Journal of Documentation, Vol. 54, No. 2, March 1998

RESEARCH BRIEF

THE CALCULATION OF WEB IMPACT FACTORS

PETER INGWERSEN
pi@db.dk

Centre for Informetric Studies, Royal School of Library and Information Science, Birketinget 6, DK 2300 Copenhagen S, Denmark

This case study reports the investigations into the feasibility and reliability of calculating impact factors for web sites, called Web Impact Factors (Web-IF). The study analyses a selection of seven small and medium scale national and four large web domains as well as six institutional web sites over a series of snapshots taken of the web during a month. The data isolation and calculation methods are described and the tests discussed. The results thus far demonstrate that Web-IFs are calculable with high confidence for national and sector domains whilst institutional Web-IFs should be approached with caution. The data isolation method makes use of sets of inverted but logically identical Boolean set operations and their mean values in order to generate the impact factors associated with internal- (self-) link web pages and external-link web pages. Their logical sum is assumed to constitute the workable frequency of web pages linking up to the web location in question. The logical operations are necessary to overcome the variations in retrieval outcome produced by the AltaVista search engine.

INTRODUCTION

IMPACT FACTORS (IF) for scientific journals are published by ISI (the Institute of Scientific Information) in the annual *Journal Citation Reports* and commonly used for evaluation purposes, although their methods of calculation and use are disputed [1]. Generalised IFs for journals covered by ISI services can be reproduced accurately online, including external- and self-citations, by applying the publicly available citation indexes [2]. We have tested whether similar online estimations of selected national, sector, and institutional impact factors for the World Wide Web (Web-IFs) are feasible and reliable. The dynamic real-time nature of the WWW suggests Web-IFs as a useful supplement to the traditional impact factors when monitoring the status of web locations. They can be seen as evidence or indicators of the relative attractiveness of countries or research sites on the WWW at a given point in time. The study demonstrates that Web-IFs are calculable and reliable with caution if processed by means of a series of inverted but identical Boolean operations in order to isolate and estimate the number of web pages pointing to the pages at a given site.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

DEFINITIONS

The generalised IF of a scientific unit is in principle defined as the number of citations given during a period T_1 to that unit's items published in a period T_2, divided by the number of citable items published by that unit in the period T_2. Citations mean the sum of self-citations and citations given by external sources. Depending on the purpose of the analysis the document types included in the denominator may vary as may the citation and publication windows, T_1 and T_2[2]. The frequently used IF for scientific journals as produced by ISI has T_1 as year Y and T_2 is (Y–1) plus (Y–2), and applies solely articles, reviews and notes in the denominator [1].

In the WWW context we have replaced the external citations by web pages external to a given site which point at least once to that site; self citations are replaced by the number of web pages internal to that site which point at least once to the same site, hereafter called 'self-link web pages'. Our definition for the Web-IF takes the logical sum of the number of external- and self-link web pages pointing to a given country or web site divided by the number of pages found in that country or web site – at a given point in time. The numerator thus consists of the number of link pages – not the number of links. The intensity of the links is hence not calculated in contrast to the traditional journal IF for which not only the number of different articles citing a journal is calculated, but *also* the frequency of the citations given to that journal and placed in those articles. This is the reason for the fact that the journal or individual IF can be artificially raised by self-citations. For the Web-IF our definition implies that, in order to count, a new self-link must be placed on a web page not already holding such a link; and then both the numerator *and* the denominator increase with the value of one. In principle, a Web-IF can only increase above 1.0 with the growth of the number of external-link pages which point, at least once, to a particular web site. Consequently, Web-IFs can only compare directly with traditional IFs for which we know the number of different sources citing a given object, for instance, that Z articles published by journal X during time T_2 were cited by Y different articles at least once during time T_1.

