Subject gateway sites and search engine ranking

Mike Thelwall

The author

Mike Thelwall is a Senior Lecturer in the School of Computing and Information Technology, University of Wolverhampton, Wolverhampton, UK.

Keywords

Internet, Information retrieval

Abstract

The spread of subject gateway sites can have an impact on the other major Web information retrieval tool: the commercial search engine. This is because gateway sites perturb the link structure of the Web, something used to rank matches in search engine results pages. The success of Google means that its PageRank algorithm for ranking the importance of Web pages is an object of particular interest, and it is one of the few published ranking algorithms. Although highly mathematical, PageRank admits a simple underlying explanation that allows an analysis of its impact on Web spaces. It is shown that under certain stated assumptions gateway sites can actually decrease the PageRank of their targets. Suggestions are made for gateway site designers and other Web authors to minimise this.

Electronic access

The research register for this journal is available at
http://www.emeraldinsight.com/researchregisters

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/1468-4527.htm

Online Information Review
Volume 26 · Number 2 · 2002 · pp. 101–107
 · ISSN 1468-4527
DOI 10.1108/14684520210424566

Introduction

The Web is a key medium for scientific communication, containing not only e-journals and conference proceedings but also information about the investigations of individual scholars, research groups and departments. Two of the most important mechanisms for finding such information are commercial search engines and subject gateways. The latter can be formal, for example an organised directory of links created as part of a government sponsored project, or highly informal, for example a links list maintained by an enthusiastic individual. While gateways and commercial search engines are both important, the way in which the two interact is not well known. Google is one of the most popular current search engines and is believed to be unique in publishing the core of its page-ranking algorithm, which is based on the link structure of the Web. As described below, it is believed that other engines use similar link-based approaches to help ranking. Given that gateway sites are link based, an important question to ask is what effect they have on Google's PageRank algorithm, and the implications of this for academic Web site designers.

There are numerous subject gateway sites in the UK and elsewhere in the world, and their importance has promoted at least one meta-level review of them (Saito and Onodera, 2001) as well as a survey in another article (MacLeod *et al.*, 1998). One further academic study focused specifically on collaboration between gateways in Europe (Huxley, 2001). Further evidence of the importance of subject gateways can be found from a recent survey of the best linked to UK academic Web sites, which found four in the top 100 as well as five other external links pages (Thelwall, 2001). The significance of PageRank has also fostered new research, for example: applying it to a standard set of data in competition with other algorithms (Hawking *et al.*, 2000); transferring its characteristic stability to another algorithm, Kleinberg's (1999) HITS (Ng *et al.*, 2001); using a related algorithm to identify topics for which a page is authoritative (Rafiei and Mendelzon, 2000); and investigating its

Refereed article received 11 December 2001
Approved for publication 7 January 2002

suitability for transferring to the computers of Web users (Haveliwala, 1999). As far as is known, however, no research has focused upon the implications of PageRank for gateway site designers, although there are general introductions available on the Web (e.g. Ridings, 2001; see also Whalen, 2001). It is, nevertheless, recognised that increasing the number of backlinks to a page may well increase its rank in many search engines, as the following quote illustrates (Sullivan, 2001):

> By analysing how pages link to each other, a search engine can both determine what a page is about and whether that page is deemed to be "important" and thus deserving of a ranking boost.

A common sense approach to gateway sites' impact on page ranking might be to believe that since they increase the link count for target pages they can only increase their ranking, but this bears further investigation.

This paper presents an explanation of the core of the PageRank algorithm and investigates the impact of gateway sites upon it under certain stated assumptions about user behaviour. The principal question addressed is the conditions under which a gateway site will increase the likelihood that a target page is found in search engines. Implications for subject gateway site designers will then be discussed.

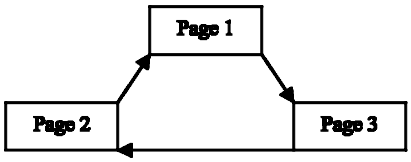
The PageRank algorithm

The core of Google's PageRank algorithm has been published by its designers and founders (Brin and Page, 1998), as well as the overall design of the search engine. It is believed that the same algorithm is in use today (Google, 2002) although it may well have secret adaptations, for example to deal with spam sites. The underlying idea of the PageRank algorithm is that a count of backlinks is a useful indication of the overall value of information in a target page. For example, a search engine that ignored links in its ranking process may use a formula based upon discovering the frequency of the keywords of a user-entered search in each potential matching document. For example, a search for "biochemistry" would be likely to return pages with this word in the document title, main headings and body, perhaps also in its URL. It would not be possible for the program to guess which was the most authoritative biochemistry page, only which was the most topic-specific. Google, on the other hand, would guess authority based upon backlink counts, perhaps ranking highest the page that was the most frequent target of links. This would make it far more likely that a genuinely respected page would be returned rather than, say, a first year biochemistry online rooming timetable.

