Extracting Macroscopic Information from Web Links[1]

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

Much has been written about the potential and pitfalls of macroscopic web-based link analysis, yet there have been no studies that have provided clear statistical evidence that any of the proposed calculations can produce results over large areas of the web that correlate with phenomena external to the Internet. This article attempts to provide such evidence through an evaluation of Ingwersen's (1998) proposed external Web Impact Factor (WIF) for the original use of the web: the interlinking of academic research. In particular, it studies the case of the relationship between academic hyperlinks and research activity for universities in Britain, a country chosen for its variety of institutions and the existence of an official government rating exercise for research. After reviewing the numerous reasons why link counts may be unreliable, it demonstrates that four different WIFs do, in fact, correlate with the conventional academic research measures. The WIF delivering the greatest correlation with research rankings was the ratio of web pages with links pointing at research-based pages to faculty numbers. The scarcity of links to electronic academic papers in the data set suggests that, in contrast to citation analysis, this WIF is measuring the reputations of universities and their scholars, rather than the quality of their publications.

Introduction

Between Citations and Backlinks

It has long been known that citations provide a useful, if problematic, source of information about journals and authors. A similar web-based phenomenon has also attracted much attention in the past five years, that of pages containing a link to the one studied, sometimes called 'backlinks' or 'sitations'. It has been pointed out that web data is inherently more unreliable and technically problematical than citations (Ingwersen, 1998; Snyder & Rosenbaum, 1999; Smith, 1999; Davenport, & Cronin, 2000; Egghe, 2000; Thelwall, 2000; Cronin, 2001a; Bar-Ilan, 2001; Björneborn & Ingwersen, 2001; Thelwall, 2001a). Such analyses have, nevertheless, been conducted on e-journals and, indeed, initiatives to add hyperlinks to references in papers originally published in print may make this much more widespread (Harnad & Carr, 2000). There has also been an attempt to use both citations and backlinks to measure the research of entire academic departments (Thomas & Willet, 2000). Web studies also have ranged further afield, producing calculations for universities, countries and top level Internet domains. An unresolved question, however, is whether any web link

[1] *This is a preprint of an article published in the Journal of the American Society for Information Science and Technology Vol 52 No. 13, 1157-1168 © copyright 2001 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

calculations can be shown to correlate with other aspects of human activity that are worth measuring, such as the production of research, or whether their utility is restricted to developing knowledge of the web itself. Much of the web link research so far has been experimental, developing the techniques and measuring web, with claims that the web or search tools may not yet be developed enough to produce reliable results (Smith, 1999; Thelwall, 2000; Björneborn & Ingwersen, 2001; Thelwall, 2001a). This paper seeks to demonstrate that associations can be made with non-electronic phenomena by providing statistical evidence that the results can correlate with an accepted measure of the activity in question. The specific hypotheses tested are that four link-based calculations (described in detail below) for British universities produce results that correlate with figures derived from the official government research assessment exercise. Because of the known problems with search engine reliability, three of these will be derived from a specialised crawler, and will be compared to one from a commercial search site. Tests will also be conducted to assess the significance of the differences between the results of the four metrics.

Validity Issues: What can be Inferred from the Existence of Links?

Is there any reason to believe that an exercise based upon counting the number of backlinks to pages on a university web site should reflect a facet of the university's research? An indiscriminate count of backlinks certainly does not *measure* any aspect of research because of the variety of uses for university web sites, only one of which is the publication of research-related information (Middleton *et al*., 1999). Would it make a difference if only backlinks to research-related pages were counted? It is clearly not possible to claim that all or the majority of research output can be measured in this way since the publication of research in public areas of university web sites is far from being the norm. Aside from technical issues about the reliability of the counting process, the motivations of a researcher in creating a link in their web site to another research-related page are likely to be at least as diverse as those for e-journal citations. These are, in turn, more diverse than those for print journal citations (Kim, 2000). In order to ascertain precisely what is being measured, it would be necessary to conduct a thorough investigation of motivation for the creation of research-related hyperlinks, including the development of a precise definition of this concept. In the context of the more mature area of patent citation analysis, Oppenheim (2000) has argued that, "one should be very cautious about drawing conclusions from [it]" and that its validity could only be fully assessed after a series of ten research questions had been addressed. Five of Oppenheim's questions concern the correlation of patent analysis results with non-patent phenomena. One of the correlation hypotheses in this paper will be for a measure based upon a (crude definition of) research related backlinks, and this will be, therefore, a contribution to the understanding of the validity of using research-related backlinks to assess an aspect of research. The remaining three metrics use data that has not been filtered in any way for content and, therefore, cannot be assessing research in any sense. The study of web links between university web sites is at an early stage, but based upon the variety of content to be expected (Middleton *et al*., 1999) and suggestions of multiple underlying trends in the patterns of links between individual institutions (Thelwall, 2001d), the answer is unlikely to be a simple one. These metrics will also be compared with research ratings, both in the belief that positive results may lead the way to practical applications, and as a step towards developing a model for that which counts of backlinks for university web sites represent.