The external-link pages can be seen to mirror social communication phenomena, such as strategic or tactical referral behaviour, and pragmatic or common semantic interest in particular sites on the web. An external Web-IF becomes a measure of the extension of the attractiveness of a given site. In addition, self-linkage also reflects the logical structures used for organising web pages in the local servers. Unlike scientific citations to journals, institutions or individuals, which may be stable or may constantly increase, the number of pages linking up to a particular web object may indeed decrease or disappear over time, for example, due to closedowns or restructuring of web sites. Thus, in contrast to the common IF calculation a retrospective Web-IF is not reproducible.

METHODOLOGY

We performed a series of online snapshots of the WWW over a month, each with a duration of less than 1.5 hours and with observation windows of less than 120 seconds per web site. The search engine's sampling method during data retrieval

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

as well as the web conditions could then be kept constant. The case study applied the advanced search mode of the operational facilities available on the web retrieval engine AltaVista [3]. This retrieval mode allows for the use of elaborate Boolean search strings and provides the exact number of pages found. AltaVista's retrieval method is in principle large scale sampling with an unknown run-time limitation. Also unknown is its mode of back tracking in time and the exhaustivity of its search for links on individual web pages. In principle almost all searches will thus be incomplete. We assume that these retrieval conditions are constant for all search sessions made in the case study. In order to avoid additional variables and incompatibility problems other web search engines than AltaVista have thus far not been used.

Two logical conditions commonly underlie the methods of isolating and calculating impact factors – including any Web-IF. First, the sum of self- and external-link pages equals the number of link pages, i.e. 'self-LP ∪ ext.-LP = link-P'. Secondly, we assume the result of the Boolean logical set operation 'a ∩ b' (named Logic A) to be equal to the result of its inverted form: 'b ∩ a' (named Inverted Logic A). In principle the latter condition implies that the order of search elements in a Boolean AND operation should *not* influence the outcome. However, we should note that the outcome from AltaVista applying the two inverted but logically identical set operations actually differs slightly. To test the reliability of a Web-IF for a given web location when in a real-time retrieval environment both conditions should thus be met with a high degree of certainty. Based on the command syntax of AltaVista the following six retrieval arguments have been applied to national and sector web pages (left column; Denmark as example) with the corresponding six arguments for institutional web sites in the right hand side column (Royal School of LIS, Denmark as example):

(1) domain:dk	(1) host:www.db.dk/
(2) link:.dk/	(2) link:www.db.dk/
(3) domain:dk AND link:.dk/	(3) host:www.db.dk/ AND link:www.db.dk/
(4) link:.dk/ AND domain:dk	(4) link:www.db.dk/ AND host:www.db.dk/
(5) link:.dk/ AND NOT (domain:dk AND link:.dk/)	(5) link:www.db.dk/ AND NOT (host:www.db.dk/ AND link:www.db.dk/)
(6) link:.dk/ AND NOT (link:.dk/ AND domain:dk)	(6) link:www.db.dk/ AND NOT (link:www.db.dk/ AND host:www.db.dk/)

Argument (1) retrieves web pages holding the national domain code 'dk' as the last part of the first segment of its address (the URL), or it retrieves the local host address 'www.db.dk/'. The slash (/) assures the retrieval of pages organised under that URL. (2) retrieves link pages which, in their text body, hold at least one link address containing the element '.dk/' as the *last part* of the *first segment* of the URL the pages are linked to – as in 'www.db.dk/'. Arguments (3) and (4) retrieve self-link pages by means of Logic A (3) and Inverted Logic A (4), that is, arguments which are supposed to yield identical results. By similarly identical but inverted logical operations the arguments (5) and (6) retrieve the number of external-link pages. In a previous webometric study [4] Almind and Ingwersen were forced to apply only the 'host:' command in order to isolate and count national web pages since the more focused 'domain:' command had not yet been launched by AltaVista.