Although PageRank is a mathematical algorithm involving finding eigenvalues of matrices, it admits a straightforward explanation. The voting metaphor used on the Google Web site (Google, 2002) and elsewhere (Lifantsev, 2000) will be used here instead of the original Brin and Page random surfer explanation but the models are equivalent. A simple link-based ranking system would be to give each Web page a vote, allowing it to split its vote evenly (in fractions) amongst all the pages it linked to. Counting votes for pages would form a ranking system, with the pages getting most votes having high numbers of backlinking pages. One criticism of this system is that it does not go far enough (e.g. Bharat and Mihaila, 2001). Gateway pages for example, will gather many votes if they are well linked to, but will only have one vote to share between their targets, which presumably contain the valuable content. It makes sense to repeat the process again, allowing each site to pass on votes acquired in the previous round to its target sites. If this voting is repeated again and again, will this guarantee that the pages that eventually get the votes have the really important content? A major practical problem with continuing the voting indefinitely is that all sites that host links will expend any votes garnered on them and may eventually run out and be unranked. In response to this, new votes can be continually added to the system so that pages with both backlinks and (out)links will still be able to retain votes, using some averaging process to prevent the total number of votes in the system increasing indefinitely. Unfortunately, there is still a flaw in the system, that of "sinks" which are groups of pages which link to each other in a circular structure (Brin and Page, 1998). These vote for each other, keeping their votes in the system, while voting that ends at a page without links gets recycled to other pages (see Figure 1). The end result of running many rounds of voting will be that

Online Information Review
Volume 26 · Number 2 · 2002 · 101–107

Figure 1 A sink created by three pages voting for each other



only sinks get a high number of votes, clearly an undesirable result. There does not seem to be an easy way around this problem, bearing in mind that the solution must be able to be translated into a mathematical algorithm that can be applied to the billions of links in the Web and so the obvious resolutions are not necessarily practical.

The solution of Brin and Page was to recycle a percentage of the votes automatically at each stage instead of sending them on to link targets. They suggested the figure of 85 percent so that at any voting stage at each URL, 15 percent of its votes would be allocated to its link targets and 85 percent were distributed evenly to all URLs in the system. The net effect of this is to stop sink systems from rapidly accumulating votes. A mathematical algorithm can implement the new regime and run it until the votes at each URL are reasonably stable, producing the desired PageRanks.

Although Google is based on PageRank, it has been subject to unknown modifications since its inception, and so the version currently used is unlikely to be as simple as described here. The official statement is, "while we have dozens of engineers working to improve every aspect of Google on a daily basis, PageRank continues to provide the basis for all of our web search tools" (Google, 2002). Modifications to PageRank have been suggested by information retrieval researchers, for example restricting the ranking calculations to topic-specific pages (Richardson and Domingos, 2001) and the modification of competitive algorithms to have PageRank-like qualities (Ng *et al.*, 2001) but the products of these do not fundamentally alter its character in a way that will qualitatively affect the discussion below. There are two potential modifications that would make a significant difference, however. The first would be to operate on the basis of Web sites rather than pages, as suggested by Lifantsev (2000). The second is to implement an early suggestion of Brin and Page (1998) to give automatically higher votes to more useful or relevant pages. As far as is known, however, these have not been implemented. Values related to Google's actual PageRanks can be obtained by installing the Google Toolbar from its Web site. See Ridings (2001) for a discussion of the relationship between the values displayed and the underlying PageRanks.

PageRank and gateway sites

We shall consider ideal systems in this section for simplicity, and discuss later the implications of likely deviations from the assumptions. Suppose that there are n pages linking to all of m content pages, in a very much larger collection of pages that PageRank is being applied to (Figure 2).

At the start of the process, each page gets an equal vote p. In voting round 1, each page is again given a vote p but any pages that link to other pages split 15 percent of their vote evenly amongst all target pages. Pages without any link targets forfeit the ability to vote for other pages and their 15 percent is lost along with the 85 percent that all pages lose anyway. In subsequent rounds the same voting patterns recur, and so the round 1 votes are a stable state for the system (see Table I).