The Web Impact Factor Metric

In order to analyse link counts, an appropriate metric must first be devised. In a common form of the Impact Factor for journals, ratings are based upon the number of citations in a selected group of other journals during a specified time period following publication (see Garfield (1994) for example). An identical calculation can be evaluated for electronic journals, but also a similar one for any defined set of areas of the web. Ingwersen (1998) proposed a link-based calculation, the Web Impact Factor (WIF). He defined the external WIF to be the total number of pages external to an area of the web with links pointing into it, divided by the total number of pages in the chosen area. This gives a measure of average external impact per page, which could be for a single university web site or all web sites in an entire country, for example. More generally, versions of the external WIF can also be calculated by only counting link pages from a specified subset of the web, for example all the set of all academic web sites outside the chosen site. Although the original WIF included link pages inside the site, the *external* WIF will be used throughout this paper because of the large impact that HTML design issues can have on internal links (Smith, 1999; Thelwall, 2001a). WIFs have normally been calculated using the advanced facilities of a search engine, a limiting factor because these cover only a fraction of the web, and one that is biased by the very page links that WIFs use (Lawrence & Giles, 1999). Other studies have also found uneven coverage of the search database and variability the results over time (Ingwersen, 1998; Bar-Ilan, 1999; Rousseau, 1999; Snyder & Rosenbaum, 1999; Thelwall, 2000), although an improvement in the reliability of AltaVista's results over time has recently been identified (Thelwall, 2001b). In order to circumvent this issue, further research used a specially constructed academic web crawler to ensure comprehensive coverage of five universities, with WIFs subsequently calculated (Thelwall, 2001a). The denominators of these calculations were problematic because they were dependent upon the form in which information was stored: a single document would produce an increase in the denominator if it was split into many different pages to improve readability, for example. In response to this, the denominator of the calculations was changed to the number of faculty members in order to provide an alternative measure of the size of an institution. This represents two shifts of focus in the calculation: towards studying the community rather than their artefacts; and towards a hybrid calculation combining web information with another source. The resulting calculations still suffered from numerator problems, particularly that many link pages included in the calculation were not research-based.

The UK was chosen as the base for this study because of the number and range of types of university and the existence of a formal, government backed exercise for assessing universities that considers all aspects of research, the Research Assessment Exercise. Figures derived from the last exercise, in 1996, give a relatively objective and reliable base line with which to compare other metrics.

Related Research and Applications in Computer Science

There is direct research in computer science showing that useful information about individual web pages and web sites can be extracted from link structures (Kleinberg, 1999; Hernandez-Borges *et al*., 1999). A close tie with citation analysis is acknowledged (Chakrabarti *et al*. 1999). Indeed, Kleinberg points out that much general information on the web is *only* available to automated processes via the link

structure, particularly information about which sources of information are most authoritative for a given information need. As an example of this, if a search term such as "Microsoft" matches a large number of web pages, it may be impossible for an algorithm to identify which match is the best from a simple content analysis of the pages. Instead, it may choose to return the most authoritative sources as the best matches, measuring authority in terms of a simple count of the number of other pages linking to the one in question. The search engine Google uses a sophisticated version of this idea to rank pages (Brin & Page, 1998) and, whilst most search engines do not publish their page ranking algorithms, it seems that using link counts in some way to help measure authority is essential to effective results. Recognition of the importance of link structure has lead to the creation of numerous tools to improve the usability of the web, see Chen (1997), McDonald and Stevenson (1998) and Amento *et al*. (1999) for example. Although the Google algorithm is globally based, all of these examples are essentially focussing on links on a *microscopic* level, targeted towards the relevance of the contents of individual web pages, and how to select and present groups of individual web pages to users. Some research into the *macroscopic* link structure of the web has also been undertaken by computer scientists (Gibson *et al.,* 1998; Broder *et al*., 2000), driven by the need to improve the computing tools available to access information on the web.