Journal of Documentation, Vol. 54, No. 2, March 1998

In order to test the first condition outlined above the retrieval result of argument (2) – called the number of 'simple link pages' – is compared for several countries and sites on the web with the results obtained from the retrieval arguments (3) and (5), added together for each site. Except for cases with only one host page as a result of argument (1), all comparisons *never* yield identical results. A typical case is, for instance, the United Kingdom. The number of national web pages as of August 20, 11 p.m. is 1,046,961 and the number of 'simple link pages' 1,067,997. A typical top-down web impact factor calculation gives 1.020 as, what we call, a 'simple WIF for the United Kingdom'. However, by means of arguments (3) and (5) (Logic A) the sum of self- and external-link pages is 1,041,941; and by means of the logically identical but inverted arguments (4) and (6) (Inverted Logic A) the sum constitutes 1,040,134 link pages. Obviously, the *differences* between the number of 'simple link pages' and the number for each of the sums of external- and self-link pages is far greater than the difference between the sums defined by Logic A and Inverted Logic A respectively. The former differences are –2.44% and –2.61% whilst the latter is -0.17% for the United Kingdom. As shown in Tables 1–2 below this pattern is quite common, also over several snapshots. Within each of the two distinct categories of link pages the study also demonstrates typical variances which seem dependent of the kind of logical operation used for retrieval. They suggest that the *order* of the statements in the Boolean argument determines for the AltaVista search engine's functional operation and the retrieval outcome.

Methodologically speaking, both the conditions outlined above can be met if one avoids the top-down approach and starts bottom-up by making use of the Logic A and Inverted Logic A operations sequentially. Since the two operations are logically identical, but do yield slightly different results, the arithmetic mean values for the external- and self-link pages respectively can be taken as the values best fit for further Web-IF calculations. Following the logic of the first condition above, the sum of the two mean values corresponds to the total number of pages linking to a particular site. This sum is consequently used as the numerator for the Web-IF calculation. The denominator is the number of web pages isolated by retrieval argument (1). For a given site at a given point in time the sequence of retrieval arguments is thus first to apply argument (1); then to enter arguments (3) and (4); and finally to activate (5) and (6). The arithmetic mean of the results obtained by (3) and (4) signifies the number of self-link pages; the average value of (5) and (6) denotes the number of external-link pages. Finally, one may retrieve the 'simple link pages' by argument (2) in order to observe the difference between the calculated Web-IF and, what we call, the 'simple WIF'. In the UK case, Table 1, the Web-IF then becomes 0.994, whilst the 'simple WIF' was 1.020; deviation: –2.52 %.

RESULTS

Table 1 shows the resulting Web-IFs in descending order for a selection of smaller and middle-size countries and large, mainly US web sectors. Table 2 displays Web Impact Factors for selected research locations and well-known scientific journals. In total, five different snapshots were applied during the month

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

August 20 to September 21, 1997, three snapshots in Table 1 and four in Table 2.

For each location in Table 1, the current web impact factor, Web-IF self-link and the Web-IF external link are displayed. The isolated number of web pages per site is shown in the last column. The deviation values are mostly negative, implying that the Web-IF has a lower value, or commonly is more conservative than the 'simple WIF'.

Each country as well as the large segments of the Net show a Web-IF with an acceptable deviation below +/–1.7% between the two intermediate arithmetic mean values isolated by the Logic A and Inverted Logic A operations. Although statistically one should not sum up the results of different samples from different snapshots we have done so in order to demonstrate a kind of 'World Web-IF' – with a relative mean of 0.899 and a deviation of –0.29%. The difference between the Web-IF and the 'simple WIF' for countries shows a deviation ranging from –7.52 to +1.88%.

