Learning Approach for Homepage Finding Task. TREC 2001, pp. 686-697. Available: http://trec.nist.gov/pubs/trec10/t10_proceedings.html.