Web Impact Factors and Search Engine Coverage[1]

Mike Thelwall
School of Computing and Information Technology
University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1SB, UK.

Email: cm1993@wlv.ac.uk

Abstract

There is an increasing amount of academic and other information on the web [1]. There is also an increasing number of online journals as well as online versions and indices of traditional journals. It seems natural therefore to expand the notion of Impact Factors for journals to web equivalents, and to use the power of search engines to carefully extend them to cover other domains on the Internet. However, search engines only index a proportion of the web, and this proportion is not determined randomly but by following algorithms that take into account the very properties that Impact Factors measure. A survey was conducted in order test the coverage of search engines and to decide whether their partial coverage is indeed an obstacle to using them to calculate Web Impact Factors. The results indicate that search engine coverage, even of large national domains is extremely uneven and would be likely to lead to misleading calculations.

INTRODUCTION

Web Impact Factors (Web-IF) are web versions of the Impact Factors (IF) published by the Institute of Scientific Information for scientific journals. A Web-IF is essentially the number of pages linking to a site or area of the Internet, divided by the number of pages in that site or area [2]. An area here can be any agreed collection of sites, such as all domain names ending in .no or all articles of an online journal inside a larger site. A high value is presumed to indicate a site with a greater impact because there are relatively many pages linking to it. This calculation includes pages that link to other pages in the same site, which is problematical because it will benefit sites using linked navigation aids on many pages, but is nevertheless an interesting concept and potentially of value to academic, government and business users of the Internet. The leap from using academic references to web links from wide ungoverned areas of the Internet is a large one: low information resources such as entertainment pages may have many links. Moreover, a link on one site may come from an informed individual's carefully considered quality judgement, but would count at the same value as any other link, including those produced by computer programs, adverts and "links exchanges", where a group of sites agree to link to every others site in the group. Nevertheless, it is possible that inconsistencies such as these average out over large areas, such as countries, or can be avoided by restricting consideration to tightly controlled domains such as journal sites. Before such projects can be considered as providing information that can be evaluated to decide upon the meaning of the results, it is necessary to ascertain the reliability of the measure.

Web-IFs have been calculated by Ingwersen [2] for web hosts and domains using advanced searches with the search engine AltaVista, following on from the

[1] Journal of Documentation, vol. 56, no. 2, March 2000, pp. 185-189

previous study of Almind and Ingwersen [3]. This method has been shown by Ingwersen to have some internal consistency at least for large segments of the web, but not for smaller institutional sites. The use of search engine technology is necessary to make this kind of calculation possible for large domains, but it has a number of drawbacks. There are inherent technological problems stemming from the fact that the search engine will not be able to recognise all links, for example links generated by a database query on the web or by some JavaScript or Java programs. A problem with a much bigger potential impact on the calculations is the actual web site coverage by the search engine used. Search engines only cover a proportion of publicly indexable web. Lawrence and Giles estimated in February 1999 that this proportion was not more than 16% for any search engine [4]. Typically two ways are used to find pages: user registrations of new sites and links from previously indexed pages. The use of links for this purpose suggests that the pages left out may well have less links to them and therefore Web-IFs created by search engines are likely to be unrealistically high. This may not be a problem for comparison purposes if the Web-IFs of areas of the web to be compared with each other have a similar coverage by search engines, but if the coverage is different then the calculations are likely to be meaningless. The actual algorithms used by search engines are not normally made fully public and are a source of public and research interest [5,6,7]. In order to compare search engine coverage of different domains, a survey was conducted to test whether the percentage of registered sites was common across different areas of the Internet or whether there were significant variations.

THE DOMAIN COVERAGE SURVEY

Our survey was intended to find sites not registered with search engines equally with those registered and so we employed a combinatorial method for finding addresses of web sites. We chose to search for sites with the main part of their domain names containing up to three letters, although up to 22 letters can normally be used. For example in the .no domain we tested from www.a.no to www.zzz.no. This gives 18278 potential domains in each case, which were tested by a specially constructed computer program and the results stored in a database. Most domains returned an error message indicating that they were unused and were subsequently ignored, the remainder progressing to the next stage of the analysis. This method was a practical way to find large numbers of sites in the correct domain. In domains where there was a significant split into categories, such as commercial, educational and government we chose the commercial domain, normally the largest, for illustrative purposes. It is clearly not a random sample and contains the potential to be biased, being restricted to short domain names. One likely impact is that in congested domains the older sites are over-represented, having adopted the more popular shorter domain names first. Different domains may also be affected in different ways due to cultural and linguistic differences of different nations. Sites with non-standard domain name structures and groups of sites hosted by a general domain name would also be left out. It is however a practical method of finding a large sample from selected domains irrespective of search engine registration. Each site found was checked manually for being genuine, many being excluded as holding pages, under construction sites or various other kinds of sites not intended for public view. Each genuine domain name found was tested, again with an automated process, to see whether AltaVista registered any pages on the site. AltaVista was chosen for comparison with the previous survey and is an appropriate choice because of its advanced search facilities, including the ability to search by domain name, and the relatively large size of its index. A negative result

would mean that no pages were included, whereas a positive one would indicate that at least one was, but not necessarily all were. The survey was conducted from 19 to 22 July 1999.

RESULTS

Selected results of the survey are summarised in table 1.