Not surprisingly, Norway performs best of the European countries in this

Table 1. *Selected national impact factors for the WWW: Web-IF – Aug. 20–Sep. 21, 1997*

Countries in rank order	*Web impact factor*	*Web-IF Self-link*	*Web-IF Ext.-link*	*Deviation % logic A/ inv. logic A*	*Deviation % Simple WIF/Web-IF*	*Number of web pages*
1. Norway	1.113	0.49	0.62	0.01	–1.2	218,141
Norway	1.127	0.50	0.63	–0.53	–2.37	212,011
2. United Kingdom	0.994	0.46	0.53	–0.17	–2.52	1,046,961
3. France	0.886	0.42	0.46	–0.18	–3.47	454,822
4. Denmark	0.886	0.52	0.37	–0.62	–0.61	144,433
Denmark	0.889	0.52	0.36	0.64	0.06	153,267
5. Sweden	0.866	0.51	0.36	0.66	–0.4	489,905
6. Finland	0.823	0.43	0.39	–0.57	–6.97	317,829
Finland	0.791	0.42	0.37	–1.69	7.52	313,085
7. Japan	0.404	0.31	0.09	1.11	1.44	1,826,051
1. Government (.gov)	1.472	0.42	1.05	–1.22	–4.61	646,585
2. Organisations (.org)	1.186	0.40	0.78	–0.59	–4.57	1,677,934
3. Business (.com)	0.942	0.59	0.35	–0.18	1.88	12,084,719
4. Academic (.edu)	0.807	0.47	0.33	–0.57	–2.2	5,390,097
Total: countr. +sectors	0.899	0.51	0.39	–0.29	–0.68	22,497,477

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

selection. From the analysis [4] we are aware of the Norwegian efforts put into marketing via the WWW, an effort which seems to pay off in impact. If the figures from the mainly US web sectors are calculated together and taken as the current estimate the relative US Web-IF is 0.943, with the number of web pages being 17,999,611 and 16,981,914 link pages. The reason for this rather low Web-IF is that the business and academic educational sectors are very large with quite low Web-IFs. However, the US Web-IF is higher than the expected value, i.e. the 'World Web-IF' of 0.899. We are aware that some servers registered in generic domains, such as .edu, are not located in the US [5]. A possible bias can be tested following the method proposed above. Further, locations may actually block the access of web crawlers to their servers. So this activity will slightly raise national Web-IFs, but decrease institutional IFs. In the national case the denominator will suffer from blocked out web pages; in the institutional case the numerator will decline in value by the omitted external-link pages.

One may note that like for the current national citation impact the Japanese Web-IF is far below the expected mean value for countries and sectors [6, 7]. This situation leads to considerations about the influence of language as well as of national cultural and social factors on the meaning and interpretation of impact factors in general – also because the Japanese Web-IF for self-linking is the lowest observed in the data set.

The variation of the Web-IF over different snapshots taken within short intervals does exist, see e.g. Finland, but can evidently be much more significant for smaller web sites, see for instance www.sciencemag.org in Table 2. National and sector Web-IFs demonstrate far more robustness in this respect, possibly due to the fact that the quite restrictive 'domain:' command cannot be applied to the local servers illustrated on Table 2. The 'host:' command used here is far more unconditional in its functionality.

We may observe that the Web-IF self-link values centre around 0.5 implying that in general half the national web pages contain self-references. In order to exceed a Web-IF value of 1.0 the results so far indicate that the external link web-IF should take a value of at least 0.6.

Table 2 demonstrates a greater – yet acceptable – variation (+/–3%) between the Logic A and Inverted Logic A results for smaller web locations. The largest in the table is the academic sector, UK (.ac.uk/) with 481,881 web pages. The only web site with a consistent number of link pages was *Nature*'s local server at the time of observation. Since only one web page was detected there cannot exist any self-link pages. The online analysis of the Royal School of LIS (www.db.dk/) reveals that the isolation method is workable for that size of web locations since the approximately correct number of known web pages was retrieved.

Most importantly, the data set in Table 2 demonstrates fluctuations and quite substantial deviations from the direct isolation of 'simple link pages' to the logical sum of mean values for self-link and external-link pages, i.e. the 'simple WIF'/Web-IF deviation: from –14.1%, through zero, to +10.5%.