Figure 3 shows the same set of source and target pages after the inclusion of a gateway

Figure 2 A link structure without a gateway site

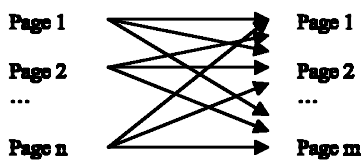
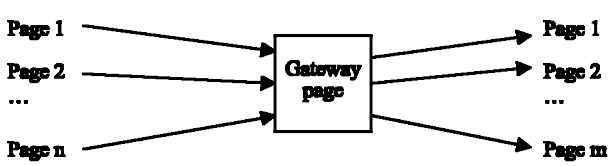


Table I PageRank calculations for system A

Pages	Start	Round 1+
Source pages 1 ... *n*	p	p
Target pages 1 ... *m*	p	$p + 0.15 \frac{n}{m} p$

Figure 3 A link structure with a gateway site *replacing* direct links



site. This operates under the assumption that all source pages switch from voting for all the target pages and vote for the gateway page instead. PageRank should compensate for the reduction in links to the target pages with an increased weight for links from the gateway page

At the start of the process, each page again gets an equal vote p. In voting round one, each page is again given a vote p and the source pages give 15 percent of their previous vote to the gateway page, which splits 15 percent of its original vote amongst all the target pages. In this round, the target pages have a low vote because the fact that the gateway page is well linked to has not fed through the system yet. In round two the same process recurs with the new votes but in this case the increased gateway site vote feeds through to the target pages. In subsequent rounds the same voting patterns recur, and so the round two votes are a stable state for the system. A related technical point is that the votes at pages that do not have links do not disappear from the system but are used to increase the value of p, but for a large system the increase will be insignificant (see Table II).

In the unorganised original system the content pages have PageRank

$$p + 0.15\frac{n}{m}p,$$

which is greater than the PageRank of the contents pages of the organised system

$$p + 0.15\frac{p + 0.15np}{m}$$

if $n > 1$, but less than the gateway Pagerank of $p + 0.15np$ if $m > 1$.

As can be seen from the calculations, the net effect of introducing a gateway site under the idealised conditions is that the gateway site will have a higher rank than any individual content pages and that the value of the additional votes of all target pages above the standard p will drop to approximately 15 percent of its original value if n is large enough to make the single vote of the gateway page insignificant. If only a proportion of pages switch from linking to all target pages to linking to a gateway site then the changes in ranking will not be as extreme as illustrated here, but will still be in the same direction. In reality, there will probably also be links to the gateway site created from pages that would not have linked to any of the target pages. These will increase the PageRank of the Gateway site and content pages but because of the 85 percent loss through intermediation, the number of these would have to be greater than $\frac{1}{0.15} = 6\frac{2}{3}$ times the number that switch to produce an overall PageRank increase in content pages.

Another likely difference from real world implementations is that each source page would probably link to only a subset of the gateway site targets, with presumably the differing overlaps giving the highest PageRank to the highest quality pages. The Gateway site, under real conditions, will tend to counteract this by allocating the same votes to each target site irrespective of quality. This aspect will not change the overall magnitude of the PageRanks for the content pages, only reduce their spread.

There is one property of the architecture of a gateway site that should be considered here: the number of levels of the site that must be traversed to get to the links to the content pages. The above calculations have been performed under the assumption that the gateway site is a single page. In reality most probably have several layers, meaning several pages that must be clicked through from the home page before external links are met. Since each additional layer will steal 85 percent of the votes, this can greatly reduce the votes given to the content pages. For example, a site with a home page, a second page giving major categories, a third set of pages giving subcategories followed by a fourth set with the external links would reduce votes transferred to approximately 0.05 percent instead of 15 percent for a single-page gateway site.

Table II PageRank calculations for system B

Pages	Start	Round 1	Round 2 +
Source pages 1 ... *n*	p	p	p
Gateway	p	$p + 0.15\ np$	$p + 0.15\ np$
Target pages 1 ... *m*	p	$p + 0.15\frac{1}{m}p$	$p + 0.15\frac{p + 0.15\ np}{m}$

Case study

The excellent Humbul Humanities Hub (http://www.humbul.ac.uk/) is one of the most visible subject based sites for the UK academic community (Thelwall, 2001) as well as having an attractive and high quality user interface, as reference to a standard text will show (Nielsen, 2001). This site is chosen for a case study because of its importance and to illustrate the issues involved in changing a site that is already well designed. Many other gateway sites will not have such a good structure already and could expect more dramatic results than those demonstrated below.

The Hub has an economical three-tier structure, with links to the 16 major categories on the home page. Each major category page contains a list of subcategories, clicking on one of which gives a page of the first 15 links in that category. Its database contains very approximately 3,000 links to other sites or pages. An advanced search with the US version of AltaVista was conducted in January 2002 in order to find out how many external pages contained a link to it. The syntax below was used:

link:humbul.ac.uk AND NOT
host:humbul.ac.uk

This returned 1,423 matches. Assuming that each of these pages links to the Humbul home page and not to any other sites, and based upon the structure of the site and average number of links per page of each type, Table III shows how the PageRank shrinks inside the site, with an enormous PageRank for the home page, but each of the estimated 3,000 target pages receiving 1.0030 *p* to 4DP (assuming no other links to them). If the external pages linked to all the 3,000 target pages directly (or shared them evenly) then this would give them a PageRank of 1.07115 *p*. Note that the lower decimal places are significant for ranking purposes since the minimum PageRank is *p*, and most are likely to be only slightly larger. The actual numbers in this example are not claimed to be precise because of the averaging and the existence of links to the home page and all major categories on each page. The navigational links were not included because they cause recursion in the calculation of PageRank, but their impact is relatively small, although it would change the decimal places shown in the table. The significance of retaining non-accurate decimal places here is in illustrating the magnitude of the differences between figures.

If the same site were to be designed with PageRank impact as the primary concern then these suggestions would be made:

- Reduce the structure to two-tier by using a Yahoo!-style listing of subcategories underneath the main categories on the home page.
- Change the redirection links to external sites to direct links, perhaps using JavaScript (see below).
- Implement the links on pages that are not to direct categories or subcategories in a way that will not result in them being identified by Google, for example using server-side image maps (Corcoran, 1996).
- Reduce the number of links per page from 15 to the smallest number consistent with usability, say ten.

A simple piece of JavaScript that can create a link that will send visitors with JavaScript

Table III PageRank calculations for the Humbul Humanities Hub

Page	Average number of external links	Approximate PageRank
A total of 1,423 external pages	1	1.00000 *p*
Humbul home page, with major categories listed	50	214.45000 *p*
Humbul major category page with subcategory list	30	1.64335 *p*
Subcategory page with first list of external links	50	1.00822 *p*
Subcategory page with subsequent list of external links	50	1.00302 *p*
Target from first page of links	–	1.00302 *p*
Target from any subsequent page of links	–	1.00301 *p*

Note: The decimal places displayed are not accurate (due to the omission of links to higher order level site pages which would cause recursion) but are included so that the trend in the data is clear

enabled in their browsers through a redirection mechanism but search engines directly to the target site is shown here. Users without JavaScript enabled browsers will still end up at the correct site, but their visit will not be logged:

```
< A HREF = "http://www.durham.ac.uk/
        corpus/" onClick = "
          document.location.href =
          "http://www.humbul.ac.uk/
          output/redirect.php?URI =
          http%3A%2F%2Fwww.
          durham.ac.uk%2Fcorpus%
          2F"; return false;
>Corpus of Anglo-Savon stone sculpture
          </a>
```

Table IV shows the same PageRank calculations if these recommendations were to be implemented.

All of the changes suggested above have drawbacks that the site designers may feel outweigh the PageRank advantages seen by comparing the two tables, and so they should be seen as points for discussion rather than definite recommendations. The first and fourth may make the site harder to use. The second will impair the service from collecting effective data on the links that their visitors are finding useful. The third requires extra server programming to implement and maintain.

Summary

The calculations have shown that the introduction of a gateway site to organise a subject area on the Web, while being a highly desirable thing in itself, can have the long term effect of reducing the link based search engine visibility of the target sites, particularly those with the highest quality. The potential decrease is subject to a trend for Web site designers to link to gateway sites instead of content sites, and is based on the calculations showing that one direct link is worth at least 6 2/3 indirect ones. A PageRank decrease will affect the ability to find information of those who use search engines rather than gateway sites, whether through choice or through lack of knowledge of the latter. It was also shown that the site architecture can impact upon the PageRank of content pages.

Given that scholars will continue to use commercial search engines in parallel with subject portals, it is contended that gateway site owners have the responsibility to take steps to ensure that their site does not harm the PageRank of its identified resources. Moreover, given that subject portals are probably unusually authoritative and effective indicators of link target quality in the context of the vast unregulated Web, it does not seem unreasonable to suggest that steps are made to increase target pages' PageRank. The following suggestions for consideration are made in line with this principle:

- The gateway site home page, as a likely recipient of a high PageRank, should be constructed so that its description, as reported in search engine results pages, is accurate and informative.
- The architecture of the site should be as shallow as is consistent with a good user interface, perhaps using the Yahoo! home page as a model for displaying categories and subcategories together. Small sites should consider operating with a single page.