Table 1. *Selected percentages of tested sites registered with AltaVista*

Country	Special domain	Domain name ending	Number of sites found	Percentage registered
Finland		fi	562	82%
US/World	com	com	8050	81%
Sweden		se	1827	75%
Japan	co	co.jp	2497	74%
France		fr	1483	68%
UK	co	co.uk	3643	57%
Norway		no	1192	57%
Denmark		dk	2762	56%
Armenia		am	42	33%

These results show large differences in percentage registration with the search engine used for the Web-IF calculations, AltaVista. We found a similar spread of differences with four other search engines that we tested, Infoseek, MSN and Hotbot: chosen for their (in some cases implicit) support for searching for the existence of domain names in their databases. This provides some evidence that the calculation of Web-IFs from search engines is problematical. To illustrate the case with one example, Finland produced a low score in Ingwersen's calculations, with an average of 0.81 but Norway scored much higher with an average of 1.12. Our figures suggest that these numbers should be much closer because of the much larger percentage of Norwegian sites missing. The missing sites would be likely to be badly linked to, since links are used by search engines to find new sites, and therefore would increase the denominator of the calculation without commensurate increase in the numerator. In this case it is unlikely that the Finnish score would overtake the Norwegian, and with any regular assumptions extrapolation would still leave a clear gap. However, for closer scores, as the majority were, search engine coverage could easily make a difference.

Academic domains tend to be smaller, older and better linked than commercial domains. The case for the ac.uk domain in comparison to the co.uk domain illustrates this point. The UK academic community has various 'gateway' sites that attempt to index all academic servers and as a result nearly all are registered with many of the search engines. The addresses of all universities for example are known to AltaVista, and if the rest of the ac.uk domain is included then 200 of the 205 members are registered, 97.5%, much higher than the 57% for our sample of the co.uk domain. Nevertheless the coverage inside each academic site is far from complete. For example for the University of Wolverhampton two main web servers AltaVista reports a total of 7764 pages but there are in fact many times this number. Aberdeen's Central web server index reports over 7216 pages but AltaVista reports 'about 2558', for Liverpool University the internal search engine reports 22,250 pages but AltaVista has indexed 'about 15083' and for the University of East Anglia its search engine indicates at least 9,508 but AltaVista reports 'about 1769'. We deduce that search

engine coverage is uneven even in the domains registered and therefore that Web-IFs are likely to be unreliable even for the relatively well linked academic domains.

CONCLUSION

For the domains that are surveyed here the exciting concept of the Web-IF appears to be a relatively crude instrument in practice due to the limited coverage of the web by search engines. A Web-IF for a domain calculated from AltaVista would be expected to be relatively accurate only if the link pages counted were between similarly well registered and covered domains. Before using a search engine for calculations on areas of the Internet checks should be made to ensure that its coverage of the actual pages is high. We believe that Web-IFs may well be unreliable even for the best case of compact groups of sites with at least one gateway site, such as academic sites, because of the limited coverage inside each site. We believe that this failure is due to the nature of search engines as they currently operate, including the inevitable time delays before indexing new pages. This builds upon the work of Snyder and Rosenbaum [9], and of Smith [10] who found major inconsistencies between search engines and even with the same search engine over short periods of time. The present survey, indicating that search engine coverage is very uneven, provides an additional cause for concern.

A fundamental problem seems to be the sheer size of the web and the inability and perhaps understandable lack of desire of the search engines to cope with it. The Holy Grail of indexing the entire web can be seen as a secondary goal to the task of getting the most relevant information to a general web user in response to their search query. For this purpose search engines use techniques to rate the relevancy of pages which downplay the importance of badly linked pages, under the assumption that better linked pages are more popular and more likely to be of interest. A by-product of this may even be a process of discarding badly linked pages from the database. It seems therefore that reliable Web-IFs from a commercial search engine are not possible at the moment. They would only be a possibility in the future if search engines manage to get ahead and stay ahead of the ever-expanding size of the web. A realistic alternative would be to use one of the numerous academic specialist search engines [8] to systematically index a limited subset of the web of recognised and verifiable value, such as academic journal sites, and to use this to calculate Web-IFs. If this can be achieved, perhaps with participating site agreeing to adhere to a common agreed format, such as one page per article and a standard method of referencing other articles, then it may be possible to achieve reliability and move on to a discussion of the meaning of the figures calculated.

REFERENCES

1. Chowdhury, G. G. The Internet and Information Retrieval Research: A Brief Review. *Journal of Documentation*, 55(2), 1999, 209-225.
2. Ingwersen, P. Web Impact Factors. *Journal of Documentation*, 54(2), 1998, 236-243.
3. Almind, T. C. and Ingwersen, P. Informetric analysis on the World Wide Web: methodological approaches to webometrics. *Journal of Documentation*, 53(4), 1997, 404-426.
4. Lawrence, S. and Giles, C. L. Accessibility of information on the web. *Nature*, 400, 1999, 107-109.
5. Pringle, G., Allison, L. and Dowe, D. L., What is a tall poppy among web pages? *Computer Networks and ISDN Systems*, 30(1-7), 1998, 369-377.

6. Schwartz, C. Web Search Engines. *Journal of the American Society for Information Science*, 49(11), 1998, 973-982.
7. Tunender, H. and Ervn, J. How to Succeed in Promoting Your Web Site: The Impact of Search Engine Registration on Retrieval of a World Wide Web Site. *Information Technology and Libraries*, 17(3), 1998, 173-179.
8. Brin, S. and Page, L. The Anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 1998, 107-117.
9. Snyder, H. and Rosenbaum, H. Can search engines be used for web-link analysis? A critical review, *Journal of Documentation*, 55(4), 1999, 375-384.
10. Smith, A. G. A tale of two web spaces: comparing sites using Web Impact Factors, *Journal of Documentation*, to appear.