Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk

All known previous Web link studies have used the Web page as the primary indivisible source document for counting purposes. Arguments are presented to explain why this is not necessarily optimal and why other alternatives have the potential to produce better results. This is despite the fact that individual Web files are often the only choice if search engines are used for raw data and are the easiest basic Web unit to identify. The central issue is of defining the Web 'document': that which should comprise the single indissoluble unit of coherent material. Three alternative heuristics are defined for the educational arena based upon the directory, the domain and the whole university site. These are then compared by implementing them on a set of 108 UK university institutional websites under the assumption that a more effective heuristic will tend to produce results that correlate more highly with institutional research productivity. It was discovered that the domain and directory models were able to successfully reduce the impact of anomalous linking behaviour between pairs of Web sites, with the latter being the method of choice. Reasons are then given as to why a document model on its own cannot eliminate all anomalies in Web linking behaviour. Finally, the results from all models give a clear confirmation of the very strong association between the research productivity of a UK university and the number of incoming links from its peers' Web sites.

Introduction

Web link research is a growing area of information science that holds out the promise of mining knowledge of a kind that is not accessible to traditional bibliometrics (Cronin, 2001). This paper is concerned with the issue of finding the most meaningful level of aggregation for counting links between Web pages in order to facilitate such mining. This is an important matter because almost all current Web link research implicitly uses the individual HTML file as the basis for counting but, as explained below, this is not necessarily the best approach.

Hyperlinks, citations and document genres

The analogy between citations and hyperlinks has motivated much research, including the design of the search engine Google's ranking algorithm (Brin & Page, 1998). Davenport and Cronin (2000) described the promise of this approach by conceptualizing links as potential purveyors of much needed trust in the unregulated Web. That there are enormous differences between the two is also acknowledged by researchers in the area (Egghe, 2000; Harter & Ford, 2000; Björneborn & Ingwersen, 2001, Cronin, 2001; Thelwall, 2001a), but the analogy is worth pursuing, if only to draw upon the traditions of bibliometrics for inspiration and informative comparisons. Ingwersen (1998) took this furthest by creating Web versions of citation metrics, the journal Impact Factors (IF) (Garfield, 1994) of the Institute for Scientific Information (ISI).

---

[1] *This is a preprint of an article published in the Journal of the American Society for Information Science and Technology Vol 53 No. 12*, 995-1005 © *copyright 2002 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

When counting citations, individual journal or conference articles are invariably the indivisible unit source, whether the object of study is an individual author or a research group, department or country (Melin & Persson, 1996). For example, if a paper is alluded to twice in the same article then it counts as one citation in the ISI database used to calculate journal Impact Factors. Two identical references in different articles of the same issue of a journal will count as two citations, however. Similarly, two references in a paper to different pages of the same article would be expected to count as one citation whereas references to separate articles in the same journal would count as two. The ISI's model is not quite symmetrical, with cites being counted to types of contribution, such as letters, that are not included in the denominator of the IF calculation. This is an unavoidable practical problem for the ISI because a reference to a letter or journalistic piece, like those found in *Nature*, is impossible to distinguished from one to an article, if just the citation itself is available (Garfield, 1994). Nevertheless, citation studies are centred around the article as the indivisible countable document, both for the origin and target of citations. This seems to be a sensible choice since each article normally represents a single coherent piece of work. Alternatives are possible, however, and perhaps multiple allusions to the same work in reality imply that greater importance should be attributed to the target. Implementing this approach may not be practical for the ISI but would be possible in electronic repositories (Goodrum *et al*., 2002; Harnad & Carr, 2000).

Web pages are fundamentally different from articles, however. The key types of physical scholarly documents are easy to identify due to a relatively stable set of norms established over a long period of time (Meadows, 1998). Examples of these are the research monograph, the journal article, the chapter in an edited volume, the conference paper. These genres have established conventions such as the use of a single clearly identifiable title, although there are subject and publication-specific variations, such as whether an abstract is included and if a standard expository structure is to be followed. On the unregulated Web, however, pages and collections of pages can conform to existing print genres, merge multiple genres together, or create new ones (Cronin *et al*., 1998; Crowston & Williams, 2000; Haas & Grams, 2000). Even something as simple as a page title is far from ubiquitous on the Web, paralleling pages in printed books.

To illustrate the problems inherent in using HTML pages as the unitary source for links, suppose that a research group has a Web site of ten pages, each of which has a standard overall design that incorporates a link to a partner research group in another institution. When counting links, this would count as ten, but all originate from one motivation, the partnership. More worryingly, the extra nine links come from the design decision to put links on all pages rather than just on the home page or on a links page. This problem could be avoided if the group's Web site were to be conceptualised as a single document: aggregating links from all pages and discarding duplicates. The issue of aggregating links from a small site is not straightforward, however, even if the question of what constitutes a Web site is temporarily sidestepped, because links from all papers in the site of an online journal should presumably be counted separately.

The research described in this paper operates on the simplistic level required to produce a workable algorithm and does not attempt to directly address the extremely complex issue of what should constitute a document on the Web, or how genres could be identified. It is hoped, however, that the quantitative results produced would be of some service to the debate over the related qualitative issue. The difficulty with attaining a workably simple definition of document or genre is illustrated by the findings of Bates and Lu (1997) of the variety of approaches to the creation of what might be a single genre type, the personal home page. Dillon and Vaughan (1997) argue for the need to combine physical form and semantic content in order to understand users' perceptions of genre, which speaks for a multifaceted model. Nevertheless, the importance of the issue can be seen in its spawning of much computer science research. For example, Rehm (2002) is building an automatic genre classification tool in the hope that a workable algorithm, even if imperfect, can aid in the development of more effective search engines.