Such unsatisfying variations suggest avoiding the use of the 'simple WIF' as an impact factor measure, both on national and institutional levels.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

Table 2. *Selected institutional Web-IFs – Aug. 20–Sep. 21, 1997*

Individual websites	*Web impact factor*	*Web-IF Self-link*	*Web-IF Ext.-link*	*Deviation % logic A/ inv. logic A*	*Deviation % Simple WIF/Web-IF*	*Number of web pages*
www.åbo.fi/	1.97	0.42	1.55	2.37	0.24	1255
www.db.dk/	2.031	0.68	1.35	1.43	0.82	484
www.dcs.gla.ac.uk/	9.1	0.38	8.72	2.41	10.47	346
Academic (.ac.uk/)	1.194	0.43	0.76	–1.43	–13.56	429,314
Academic (.ac.uk/)	1.068	0.39	0.68	–2.82	–14.14	481,889
meetings.nature.com/	51	0	51	0	0	1
meetings.nature.com/	51	0	51	0	0	1
www.sciencemag.org/	23.762	0.06	23.7	0.835	2.23	86
www.sciencemag.org/	28.846	0.03	28.82	–0.11	3.08	65

CONCLUSION

As observed previously, the duration of the observation windows and snapshots, the logical retrieval operations and the form of the search arguments are crucial when generating the data, principally due to temporary closedowns, reorganisation, size and structure of web servers, and the search engines' sampling methods in real-time. One may also point to the fact that page revision dates can be added to the retrieval arguments proposed above. Various publication or linkage windows, e.g. the last two years' linkage to a particular web segment, may consequently be put to analysis as done recently by Rousseau [8].

One may detect at least three spin-off effects of this and similar webometric studies. Firstly, they may in turn provide novel insights into the retrieval process on the WWW. For instance, clusters of web sites can be detected by means of link page co-occurence. Secondly, the proposed analysis method can be regarded as a tool for measuring the accuracy of web search engine performance and web site organisation, linking, and structuring of pages. Thirdly, web impact factor studies may open up a Pandora's box concerning the validity of the matter, in particular because most impact factor analyses are contested. More detailed qualitative investigations of the nature of intra-web linkage may uncover the significance and properties of Web-IFs.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

In conclusion we observe that the proportional distribution of web pages between the Nordic countries presented in this study conforms to the results obtained in earlier webometric analyses [4]. We are confident that analyses of national, sector and larger web segments' or sites' Web-IFs are reliable. For smaller institutional sites of the WWW the web impact factors are less dependable – however within reach following the proposed method of calculation. In the same way as traditional IFs, comparisons should be performed with caution, and preferably be carried out within the same snapshot.

REFERENCES

1. Moed, H.F. and van Leeuwen, Th.N. Impact factors can mislead. *Nature,* 381, 1996, 186.
2. Hjortgaard Christensen, F. and Ingwersen, P. Online citation analysis: a methodological approach. *Scientometrics*, 37, 1996, 39-62.
3. AltaVista Help page. http://www.altavista.digital.com/ (visited 20–8 1997).
4. 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.
5. Press, L. Tracking the global diffusion on the Internet. *Communications of the ACM*, 40(11), 1997, 11-17.
6. Swinbanks, D. and Nathan, R. Western research assessment meets Asian cultures. *Nature,* 389, 1997, 113-117.
7. Bonitz, M. Bruckner, E. and Scharnhorst, A. Why and how could we measure the Matthew Effect for countries? In: Ingwersen, P. and Pors, N.O., eds. *Information science: integration in perspective.* Copenhagen: Royal School of Librarianship, 1996, 185-199.
8. Rousseau, R. Citations: an exploratory study. *Cybermetrics*, 1(1), 1997. http://www.cindoc.csic.es/cybermetrics/vol1iss1.html (visited 8 Dec 1997).

(*Revised version received 7 November 1997*)

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib