Classifying links for substantive Web Impact Factors

Alastair G Smith
School of Information Management
Victoria University of Wellington
New Zealand
Alastair.Smith@vuw.ac.nz

Abstract

This exploratory study develops a classification of links, based on the nature of source and target pages, and the reason for linking. This classification is applied to a sample of research oriented websites: those of universities, professional institutes, research institutes, electronic journals, and individual researchers. Some tentative conclusions are drawn about the nature of the links in the study. A definition of substantiveness of links is proposed, based on the classification. This indicated that overall, 20% of the links in the study could be regarded as substantive research links.

1 Introduction

This study examines the nature of links between web sites, and proposes a classification of links which classifies the source page, the target page, and the reason for the link. The study also examines the implications for calculating Web Impact Factors (WIFs). The WIF has been proposed as a measure of the influence of a website or domain (Ingwersen, 1998). The WIF is analogous to the Journal Impact Factor, and is the ratio of the number of links made to the site to the number of pages at the site. However, web links are not equivalent to citations: they are made for different reasons. The “raw” count of links to a site may not be a useful indication of the influence of the site. This study examines the nature of links, and the concept of a substantive WIF, in which only substantive links, that indicate the significance of the site, are taken into account.

Citations in conventional print publications are generally between research publications; while web links may be between a wide variety of publication types: personal home pages, subject resource guides, etc. Studies of conventional citations have tended to examine motivation. For example Egghe and Rousseau (1990) review a number of studies of citation motivation. Also, Case and Higgins (2000) found that citations might be made because the cited work (1) was a “concept-marker” (2) promoted the authority of the citing work or (3) deserved criticism. Links on the web, on the other hand, display different characteristics. Kim (Kim, 2000) in studying links from web based scholarly publications found a wide range of motivations for linking between web pages, including publicity, credit to an institution, and an editorial policy of encouraging hyperlinks. Wilkinson *et al* (Wilkinson, Harries, Thelwall, & Price, 2003) found that less than 1% of links to university department web sites were formal research citations.

Alastair G Smith, *Classifying links for substantive Web Impact Factors*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 305-311.

If a suitable methodology can be developed for determining the substantiveness of links, it should be possible to use the concept of substantive links to calculate Web Impact Factors that more realistically reflect the value of sites. An earlier study (Smith, 2002) showed that there was a positive relationship between the substantive WIF of electronic journals (using the simple definition of a substantive link being one made to a specific article, rather than to the journal as a whole) and the extent of metadata usage at the site. Future research could establish whether substantive WIFs correlate to other measures of effectiveness, e.g. research ratings of universities, or conventional citation counts for e-journals.

2 Methodology

The study proposes a classification of links based on the dimensions:

- Nature of source page, i.e. the page containing the link.
- Nature of target page, i.e. the page linked to.
- Reason for linking.

Initially it was intended to classify links as positive (existence indicates quality), negative (existence indicates criticism), or neutral (link is not discriminating; for instance a link to a university from a directory which includes all universities in the particular area). However in practice few links were clearly positive or negative.

Source and target pages were classified according to the schedule in Table 1.

Code	Type of page
1	General information resource, not otherwise classified
2	Research information
3	Teaching resource
4	Administrative information
5	Student assignment
6	Links list
6.1	• Bibliography/publication list
6.2	• Directory/ subject guide
6.3	• Related/useful links
6.4	• Events list
7	Discussion list archive, Blog
8	Formal publication
8.1	• Technical paper, report
8.2	• E-journal article
8.3	• Conference paper
8.4	• News item
8.5	• E-journal
8.6	• Conference
8.7	• News source
9	Personal web page
10	Main page of organisation
10.1	• Hosted or subsidiary organisation
11	Software source
12	“About” page

Table 1: page types

Reasons for linking were classified according to the schedule in Table 2:

code	**reason for link**
1	General information link
1.1	• teaching/learning
1.2	• administrative
1.3	• Research funding
1.4	• Dissemination of Research
1.5	• Employment
2	Formal research citation
4	sponsor/acknowledgement of support
5	self link to more information about creator
6	Related pages
6.1	• Related individual
6.2	• Related organisation
7	Information about geographic area
8	Advertising
9	Software download

Table 2: Reasons for linking

Links to 5 types of research oriented sites were studied; three examples of each type were chosen, making 15 sites altogether:

- Universities (Victoria University of Wellington vuw.ac.nz, Australian National University anu.edu.au, and MIT mit.edu)
- Professional institutes (The Royal Society of New Zealand rsnz.govt.nz, Institute of Professional Engineers of New Zealand ipenz.org.nz, and the New Zealand Law Society nz-lawsoc.org.nz)
- Research institutes (National Institute of Water and Atmospheric Research niwa.cri.nz, Institute of Veterinary, Animal and Biomedical Sciences ivabs.massey.ac.nz, and the New Zealand Institute for Economic Research nzier.org.nz)
- Electronic Journals (*British Medical Journal* bmj.com, *Journal of Internet Law and Technology* elj.warwick.ac.uk/jilt, and *Antepodium* vuw.ac.nz/atp)
- Individual researchers' web pages (an NZ artificial intelligence researcher mcs.vuw.ac.nz/~pondy/, a US library and Information management researcher simmons.edu/~schwartz, and a US mathematical statistician www-stat.stanford.edu/~donoho)

A listing of source and target sites for a domain or site can be obtained from search engines that implement a command which searches for links to a domain or site, for instance the AltaVista `link:` command. With most search engines, sites are presented in a ranked order, and it may not be possible to view the lower ranked sites in order to obtain a random sample of links. AltaVista in its advanced mode, however, appears to present sites in unranked order, and was used in the current study for sampling linking sites to determine the nature of the links.

Links to a target site were determined by using the AltaVista command in advanced mode (http://altavista.com/web/adv) :

```
link:xxx and not host:xxx
```

where xxx is the target domain. This excluded links made within the site itself. Site collapse was off, and the search was set to be world wide, for documents in English (to avoid having to classify sites in languages the researcher was not familiar with). Every 20th item on the list was examined to a total of 10 linking sites, except where fewer than 200 linking sites were found, and every 10th site was examined. If a site or link was no longer valid, the next on the results list was chosen. 150 links were studied in total. Searches were carried out in January 2003. Classification was carried out by the researcher.

3 Results

Frequency tables for source and target page types are given in Appendix 1, and for link types in Appendix 2. A summary is shown below.

Frequency of page types

Code	Type of page	src %	targ %
1	General information resource	9	5
2	Research information	4	2
3	Teaching resource	3	2
4	Administrative information	1	0
5	Student assignment	0	0
6	Links list	0	0
6.1	Bibliography/publication list	9	1
6.2	Directory/ subject guide	48	11
6.3	Related/useful links	3	0
6.4	Events list	1	0
7	Discussion list archive, Blog	2	0
8	Formal publication	0	0
8.1	Report	2	4
8.2	E-journal article	1	5
8.3	Conference paper	0	0
8.4	News item	1	1
8.5	E-journal	1	13
8.6	Conference	1	2
8.7	News source	4	3
9	Personal web page	5	10
10	Main page	2	36
10.1	Host/subsidiary	1	3
11	Software source	0	1
12	“About” page	3	1

Frequency of link types

code	reason for link	%
1	General	39
1.1	teaching	7
1.2	administrative	0
1.3	Research funding	1
1.4	Dissemination of Research	7
1.5	Employment	2
2	citation	10
4	sponsor	7
5	information about creator	1
6	Related pages	1
6.1	Related individual	6
6.2	Related organisation	17
7	geographic area	1
8	Advertising	0
9	Software download	1

This is an exploratory study, of relatively small samples, so not a lot of weight can be placed on the specific results, but nonetheless some interesting patterns emerge. The most common source page overall was the directory or subject guide (page type 6.2) with 72 links (48% of all links) made from these. This reinforces the role of directory and subject guide in facilitating surfing on the web. Interestingly, bibliography and publication lists (page type 6.1), as well as general information sources (page type 1), were significant sources of links, with 13 (9% of all links) each. Formal publications (page types 8.1-8.7, combining technical papers, e-journals, conferences, and news sources) were also significant, with 13 links (9% of all links).

The most common target page type is the main page of an organisation (page type 10, 54 links, 36% of all links). "Entry points" (main page of organisation, page type 10, or subsidiary/hosted organisation, page type 10.1, personal web pages, page type 9, and e-journal home pages, page type 8.5) were the targets of 101 links (67% of total links). Directory/Subject guides (page type 6.2), together with publication lists (page type 6.1), were the target of 19 links (12% of total links), indicating that these pages are useful to attract links to a site. On the other hand, specific formal publications (page types 8.1-8.4) were the targets of 15 links (10% of total links). Target pages that could be regarded as "information content" (page types 1-4, 8.1-8.4, and 11) as opposed to entry or directional, accounted for 30 links (20% of total links).

The most common reason for linking was a general information link (59 links, 39% of the total links). There were 84 "information" links (link types 1-1.5, 56% of total links) in all. Links to related individuals or organisations numbered 35 (23% of total links). Interestingly, from the point of view of comparing web links with print citations, there were 15 formal research citations (10% of total citations). Of these, 9 were to e-journals, 4 to individual researchers sites, and 2 to other university sites.

Professional organisations and research institutes were the target of the bulk of links to related organisations (link type 6.2) . Interestingly, there were no related organisation links made to general university sites – this may reinforce the view of universities as "ivory towers". Not surprisingly the bulk of "teaching and learning" links were made to universities (7 out of 11 links of link type 1.1).

A significant number of target pages for the researchers web sites were directory subject guides (10 out of 30 were page type 6.2) indicating that researchers may perform a valuable function in providing subject guides on the Internet. Of target pages at e-journals, 7 were specific articles (page type 8.2), while 19 were to the e-journal as a whole (page type 8.5). This reinforces the view that web links to e-journals are not exactly equivalent to print citations.

4 Discussion

The classifications used proved to be robust, although future research should be undertaken to determine if consistent classification can be obtained between different individuals acting as classifiers. The class of directory and subject guides (6.2 in the page types) could possibly be broken down into overall subject guides (Yahoo! etc), specialised subject guides (e.g. SOSIG) and informal subject guides (researchers' personal lists of interesting sites). In the current study, any site listings that were organised in some way were classified as directory and subject guides, and this may have been too inclusive.

This study proposes a method determining "substantive" links to a website, by sampling links to site, and determining if they are substantive, i.e. whether the link is a true indication of the significance of a site. This can then be used to determine a substantive WIF. The definition of substantiveness may depend on purpose for which WIF is being used.

The criteria for a link to be substantive will vary according to the nature of the site. For a university, a link from an organisation listing to the main site may not indicate the standing of the university, while a link to a research resource would. Links to an electronic journal might be substantive if made to specific articles, but not if made to the journal as an entity, for instance in an unselective listing of e-journals.

In the current study, some classifications of target pages that could be regarded as indicating substantive research links are:

- Research information (page type 2)
- Technical papers, reports, e-journal articles, and conference papers (page types 8.1-8.3)
- Software sources (page type 11)

Software sources (in this sample, only 2 links) are included since in computer science, software is often seen as a research output. However it could be argued that a future version of the classification should distinguish between software originating from the organisation, and software being redistributed.

19 links (13% of total links) fitted this definition of substantive links. Not surprisingly, e-journals were the target of almost half of these links (8 links, 5% of total links).

Reasons for linking that that could be regarded as indicating substantive links are:

- Dissemination of research (link type 1.4)
- Formal research citation (link type 2)
- Software download (link type 9)

26 links (17% of total links) fitted this definition of substantive linking.

Table 3 shows the numbers of links that satisfied one or other of these definitions (target page type 2,8.1-8.3,11; or link type 1.4, 2, or 9). Percentages are of the possible links, e.g. 8 of the 30 links to universities were substantive, or 27%. Over all the types of sites, 30 of the 150 links (20%) were substantive. University sites and electronic journals had relatively high proportions of substantive links made to them; while the professional institute and research institute sites had relatively low proportions. However the sample in this study is too small to put much weight on these differences.

	Number	%
Universities	8	27
Professional institutes	3	10
Research institutes	4	13
Electronic Journals	9	30
Researchers	6	20
All sites	30	20

Table 3: substantive links

The definition of substantiveness used above might well be regarded as too restrictive. For instance links made for teaching and learning purposes (link type 1.1) may indicate valuable resources. If the website is the target of links made from subject guides or directories (page type 6.2) this could be an indication of the value of the site. If the substantiveness of links is to be used in intersite comparisons, such as the calculation of Web Impact Factors, the purpose of the comparison needs to be taken into account in deciding on the definition of substantiveness. The definition used above may be useful for a research evaluation exercise; for other purposes different classes of links may need to be used.

A more efficient method would be to automatically examine sites to determine substantiveness; however an algorithm for this is problematic. A possible approach for determining substantive links to e-journals, for example, may be to count links that are

- to points other than the main page of the journal
- from pages that include the terms “references” or “bibliography”
- from another electronic journal

These are factors that could be determined by a suitable algorithm, however would cause some false drops, and would not identify all the substantive links found by a human classifier.

This study concentrated on research-oriented sites; future research could usefully examine links to other types of sites, for instance government and e-commerce sites. E-commerce links appear to be more likely to be to the main site, possibly because the design of e-commerce sites (often database driven) may make it difficult to reference specific products or services. It is likely that links to e-commerce sites may be both positive and negative in nature. E-commerce sites also may host information resources, which are likely to have similar patterns of linking to information resources at other types of sites. Shaw (Shaw, 2001) found that .com sites tended to have fewer external links, presumably in order to promote “stickiness”, i.e. keeping users on the same site.

5 Conclusion

This study proposes a classification for links, based on classifying source and target pages, and the reason for linking. The classification method was applied to a small sample of research-oriented sites, and produced some indications of the types of links made to these sites. Not surprisingly, a high proportion of links were made from directory or subject guides. Formal publications (technical reports, e-journal articles, conference papers) were significant as sources and targets of links. Formal research citations were significant, but only amounted to 10% of all citations. Links to an e-journal as a whole were more common than to specific articles. Future research could refine this classification, for example by differentiating between different types of directories and subject guides, and see if these tentative conclusions were borne out for larger samples of sites.

A definition of substantiveness of links, based on the classification, is proposed. This would be suitable for a research evaluation exercise, and 20% of the total links in the study were substantive by this definition. Future research could apply this definition to larger samples of sites in order to make meaningful comparisons between sites.

This study examined the nature of links to research oriented sites, and it would be valuable to extend the methodology to other classes of sites, for example e-commerce or government sites. While this study took as its starting point traditional citation analysis, it has shown that the nature of web links are far more varied, and relatively few of the links fall into the research oriented classes that traditional citation analysis has studied.

References

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behaviour? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Egghe, L. L., & Rousseau, R. (1990). Citations and citer's motivations, *Introduction to informetrics : quantitative methods in library, documentation, and information science* (pp. 211-227). Amsterdam: Elsevier Science Publishers.

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

Kim, H. J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science, 51*(10), 887-899.

Shaw, D. (2001). Playing the Links: Interactivity and Stickiness in .Com and "Not.Com" Web Sites. *First Monday, 6*(3). http://firstmonday.org/issues/issue6_3/shaw/ [URL checked 14 February 2003]

Smith, A. G. (2002). Does metadata count? A Webometric investigation, *Proceedings of DC-2002, Florence, 14-17 October 2002*.

Wilkinson, D., Harries, G., Thelwall, M., & Price, L. (2003). Motivations for Academic Web Site Interlinking: Evidence for the Web as a Novel Source of Information on Informal Scholarly Communication. *Journal of Information Science, 29*(1), 49-56.

Appendix 1: frequency of page types

Code	Type of page	*uni src*	*uni targ*	*proff src*	*proff targ*	*RI src*	*RI targ*	*ej src*	*ej targ*	*rschr src*	*rschr targ*	*tot src*	*tot targ*
1	General information resource, not otherwise classified	3	4	5	2	4	1	0	0	1	0	13	7
2	Research information	1	1	0	0	4	2	0	0	1	0	6	3
3	Teaching resource	0	3	0	0	3	0	1	0	1	0	5	3
4	Administrative information	0	0	0	0	2	0	0	0	0	0	2	0
5	Student assignment	0	0	0	0	0	0	0	0	0	0	0	0
6	Links list	0	0	0	0	0	0	0	0	0	0	0	0
6.1	Bibliography/publication list	1	0	2	0	0	0	6	0	4	2	13	2
6.2	Directory/ subject guide	20	6	14	0	9	1	16	0	13	10	72	17
6.3	Related/useful links	0	0	1	0	2	0	1	0	0	0	4	0
6.4	Events list	2	0	0	0	0	0	0	0	0	0	2	0
7	Discussion list archive, Blog	1	0	0	0	0	0	0	0	2	0	3	0
8	Formal publication	0	0	0	0	0	0	0	0	0	0	0	0
8.1	Technical paper, report	1	0	0	1	0	1	1	1	1	3	3	6
8.2	E-journal article	0	1	0	0	0	0	1	7	0	0	1	8
8.3	Conference paper	0	0	0	0	0	0	0	0	0	0	0	0
8.4	News item	0	0	0	0	1	0	0	1	0	0	1	1
8.5	E-journal	1	0	0	1	0	0	0	19	0	0	1	20
8.6	Conference	0	2	0	1	0	0	0	0	1	0	1	3
8.7	News source	0	0	1	2	2	1	3	1	0	0	6	4
9	Personal web page	0	1	1	0	0	0	1	0	5	14	7	15
10	Main page of organisation	0	10	1	20	2	24	0	0	0	0	3	54
10.1	Hosted or subsidiary organisation	0	0	1	3	0	0	0	0	1	1	2	4
11	Software source	0	2	0	0	0	0	0	0	0	0	0	2
12	“About” page	0	0	4	0	1	0	0	1	0	0	5	1

Appendix 2: frequency of link types

code	**reason for link**	*uni*	*proffl*	*RI*	*ej*	*reschr*	*total*
1	General information link	12	7	13	18	9	59
1.1	teaching/learning	7	0	1	1	2	11
1.2	administrative	0	0	0	0	0	0
1.3	Research funding	0	1	0	0	0	1
1.4	Dissemination of Research	4	2	2	0	2	10
1.5	Employment	0	2	1	0	0	3
2	Formal research citation	2	0	0	9	4	15
4	sponsor/acknowledgement of support	3	3	4	0	1	11
5	self link to more information about creator	0	0	0	0	2	2
6	Related pages	1	0	0	0	0	1
6.1	Related individual	0	0	0	0	9	9
6.2	Related organisation	0	14	9	2	1	26
7	Information about geographic area	0	1	0	0	0	1
8	Advertising	0	0	0	0	0	0
9	Software download	1	0	0	0	0	1