Previous Web link research

Research into the patterns by which hyperlinks interconnect the Web goes back at least as far as 1996 (Larsen, 1996) and has included work on assessing bibliographic distributions in an electronic context (Rousseau, 1997). Computing research has used links to improve information retrieval algorithms, Kleinberg's (1999) topic driven algorithm in addition to that of Google. The link structure of the Web itself (at least that part of it in AltaVista's databases) has also been described (Broder *et al*., 2000) and individual types of linking phenomena explored (Björneborn, 2001a). Other research has investigated the use of links to help rate Web sites and their content (Cui, 1999; Darmoni *et al*., 2000). One key problem of the past few years has been the development and assessment of techniques to measure counts of links between academic Web sites. Ingwersen (1998) introduced the Web Impact Factor (WIF) for this, in one version measuring the 'impact' of a Web space through the ratio of external pages containing a link to any page in the target space divided by the number of pages in that target space, which could be a Web site or even a national or international domain. Since then the WIF has undergone modifications in response to early poor results (Smith, 1999; Thelwall, 2000; Thelwall, 2001f) and now appears to be measuring something closely correlating with research productivity of universities, at least in the UK, Australia and China (Thelwall, 2001a; Smith & Thelwall, 2002; Tang & Thelwall, 2002). This highlights the possibility to investigate the Web projection of concerns traditionally associated with bibliometrics and scientometrics, including patterns of informal communication, research dissemination and collaboration (Cronin *et al*., 1998, Leydesdorff & Curran, 2000; Wikgren, 2001). A significant relationship has also now been found at a lower level: link counts and rankings for library schools (Chu *et al*., 2002). A second approach made possible by improved understanding of the modified WIF is the deeper mining of university Web link data through the factoring out of the research productivity-related component. This has lead to the discovery of new patterns that are not present in the raw data, such as a geographic trend in university Web site interlinking (Thelwall, 2001b). This approach has the promise of being able to provide a mechanism for researchers interested in the Web to be able to test their own hypothesis about patterns of use between communities through Web link analysis facilitated by free online databases and tools such as those at http://cybermetrics.wlv.ac.uk/database and http://www.archive.org.

There are essentially three different implicit document models in the link research described above. If a search engine is used for raw data in order to count pages that link to a given URL then the document model is the HTML file, at both the source and target level. Some of the research, however, has used advanced searches, such as available from AltaVista, to find pages that link to a domain. In this case the model is hybrid: HTML files for the link source but whole domains for the target. This would lead to situations, for example, where a page that contained links to two different pages hosted by the target domain only counted as one matching page. A variant of this has also been used, where the URL of the target directory is the URL to find a link to, catching all pages that link to any file in the directory. This is again a hybrid model for the same reasons. The research that has used a Web crawler to collect raw data has also used the HTML file model, but in this case has been able to count separately multiple links in the same page to different target URLs.

The research question

Four different heuristic-based models of Web documents will be defined and put into practice on a large data set, that of UK university Web sites in July 2001. These are extensions of Björneborn's (2001b) technique for aggregating links at the source domain level as a data filtering tool. The relative efficacy of the different approaches will then be assessed under the assumption that an effective definition will produce results that correlate closely with a measure of institutional research productivity. The rationale for this is that previous studies have found just such an association (Thelwall, 2001a; Smith & Thelwall, 2002). Results will be analyzed both in terms of total links to university Web sites, and total links between university Web sites.

The concept of the document on the Web

This section will focus on the inherent problems with identifying coherent documents on the Web. A discussion of this will be preceded by an introduction to two related concepts: the Web page and the Web site.

The Web page

The Web page is not a clearly understood and defined entity. There are publicly available seemingly authoritative definitions that incorporate fundamental disagreements (Thelwall, 2002a). The following are all plausible alternatives to five of the major components of a definition, with illustrative alternatives.

- **File format** An electronic file validly encoded in the language of the Web, HyperText Markup Language (HTML) - or - any file type accessible through a modern Web browser including non-HTML formats such as plain text, PDF (Portable Document Format) and Microsoft Word.
- **Access mechanism** Requests made using official 'port number' of the Web, 80 – or – requests made using the official computer request language of the Web, the HyperText Transfer Protocol (HTTP, as seen at the start of many URLs) – or – requests made using any mechanism available to a modern Web browser, including common non-Web protocols such as FTP (File Transfer Protocol).
- **Scope** Public Web pages that are available to all Web users – or – public and private Web pages, including password protected pages and Intranet and Extranet pages.
- **Permanence** Static resources only – or – all resources, including dynamically-created Web pages such as search engine results pages.
- **File numbers** Should pages built up from separate files using the HTML frameset feature count as one combined page or one page per file?

The phrase "Web page" is almost certainly used in practice with a variety of meanings, perhaps including most combinations of those mentioned above, and its actual meaning in any situation will be dependant on the context and technical background of the communicators. Probably the need to be precise about exactly what constitutes a Web page is actually very rare outside of counting exercises.

The Web site

A much easier question, but also subject to context-sensitive interpretations, is that of what can be referred to as a Web site. These are probably most frequently associated with domain names. For example the Google home page is at http://www.google.com/ but its Web site would probably be recognized as all URLs with domain name ending in google.com. Other plausible definitions would include all URLs with identical domain names and the same first few directories. As an example of the latter, all URLs beginning with http://www.scit.wlv.ac.uk/~cm1993/ would probably be accepted as being the author's "Web site", although some of the collections of pages matching the URL might be seen as small sites in their own right. In practice, Web sites can also be identifiable even if based on multiple domains as long as there are sufficient cues in the interface. Interestingly, one of the major Web editors, Microsoft's FrontPage, describes collections of related pages as just a "Web" and recognizes standard types of "Web" such as the "Personal Web" and the "Customer Support Web". This may be an attempt to avoid confusing users who may have a different understanding of the more standard term "Web site".

The Web document

There are varied organizational schemas for coherent collections of content on the Web: single/multiple pages; single/multiple domain names; single/multiple file directories used; hierarchical/sequential/structured link navigation; database driven/queried. Moreover, in terms of content alone it is a struggle to find a sufficiently general metaphor for the Web as the scope of its coverage of the classes of activity below testifies to.

- **Print media style publications** The Web contains e-journals, e-books, online conference proceedings, digital libraries.
- **Information retrieval** The Web hosts online library catalogues and general search engines.
- **Electronic media** The Web hosts photographs, films, music and even three-dimensional environments such as archaeological reconstructions.
- **Informal communication** Email archives are found on the Web, bulletin boards and the real-time discussion forum.

The document is therefore an awkward concept to give a general Web definition for. Additionally, recognizable genres of print and electronic document are commonly found in pieces in the Web. For example, PowerPoint slides can be automatically converted into a set of HTML files and associated image files. Books can be found on the Web too, some as a single huge PDF or HTML file, others in thousands of individual sections. Questions must then be asked of the new genres. Presumably a frequently asked questions list is a coherent document, irrespective of whether it is a single file or broken up, or with single or multiple authors. But if, say, a small research group Web site contains multiple types of resource, perhaps a home page, a background information page, twenty PDF project findings and papers as well as a links list, is this a single document or multiple documents, and if the latter, how many? The purpose of having a definition of the Web document being counting links, it is appropriate to include a discussion of why two different links should count separately as two or together as one. The object of the exercise is to be able to model the average behaviour of Web authors and so it makes sense to focus on these and allow one link per recognisable coherent body of work. Taking this into account in addition to the concerns above, the following is proposed as a working definition.

> A Web document is a body of work with a consistent identifiable theme produced by a single author or collaborating team. It may consist of any number of part or whole unrestricted access electronic files retrievable over the Web using a modern browser.

This definition is clearly far from being a razor for separating document and non-document collections because of the inclusion of imprecise concepts such as “identifiable theme” and “collaborating team”. One natural interpretation would be to equate the Web document with a small Web site or a small subsite of a larger site. The problem remains of how to make the definition more prescriptive in order to implement it. Two possible solutions are: to allow panels of human experts to cluster pages or parts of pages, into documents, or to develop an heuristic for page aggregation, accepting that it represents a simplistic model. Two approaches to automatic clustering are possible. The first is to develop an algorithm to automatically merge ‘similar’ Web pages into documents for counting purposes, probably based upon both their content and enveloping link structure. The second possibility, pursued in this paper, is to develop simple URL-based heuristics to automatically merge Web pages for counting purposes. Many Web editors, including Microsoft FrontPage, have the default setting of storing all related documents in the same folder, perhaps using subfolders for auxiliary files such as images. This makes the directory or folder a plausible level of aggregation of HTML files into documents, in addition to the other natural levels of domain name and entire university site, as defined in Table 1. Whichever methodology is adopted, additional research is needed to assess its efficacy.

Table 1 Document-based models of Web organization

Model	Description
Individual Web page	Each separate HTML file is treated as a document for the purposes of extracting links. Each unique link URL is treated as pointing to a separate document for the purposes of finding link targets. URLs are truncated before any '#' character found to avoid multiple references to different parts of the same page.
Directory	All HTML files in the same directory are treated as a document. In other words target URLs are automatically shortened to the position of the last slash, and links from multiple pages in the same directory are combined and duplicates eliminated.
Domain name	As above except all HTML files with the same domain name are treated as a document.
University	As above except that all pages belonging to a university are treated as a single document.

Methodology

The system of universities chosen for the study is that of the UK because of its relatively definitive government Research Assessment Exercise (RAE), the latest having been published in 2001 to cover the period 1996-2000, ideal for the present study. It comprised 68 separate subject-based groupings, each controlled by a panel of experts. Universities submitted sets of faculty to each appropriate panel, and their average research contribution over the time period was then assessed using peer review by the panel members, based primarily on the best four publications produced by each submitted scholar (http://www.qaa.ac.uk, http://www.rae.ac.uk). The results were to be used to determine the destination of a major part of official government research funding until the next exercise, and so it was taken very seriously by all concerned. It is possible to use these figures to calculate the average research quality per faculty member, a number with a theoretical minimum of 0 (no research conducted in the institution) and maximum of 7 (all faculty conduct internationally excellent research). The average figures are based upon the assumption of a linear scale for the 7 categories, but the results are believable and probably on an international level the most plausible statistics of their kind in the world. Multiplying the weighted average RAE score by total faculty gives what will be called the *research productivity* of an institution.

The link structure data

The link structure of the UK universities was obtained from a publicly available database of 108 major university institutions as of July, 2001 (Thelwall, 2001d) created by a specialist information science Web crawler (Thelwall, 2001c), version 2 (http://cybermetrics.wlv.ac.uk/data) in order to avoid the known problems of reliance upon commercial search engines for data (Rousseau, 1999; Mettrop & Nieuwenhuysen, 2001), even if it can be accessed directly (e.g. from archive.org). The database represents not the entire Web site in each case but only those pages that can be found by following links from the home page, excluding recognized mirror sites and pages from which robots are banned. In the case of universities with non-HTML links on the home page, the starting point chosen was a list of departmental home pages instead. The term 'university Web site' in this paper will be used to refer to this collection of pages.

The link structure database consists of a separate text file for each institution, giving a list of the URLs of all pages crawled together with the URLs of all identified referred to URLs in the page (standard HTML links, client side image maps, server and meta tag redirects) with duplicate URLs removed and all URLs truncated the first '#' character, when present. This last point means that in one page there cannot be links to two or more parts of a common target page.

Obtaining counts based upon the document definitions

Each of the definitions of document described in table 1 was realized through the construction of a computer program to process the database. Each program counted the number of links from documents in each institution to all others. Links between different documents in the *same* institution were never calculated because these are not likely to carry useful information in this context (Thelwall, 2001a). A link was identified as targeting a recognized other UK university by extracting the domain name and comparing it to a list of recognized domain names. A match was recorded if the link domain name was either equal to a recognized root domain name or terminated in a dot followed by a recognized domain name. In the example of Edinburgh, with recognized roots ed.ac.uk and edinburgh.ac.uk, the following URLs would match:

```
http://[anything].ed.ac.uk/[anything]
      http://ed.ac.uk/[anything]
http://[anything].edinburgh.ac.uk/[anything]
      http://edinburgh.ac.uk/[anything]
```

but the following would not, despite its inclusion of ed.ac.uk in its domain name.

```
http://qmced.ac.uk/
```

During the link counting process described below, all link target URLs were automatically converted to canonical name form, e.g. http://www.doc.edinburgh.ac.uk/index.html would be converted to http://www.doc.ed.ac.uk/index.html. Domain names starting with "www." were also truncated to cut out this part to avoid duplication through the common practice of having multiple equivalent versions of Web domain names with this pattern, particularly in previous years. For example wlv.ac.uk and www.wlv.ac.uk used to give the same site. URLs ending in standard file names such as index.html and main.htm had these removed in order to avoid double counting links to home pages.

- For model 1, the link count from institution A to institution B was a simple count of URLs in the database of the link structure of university A that were from B.
- For model 2, the link count from institution A to institution B was calculated in two steps. Firstly a new link structure file was constructed for institution A with all link URLs truncated immediately before the final slash, if present. At the same time the source page URLs were truncated in the same way and all links associated with the same truncated source URL merged and duplicates eliminated. The model 1 program was then run on this new directory level document file to obtain the link count.
- For model 3 the same process was conducted as in model 2 except that each source page and target URL were truncated at the first slash following the domain name, if present.
- For model 4 the same process was conducted as in model 3 except that each source page and target URL were additionally cropped to the canonical form of the domain name, discarding any additional domain sections. The end result of this is to allow a maximum of one link between any pair of universities.

Model 2 is actually (unavoidably) flawed in its execution, although it is believed that this will only have a small impact on its use in practice. The problem is that URLs can point to a directory rather than a particular file in the directory and be served with its default file. For example, a request for the URL http://www.scit.wlv.ac.uk/~cm1993/ will retrieve the file named index.html from the virtual folder ~cm1993, being equivalent to http://www.scit.wlv.ac.uk/~cm1993/index.html. In this case it can be easily seen that ~cm1993 is a folder rather than a file because of the slash at the end of the URL. Unfortunately, the originator of the link can also omit the terminating slash, as in http://www.scit.wlv.ac.uk/~cm1993 and still be served with the same page because of the intelligent behaviour of the Web server. The algorithm, as implemented, would see ~cm1993

as a file name, because it is not followed by a slash and in model to would truncate it incorrectly to http://www.scit.wlv.ac.uk/. It would be theoretically possible to check all URLs to see if the page referred to would be unchanged if a slash were to be added (e.g. see if http://www.scit.wlv.ac.uk/~cm1993 fetched the same page as http://www.scit.wlv.ac.uk/~cm1993/) but this would be a large task and error-prone when the target page was already gone and replaced by an error message or contained a variable component such as a text based page counter, as the page associated with the URL in this example does.

Results - total external links to universities

Table 2 shows Spearman correlation coefficients for the counts of links to each university against research productivity and Figures 1-4 show the raw data from which these were derived. Spearman is used instead of Pearson because although a linear trend is evident in the first three graphs the independent variable has a significantly non-normal distribution, moreover the dependant variable variance is clearly not uniform across the spectrum of its values.

Table 2. Spearman correlations between counts of links to each university and their research productivity for 108 UK university institutions. All are significant at the 0.1% level.

Model	Correlation
File	0.920
Directory	0.925
Domain	0.923
University	0.807

FIG. 1. Link counts based upon the file model against estimated research productivity for UK universities

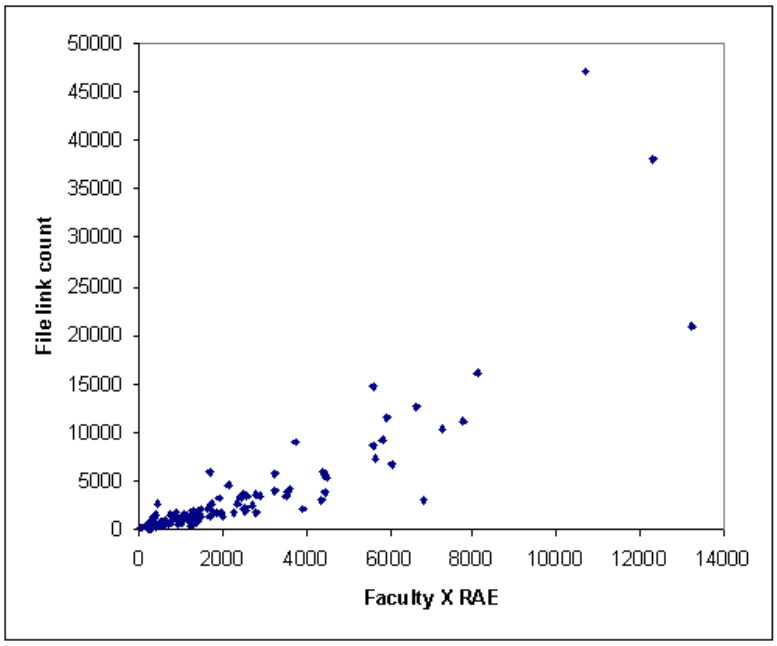


FIG. 2. Link counts based upon the directory model against estimated research productivity for UK universities

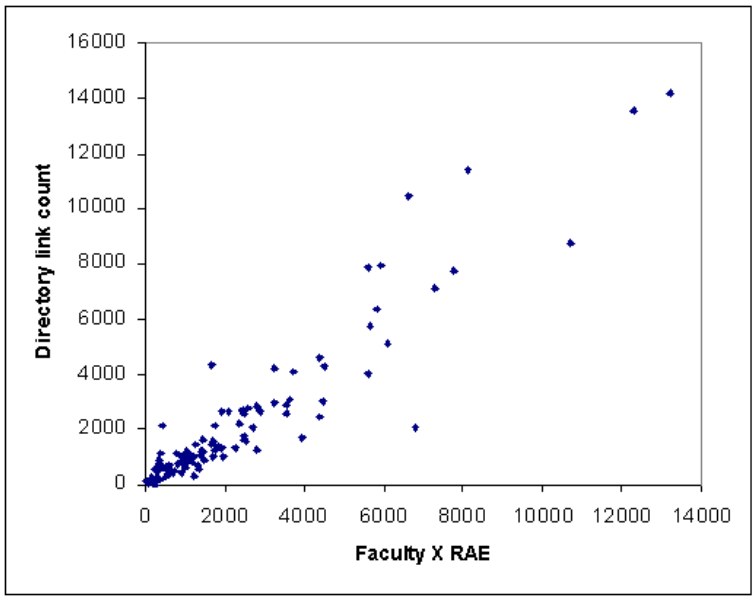


FIG. 3. Link counts based upon the domain model against estimated research productivity for UK universities

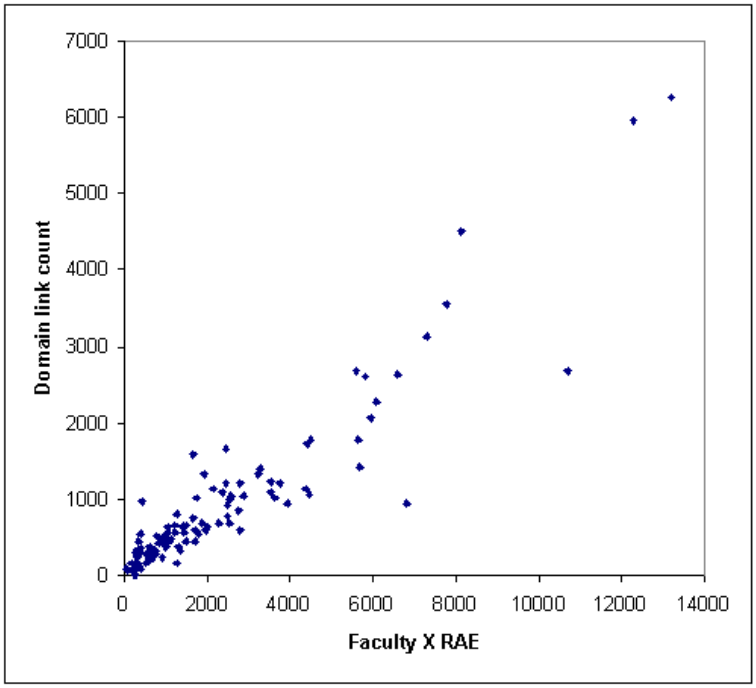
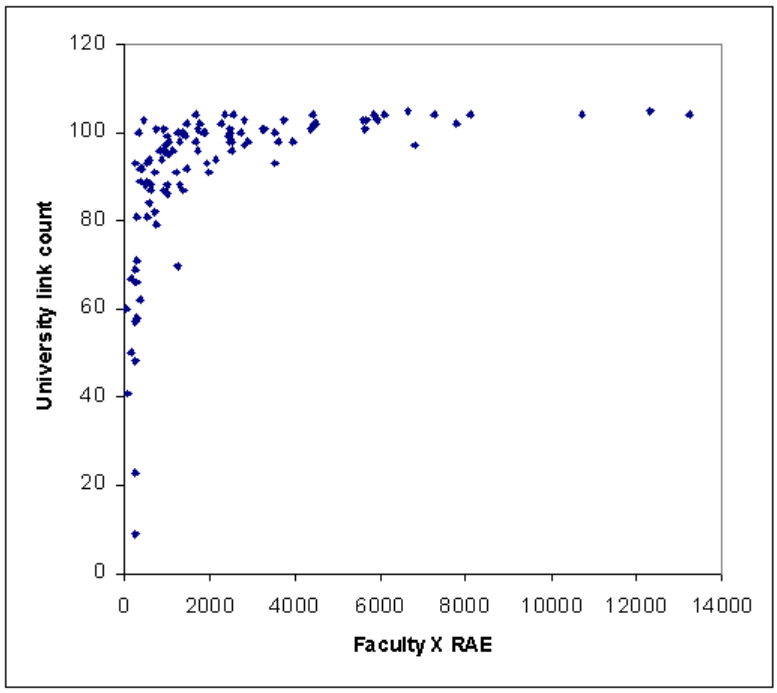


FIG. 4. Link counts based upon the university model against estimated research productivity for UK universities



From the high first three correlation coefficients and striking linear trends in Figures 1 to 3 it can be seen that the first three models are all extremely plausible and very similar in overall terms, although different for individual universities. The second and third models show a slight improvement over the first, but do not remove all outliers. The change in correlation coefficients between the file and directory model does not reflect the increased linearity of the graph because Spearman only uses the rank of the data. In particular, the increased conformity of the rightmost three points is not reflected at all. Pearson correlation coefficients would show this (file: 0.848; directory:0.934) but for the reasons given above these are not sufficiently reliable to be trusted for comparison purposes. The most extreme cases will be analysed to see which universities suffered the biggest relative change in values between models, and reasons will be sought for the changes through analysis of the relevant data files.

Outliers

Three main outliers can be seen from an examination of the first three graphs – universities that have higher or lower link counts than would be expected for their research productivity. One institution with a low link count in these graphs is the School of Oriental and African Studies (SOAS), an autonomous institute of the large federal University of London (research productivity: 6808, links: 2495 in Fig. 1). This, and most of the other institutions in the University of London are unusual in that they are not fully multidisciplinary (O'Leary, 2001). Given differential Web use between disciplines (Kling & McKim, 2000; Jacobs, 2001) it is not surprising that some of these appear as outliers. It seems reasonable to speculate that the disciplines in SOAS use the Web less heavily than the average for scholarly research. The second consistent outlier is the University of Wolverhampton (447, 2354) which has a high count because of its hosting of a very popular Web link resource, the UK clickable map of universities, which almost all universities link to at least once. The third is Heriot-Watt (1675, 5415) because of shared sites with the University of Edinburgh. All of these have been remarked upon before in previous papers (Thelwall, 2002b), but it is noteworthy that the document models have not been able to eliminate these outliers and they still represent fundamental anomalies.

Analysis of changes in results

Some universities had particularly large changes in relative numbers of link counts in the move from model to model. The largest changes are shown in tables 3 and 4 and individual large changes are discussed.

TABLE 3. Universities with link counts that decrease the most in relative size between the file model and the directory model

Name	File model link count	Directory model link count	File model to directory model ratio
University College London	47,205	8,727	5.4
University of Cambridge	38,207	13,525	2.8

- For University College London, a constituent college of the University of London, the decrease was due to database links from Cambridge, as described below Table 6.
- For Cambridge, there were a few sites that used its Web long analyzer package, the output of which is a large number of Web pages saved in the same folder, all with a credit link to the Cambridge site at http://statslab.cam.ac.uk/~sret1/analog/. The directory model reduced the number of credit links to one per directory of results.

TABLE 4. Universities with link counts that decrease the most in relative size between the directory model and the domain model

Name	Directory model link count	Domain model link count	Directory model to domain model ratio
University of Leeds	10433	2630	4.0
University of Sheffield	5743	1432	4.0
University of Birmingham	7950	2067	3.8

- For Leeds, the relatively large decrease from the directory to the domain model was attributable to two target URLs associated with a single resource, known as LaTeX2HTML, which attracted a large number of links. This is a program that generates HTML files that contain credit links to the owner's (old) site. The implication is that there are a lot of files scattered across different directories of different Web sites created by this software.
- For Sheffield, it was found that its pages are unusually clustered on the main domain. For example 923 of its 1429 external links (domain model) come from the main domain (shef.ac.uk). This is a serious concern for the domain model because it is an example of an unusual information organization producing undesirable results.
- For Birmingham, there were very many links to a copy of the Web site of CatchWord Ltd., which claims to host over 1,100 journals from over 65 publishers (CatchWord, 2002). It is organized in multiple directories but a single domain, but there were enough resources in common directories to give an overall decrease to the Birmingham results.

Results - total links between universities.

In Thelwall (2001e) it was found that links between individual universities can also correlate quite highly with the product of the research productivity of both source and target institution. Greater variability and a lower overall correlation is to be expected, however, because the lower number of links involved gives more scope for the law of averages to apply. Table 5 shows Spearman correlation coefficients for the data. The Kruskal-Wallis test is perhaps more appropriate for the two value university model data, but this gives a result of the same significance. The Spearman result is used in the table for ease of comparison.

TABLE 5. Spearman correlations between counts of links between pairs of universities and the product of their research productivity for 11,556 ordered pairs of UK university institutions. All values are significant at the 0.1% level.

Model	Correlation
File	0.223
Directory	0.739
Domain	0.832
University	0.186

FIG. 5. Link counts between institutions based upon the file model, against the product of the estimated research productivity for source and target university

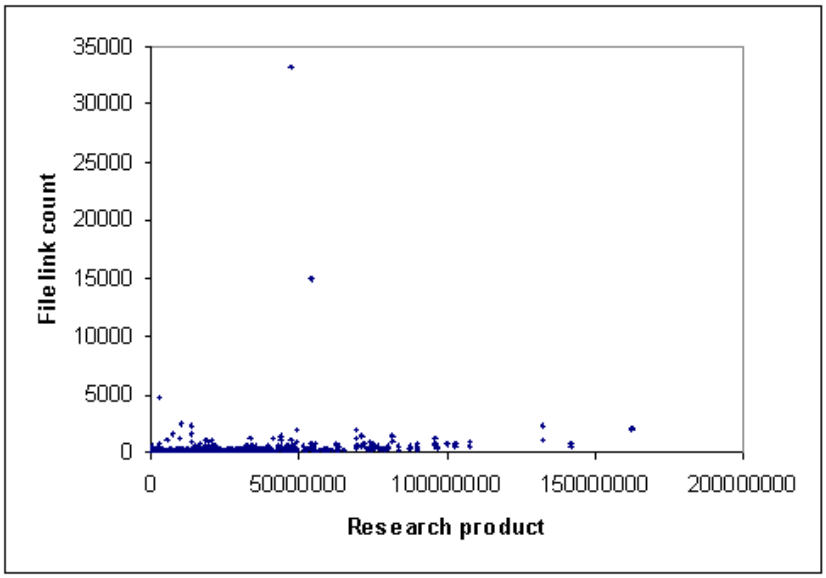


FIG. 6. Link counts between institutions based upon the directory model, against the product of the estimated research productivity for source and target university

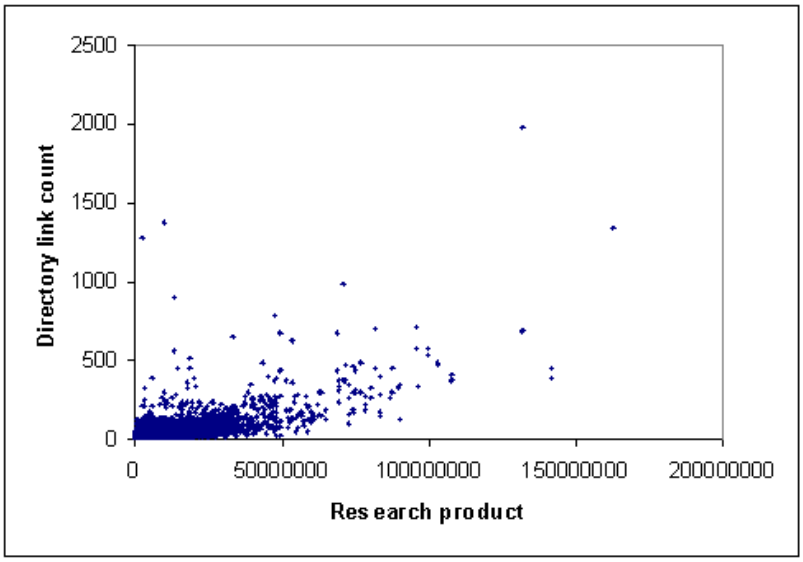


FIG. 7. Link counts between institutions based upon the domain model, against the product of the estimated research productivity for source and target university

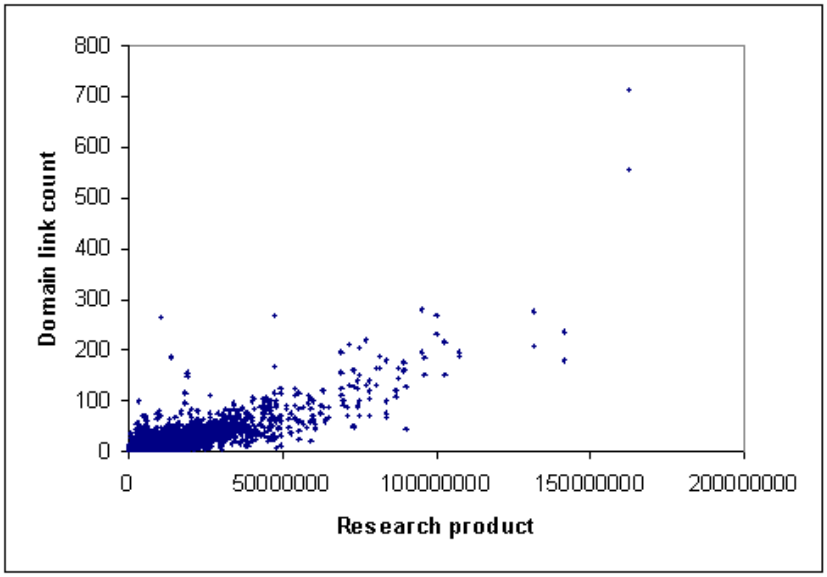
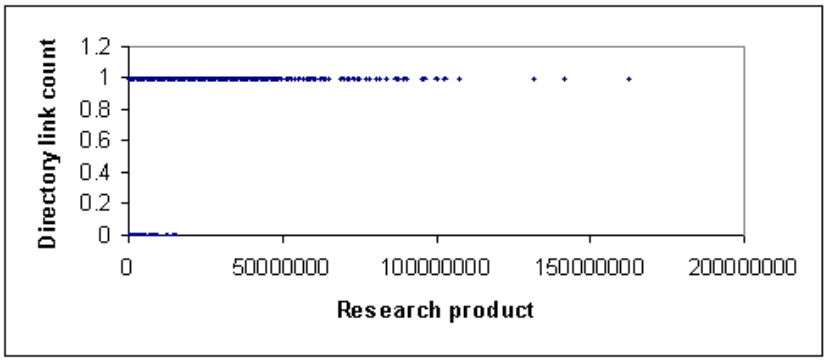


FIG. 8. Link counts between institutions based upon the university model, against the product of the estimated research productivity for source and target university



From the graphs and correlations it can again be seen that the domain and directory models both give dramatic improvements over the standard file model, with a reduction of the numbers of obvious outliers in the graphs. The changes in the correlation values perhaps reflect a reduction in outliers more than a more effective ordering within the main cluster of the data. This provides clear evidence that the domain model is the most robust of the four. Figure 8 is an impressive demonstration that beyond a certain research productivity point, all pairs of universities will link to each other.

All of the graphs suffer from having to display too many points for the space available, and so it is not possible to visually interpret the density of points in the main areas, particularly in Figure 5. Also, in the first two graphs there are coincident points for the extreme right hand value, which may confuse the reader since all points should be displayed in pairs, there being two values for each pair of universities.

Outliers

Since the outliers are less pronounced in Fig. 7 than in 5 and 6, only those from the domain model will be analyzed, being the most persistent. Many of the other outliers for earlier models will be caught in the next section, however.

The shape of the domain link count graph contains the hint of a curve rather than a straight line and this has implications for the identification of outliers. In this context the pair of point on the right of the graph, representing links from Cambridge to Oxford and back,

may or may not be outliers. One possible explanation for the apparently high interlinking is their geographical proximity, known to impact on Web link counts (Thelwall, 2001b).

The following are a selection of high link count anomalies from the domain model.

- South Bank University to Oxford University, (research product 10233426, links 264) caused by extensive credit links on sites created by a lecturer associated with both universities to his own pages hosted by sites in Oxford.
- Heriot–Watt to Edinburgh (13603248, 188), caused by a variety of links to different areas, including 15 to the main site and 10 to the Department of Geography site. Heriot-Watt is also located in Edinburgh, but is only approximately a third of the size of Edinburgh University. The extensive linking provides evidence of the influence of the larger university and a perhaps inevitable close working relationship for two similar institutions in a relatively isolated part of the country.
- Edinburgh to Heriot–Watt (13603248, 184), caused by high numbers of links to a few sites, for example 31 to the main site, 28 and 20 to two different domains of the Edinburgh Business School. The relatively few sites may reflect the smaller size of the target university, but its business school seems to have created a significant impact.

The pair of universities with the highest research productivity product but without any links was Cranfield to Oxford. Cranfield is a slightly unusual university, having a postgraduate focus as well as a specific mission to provide research for industry.

Analysis of changes in results

Tables 6 and 7 show the link counts with the greatest proportional reduction between models.

Table 6. Link counts between pairs of universities that decrease the most in relative size between the file model and the directory model

Source university	Target university	File model link count	Directory model link count	File model to directory model ratio
University of Warwick	University of Central London	33228	55	604
University of Warwick	University of Cambridge	14929	232	64
University of Coventry	University of Cambridge	1593	31	51

- For Warwick to both UCL and Cambridge, the high number of links using the file model was due to thousands of automatically created links to biochemistry database queries on the target site. Each query was expressed as a unique URL and was recorded as a link to a page. With the directory model, these links reduced to just one.
- For Coventry to Cambridge, the decrease was due to large sets of log files on the Coventry Web site that used the Cambridge Web log analyzer, described above.

Table 7. Link counts between pairs of universities that decrease the most in relative size between the directory model and the domain model

Source university	Target university	Directory model link count	Domain model link count	Directory model to domain model ratio
University of Bath	University of Cranfield	210	4	52
Leeds Metropolitan University	University of Birmingham	212	5	42
University of Hertfordshire	University of Birmingham	386	16	24
Northampton University College	London Guildhall University	46	2	23
University of Huddersfield	University of Leeds	321	14	23
South Bank University	University of Reading	1274	57	22
University of Greenwich	University of Leeds	152	7	22

- Bath's library site contains a large number of links to another copy of the Web site of CatchWord Ltd.. The directory structure used in the database means that aggregation at the domain level is effective in pooling together links into this site.
- Leeds and Hertfordshire both have libraries with extensive links to CatchWord's database on the Birmingham university Web site.
- Northampton has one unusual site: a set of pages giving links to resources for some staff teaching and learning modules. A large number of these links are to different pages on a single extensive multiple directory digital libraries site: DeLiberations, which explains the reduction in links at the domain level.
- Huddersfield has one domain that hosts a large number of lecture notes and other resources converted into HTML using the LaTeX2HTML program created at Leeds. Greenwich exhibits a similar phenomenon.
- South Bank contains a single large maths resource site with many pages having credit links to Reading, the former employer of the creator, and presumably the former home of the resource.

Discussion and Conclusions

Based upon the theoretical analysis, the qualitative study of the outliers and the statistics - particularly for the counts of links between pairs of universities - there is now considerable evidence that the both the directory and domain models are better than the default file model from the perspective of counting links between sites. The domain model is suggested to be the least susceptible to spurious perturbations in its results, although it is not immune. A weakness in the methodology, however, is that only one country has been used for the tests. Although it seems likely that the directory and domain document models would be generally applicable, it is not inconceivable that there would be countries to which it would not apply, for example if an URL, directory or domain structure was commonly used that was substantially different to that used in the UK.

Based upon the analysis of outliers and individual large changes in values, the advantage of the directory model appears to be primarily in reducing the impact of credit links on multiple pages of a single site and mass collections of database links, whereas the domain model reduces the impact of both widely linked to individual resource-related pages and widely linked to entire sites. Neither could prevent the continued presence of clear outliers, one of which was the result of an individual very popular page hosted on the sites of a low research productivity institution. It can easily be seen that no pure document model will be able to cope with this type of anomaly. The new models appeared to be eliminating sources of anomalies that it would be desirable to reduce and also producing overall increases in the correlation measure from both perspectives. It is perhaps surprising how little the new models improved the first three graphs, given the improvements in Figures 5 through 7, and a

possible explanation for this, in addition to the problems with using Spearman in this context discussed above, is that it is likely that some individual universities had consistently high or low link counts from most other universities and that these individual small anomalies accumulated into the bigger anomalies seen on Figures 1 to 3.

It is interesting that the extraordinarily high correlation found for the first three models applies to Web sites "Warts and all". In other words link targets are unfiltered for content unrelated to research, which they certainly do have (Thelwall, 2001a). It seems very likely that Web activities unrelated to research are influenced by it directly or indirectly, for example through the availability of computing resources or the development of the technological know-how. One further potential approach to extracting more information would be to develop metrics focusing on the spread of targets, rather than the absolute quantity, under the assumption that multiple links to the same target may have common underlying causes. A second, more difficult but potentially revealing more information, would be to attempt to classify types of targets in a way that would be enlightening about the extent of use of "new modalities of scholarly communication" (Cronin *et al*., 1998). In this context, a potential negative outcome of the paper, in practical terms, is that it commends two methodologies for link counting that are not available through commercial search engines, as used in previous studies (Ingwersen, 1998; Smith, 1999; Thelwall, 2000). In response to this issue it is stressed that the comparable high correlation for the file model supports its continued use for research, provided that methodological safeguards are in place. In addition to this, the code necessary to apply the four models will be made freely available on the Web site hosting the national university system crawl data (http://cybermetrics.wlv.ac.uk/database/). It is hoped that this, perhaps in conjunction with the historical archive of Web pages (http://www.archive.org), will form a valuable resource for Web link researchers.

The approach described in this paper risks accusations of being simplistic because of the aggregation entire university Web sites. Its ultimate objective, however, is the opposite: to improve metrics so that they can provide a reliable baseline for research that would give quantitative support for hypotheses concerning differential Web use and linking (Thelwall, 2002b), a possibility also illustrated by the example of SOAS above.

Acknowledgements

The author is grateful for the useful comments and suggestions made by the reviewers.

References

Bates, M. J. & Lu, S. (1997). An Exploratory Profile of Personal Home Pages: Content, Design, Metaphors, Online and CDROM Review, 21, 331-340.

Björneborn, L. (2001a). Small-world linkage and co-linkage. In: Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (pp. 133-134). New York: ACM Press.

Björneborn, L. (2001b). Necessary data filtering and editing in webometric link structure analysis. Royal School of Library and Information Science.

Björneborn, L. & Ingwersen, P. (2001). Perspectives of webometrics. Scientometrics, 50(1), 65-82.

Brin, S. & Page, L. (1998). The Anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30, 107-117.

Broder, A., Kumar, R., Maghoull, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the web. Computer Networks, 33(1-6), 309-320.

CatchWord, (2002). Profile. Available: http://dandini.cranfield.ac.uk/profile.htm Accessed 22 January, 2002.

Chu, H., He, S. & Thelwall, M. (2002, to appear). Library and Information Science Schools in Canada and USA: A Webometric Perspective, Journal of Education for Library and Information Science.

Cronin, B. (2001). Bibliometrics and Beyond: Some thoughts on web-based citation analysis. Journal of Information Science, 27(1), 1-7.

Cronin, B., Snyder, H.W., Rosenbaum, H., Martinson, A. & Callahan, E. (1998). Invoked on the web. Journal of the American Society for Information Science, 49(14), 1319-1328.

Crowston, K. & Williams, M. (2000). Reproduced and emergent genres of communication in the world wide web, Information Society, 16(3), 201-15.

Cui, L. (1999). Rating health Web sites using the principles of citation analysis: a bibliometric approach. Journal of Medical Internet Research, 1(1), e4. Available: http://www.jmir.org/1999/1/e4/index.htm

Darmoni S. J., Thirion B., Douyère M., Challoub C. & Leroy J. P. (2000). Mesure de l'impact des sites Web : le Web Impact Factor. L'exemple des CHU français. Revue du Praticien - Médecine Générale, 14(516), 2079-2080

Davenport, E. & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In: Cronin, B. & Atkins, H. B. (eds.). The web of knowledge: a festschrift in honor of Eugene Garfield. Metford, NJ: Information Today Inc. ASIS Monograph Series, 517-534.

Dillon, A. & Vaughan, M. (1997). "It's the journey and the destination": Shape and the emergent property of genre in evaluating digital documents, New Review of Multimedia and Hypermedia, 3,91-106.

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. Journal of Information Science, 26(5), 329-335.

Garfield, E. (1994). The impact factor, Current Contents, June 20. Available: http://www.isinet.com/isi/hot/essays/journalcitationreports/7.html

Goodrum, A. A., McCain, K. W., Lawrence, S. & Giles, C. L. (2001). Scholarly publishing in the Internet age: a citation analysis of computer science literature. Information Processing and Management, 37(5), 661-676.

Haas, S. W. & Grams, E. S. (2000). Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages. Journal of the American Society for Information Science, 51(2), 181-192.

Harnad, S. & Carr, L. (2000). Integrating, navigating, and analysing open eprint archives through open citation linking (the OpCit project). Current Science, 79(5), 629-638.

Harter, S. P. & Ford, C. E. (2000). Web-based analyses of e-journal impact: approaches, problems and issues. Journal of the American Society of Information Science, 51(13), 1159-1176.

Ingwersen, P. (1998). The calculation of Web Impact Factors. Journal of Documentation, 54, 236-243.

Jacobs, N. (2001). Information technology and interests in scholarly communication: A discourse analysis, Journal of the American Society for Information Science, 52(13), 1122-1133.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment, Journal of the ACM, 46(5), 604-632.

Kling, R. & McKim, G. (2000). Not Just a Matter of Time: Field Differences in the Shaping of Electronic Media in Supporting Scientific Communication. Journal of the American Society for Information Science, 51(14), 1306-1320.

Larson, R. R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace, Proceedings of the AISS 59th annual meeting.

Leydesdorff, L. & Curran, M., (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy, Cybermetrics, 4. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Meadows, A. J. (1998). Communicating research, Academic Press.

Melin, G. & Persson, O. (1996). Studying research collaboration using co-authorships, Scientometrics, 36(3), 363-377.

Mettrop, W. & Nieuwenhuysen, P. (2001). Internet search engines – fluctuations in document accessibility, Journal of Documentation, 57(5), 623-651.

O'Leary, J. (2001). The Times good university guide 2002, London: HarperCollins.

Rehm, G. (2002). Towards Automatic Web Genre Identification - A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. In: Proceedings of the Hawaii International Conference on System Sciences, January 7-10, 2002, Big Island, Hawaii.

Rousseau, R., (1997). Sitations: an exploratory study, *Cybermetrics*, 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R., (1999). Daily time series of common single word searches in AltaVista and NorthernLight, Cybermetrics, 2/3. Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Smith, A. G. (1999). A tale of two web spaces: comparing sites using Web Impact Factors. Journal of Documentation, 55(5), 577-592.

Smith, A. & Thelwall, M. (2002, to appear). Web Impact Factors for Australasian Universities, Scientometrics, 54(1-2).

Tang, R. & Thelwall, M. (2002). Exploring the pattern of links between Chinese university Web sites, University at Albany, SUNY.

Thelwall, M. (2000), Web Impact Factors and Search Engine Coverage, Journal of Documentation, 56(2), 185-189.

Thelwall, M. (2001a). Extracting macroscopic information from web links. Journal of the American Society for Information Science and Technology, 52 (13), 1157-1168.

Thelwall, M. (2001b). Evidence for the existence of geographic trends in university web site interlinking, University of Wolverhampton.

Thelwall, M. (2001c). A web crawler design for data mining, Journal of Information Science, 27(5), 319-325.

Thelwall, M. (2001d). A publicly accessible database of UK university website links and a discussion of the need for human intervention in web crawling, University of Wolverhampton.

Thelwall, M. (2001e). A Research and Institutional Size Based Model for National University Web Site Interlinking, University of Wolverhampton.

Thelwall, M. (2001f). Results from a Web Impact Factor crawler, Journal of Documentation, 57(2), 177-191.

Thelwall, M. (2002a, to appear). Methodologies for Crawler Based Web Surveys, Journal of Internet Research, 12.

Thelwall, M. (2002b). A comparison of sources of Links for academic Web Impact Factor calculations. Journal of Documentation, 58, 60-72.

Wikgren, M. (2001). Health discussions on the Internet: A study of knowledge communication through citations, Library & Information Science Research, 23(4), 305-318.