Table IV PageRank calculations for a Humbul-like site created to optimise PageRanks

Page	Average number of external links	PageRank
A total of 1,423 external pages	1	1,423.00000 *p*
Home page with major categories (with server links) and 80 subcategories	80	214.45000 *p*
Subcategory page with first list of ten external links and ten links to subsequent pages of links from the same subcategory	20	1.40209 *p*
Subcategory page with subsequent list of ten external links and ten links to other links pages	20	1.01052 *p*
Target from first page of links	–	1.01052 *p*
Target from any subsequent page of links	–	1.00758 *p*

Note: The decimal places displayed are not accurate but are included so that the trend in the data is clear

- The internal and external links on the site should be implemented in a manner that is transparent to search engines, i.e. standard HTML rather than JavaScript, Java, Flash, database-driven dynamic links or server redirection-based links. JavaScript can be used to enable the server to gather data on which sites users visit, however (see above).
- Links that are not intended to confer recognition on target pages (e.g. navigational or credit links repeated on every page) should be implemented using the server-side image map mechanism (Corcoran, 1996) that should prevent search engines from recording them. This will help to reduce the PageRank loss due to large numbers of outwardly pointing links per page.
- Consider hosting fewer numbers of links per page in larger multi-page sites.
- Consider putting the most important links on the first page of links, rather than using alphabetical order, because the first page will have a higher PageRank.

As a final recommendation, individual Web page authors should continue to create links to subject gateways that organise useful resources because these provide direct pointers to new visitors and also help search engine visibility. They should also continue to link directly to pages considered to contain high quality content in order to help differentiate them in search rankings.

References

Bharat, K. and Mihaila, G.A. (2001), "When experts agree: using non-affiliated experts to rank popular topics", *Tenth International World Wide Web Conference*, available at: www.www10.org/cdrom/papers/474/index.html

Brin, S. and Page, L. (1998), "The anatomy of a large scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 30 Nos 1-7, pp. 107-17, available at: citeseer.nj.nec.com/brin98anatomy.html

Corcoran, P. (1996), "Piecing together server-side image maps", available at: hotwired.lycos.com/webmonkey/html/96/39/index2a.html

Google (2002), "Our search: why use Google", available at: www.google.com/technology/index.html

Haveliwala, T. (1999), "Efficient computation of PageRank", *Stanford University Technical Report*, available at: 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, pp. 307-15.

Huxley, L. (2001), "Renardus: fostering collaboration between academic subject gateways in Europe", *Online Information Review*, Vol. 25 No. 2, pp. 121-27.

Kleinberg, J. (1999), "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46 No. 5, pp. 604-32.

Lifantsev, M. (2000), "Voting model for ranking Web pages", in Graham, P. and Maheswaran, M. (Eds), *Proceedings of the International Conference on Internet Computing*, CSREA Press, Las Vegas, NV, pp. 143-8.

MacLeod, R., Kerr, L. and Guyon, A. (1998), "The EEVL approach to providing a subject based information gateway for engineers", *Program*, Vol. 32 No. 3, pp. 205-23.

Ng, A.Y., Zheng, A.X. and 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)*, ACM Press, New York, NY, pp. 258-66.

Nielsen, J. (2001), *Designing Web Usability: The Practice of Simplicity*, New Riders, Indianapolis, IN.

Rafiei, D. and Mendelzon, A.O. (2000), "What is this page known for? Computing Web page reputations", *Computer Networks*, Vol. 33 Nos 1-6, pp. 823-35.

Richardson, M. and 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: www.cs.washington.edu/homes/mattr/doc/NIPS2001/qd-pagerank.pdf

Ridings, C. (2001), "PageRank explained", available at: www.goodlookingcooking.co.uk/PageRank.pdf

Saito, E. and Onodera, N. (2001), "The use of metadata for science resources on the Web", *Journal of Information Processing and Management*, Vol. 44 No. 3, pp. 174-83.

Sullivan, D. (2001), "How search engines rank Web pages", available at: www.searchenginewatch.com/webmasters/rank.html

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

Whalen, J. (2001), "PageRank summary", *Successful Online Copywriting & Search Engine Optimisation*, Vol. 70, available at: www.rankwrite.com/archives/issue070.htm#seo