In praise of Google: finding law journal Web sites[1]

Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton,
35/49 Lichfield Street, Wolverhampton, WV1 1EQ, UK.
Email: m.thelwall@wlv.ac.uk

Abstract
Google is a highly regarded and widely used search engine, particularly in academia. In this short note we comment on its remarkable ability to find journal web sites, using a case study of law.
Keywords: Google, search engines, PageRank.

Introduction

Whilst early search engines attempted to identify pages satisfying an information request by matching the query text with pages in their database, Google introduced an algorithm that used the link structure of the web to predict the best quality matching pages. The seminal article of Google's founders describing its ranking mechanism (Brin & Page, 1998) did not contain convincing experimental evidence for its worth, but its widespread use testifies to its ability to satisfy user needs. In fact the efficacy of Google's PageRank algorithm seems to be taken as a given by related research that has borrowed aspects of its functionality (Ng *et al*., 2001; Thelwall, 2002), despite poor results under experimental conditions in one study, (Hawking *et al*., 2000) although information retrieval requirements in this case were probably very atypical for the Web. This study investigates how valuable Google is for a specific type of information request: finding the Web site of a journal. In particular, law journals will be considered. The purpose is not to compare Google with other search engines, but only to comment on its ability to return the desired pages from amongst the (frequently) many that match the search text but not the search intention.

Methods

A list of law journals was extracted from the ISI Web of Science. In order to consider journals that exhibited a reasonable life span, the list was restricted to just those journals that were in both the 1997 list and the 2000 list.

Each journal was then searched for in Google using its full name encapsulated in quotes. The position in the list of the first page from the journal's site(s), normally the home page, was then recorded. In many cases, Google returns two pages from a site: the site home page and one other page within the site, indented in the results page. In the classification scheme used, if the second of a pair of this type was from the journal but the first was from the hosting organisation, it was counted as the first one matching since the difference is really a results presentation issue. In cases where no match was found on the first fifty pages, a site for the journal was sought by other means, including AltaVista and lists of law journal home pages on the Web. If no URL could be found then the journal was classified as not being on the Web. Password protected online databases such as that of Westlaw were not counted. The queries were conducted during November, 2001.

---

[1] Thelwall, M. (2002). In praise of Google: finding law journal Web sites. *Online Information Review*, 26(4), 271-272.

Results and Discussion

The results shown in Table 1 are a remarkable tribute to Google's information retrieval ability. On the Web there are normally many pages containing the query string, but Google was very successful in obtaining the most important one for the query. For example for the journal title, "Law Library Journal", its home page was number one in a list of "about 2,460" results. The remainder of pages seemed to be predominantly including the journal's name in either lists of journals or as part of a citation.

Table 1. The ranking of law journal home pages in a Google search for the quoted full journal name.

Rank	1	2	3	4	5	6	14	Not found
Number	77	2	4	1	3	1	1	2

The two journals not found were both hosted by Kluwer Academic Publishers. It is believed that this site used to ban search engines (WayBack Machine, 2002), which would exonerate Google from any 'blame'. When re-checked two months later, the site was crawlable, the pages were indexed and both appeared at number one for their respective searches.

Of the journals indexed by Google but not highly ranked, one reason was that the journal's name was a commonly occurring phrase and that the lower ranking was due to competition with pages not associated with it. For example, the lowest ranked journal found was *Family Law Quarterly* and the number one match for this journal was *Child and Family Law Quarterly*. Other journals also were also topped by the home pages of other journals with a longer version of their name. Another general reason may be page or site design that is not optimised for search engines. The home page of *Family Law Quarterly* (http://www.abanet.org/family/familylaw/) does badly in this respect because although the page's title contains the journal name, in the body of the page it is not repeated, with a non-indexable graphic used instead, although the caption of the graphic was the journal name. It may be that as a consequence of this, Google's algorithm did not judge the page to be highly relevant to the specific query. In fact, Google's highest match from the site was a page from an issue in 1996 (http://www.abanet.org/family/familylaw/current.html) that did contain its title in both the HTML title tag and in the text of the page. It is presumed here that Google does have reasonably developed additional query matching techniques in addition to its PageRank algorithm.

In conclusion, Google can be highly successful at finding the most important Web sites for matching text but may be less successful when the query text overlaps with significant other types of information, or the site design obscures its content from the text based crawlers.

References

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. 107-117 Available at http://citeseer.nj.nec.com/brin98anatomy.html

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.

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),* New York: ACM Press, pp. 258-266.

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

WayBack Machine (2002) Robots.txt Query Exclusion. Available at: http://web.archive.org/web/*/http://www.wkap.nl/journalhome.htm/1011-6702