Citations and Backlinks

Journals and E-journals

There have been several attempts to extend citation analysis to e-journals, a phenomenon now apparently reaching maturity and at least partial acceptance (Fosmire & Yu, 2000). References are known to be made for a diverse set of reasons, not all positive, yet they can yield useful information (Garfield, 1979). The causes for concern over the reliability of traditional citation counts are numerous, but include their use for criticism of previous work (Case & Higgins, 2000) and that the figures are a potential source of manipulation (Gowrishankar *et al*., 1999). With refereed online journals, moreover, new motivations for referencing can be ascertained that have not been observed in traditional journals (Kim, 2000). These include convenience in accessing material and the desire to link to illustrative graphics. With, in addition to this, the relationship between scholarly documents undergoing changes in the digital era (Borgman, 2000), it is not immediately apparent that electronic backlinks or citations will yield similar information. It should be noted, however, that the distinction between electronic and paper publication is becoming increasingly blurred in the era of digital libraries, and with initiatives such as the OpCit project adding hyperlinks to traditional articles, allowing new online informetric analyses (Harnad & Carr, 2000). In fact, web techniques may add significant value to citation studies, seen as under threat in some areas from problems associated with multiple authorship (Cronin, 2001b) if a suggestion like the one of Davenport and Cronin (in press) for the inclusion of detailed authorial information in an XML-like form is adopted.

Smith (1999) and Harter and Ford (2000) have studied backlinks in multiple separately hosted web based e-journals. This refers to journals with their own separate public web site, rather than a set of normally print-based journals published electronically by a publisher. Smith's study was of 22 refereed Australasian e-

journals, three of which had print equivalents, and found that calculations based upon such counts were unreliable. The later Harter and Ford study of 39 very diverse journals arrived at the same result. In both cases there were methodological problems identified by the authors, for example the use of different URLs for some journals, and the reliance upon search engines for counts. Both studies counted all web links found by AltaVista, and so a fundamental difference between these and traditional citation counts was the extensive use of information from unrefereed sources, with Harter and Ford finding scholarly works a 'very small percentage of backlinking files'. Such studies do not prove, however, that web link analyses are necessarily unproductive, but do serve to illustrate the problems. Other methodologies or the passage of time may provide different results. Examples of different approaches that could be tried include a restriction on the domains from which links may be counted, a more uniform selection of journals (if and when there are enough online journals), or the avoidance of reliance upon search engines for site coverage.

Academic Institutions and Departments

University web sites form a promising area to trial the extension of backlink research to non-refereed pages because of the relative maturity of the web in academia. Web link analysis for university web sites, whilst promising in general terms, does suffer from the fact that academic web sites are populated by pages designed for a mixture of purposes and targeted at different audiences (Middleton *et al*., 1999). This makes web links a more complex phenomenon than journal citations, for example. The web is known to contain much information useful to researchers, including online journals, information about research groups, profiles of individual academics, non-refereed articles and various discussion forums. One study, for example, showed that useful information not available in other forms can be extracted by researchers from the web using search engines, at least for informetrics (Bar-Ilan, 2000). This study found that, whilst there was very little actual information content in the pages found, there were many hyperlinks to other web pages on Informetrics and many formal bibliographic references. Although the majority of university web pages are probably not primarily aimed at research, this seems the logical aspect of the role of a university to compare with link counts. Since there are links between universities for research-related reasons, it may be possible to extract research-related information on a macroscopic level from the disorderly environment created by the number of links used for other purposes. The question, then, is whether counting links between different institutions will correlate in any way with accepted indicators of research quality.

There have been successful attempts in the UK to correlate research ratings with traditional citation counts for four subject areas, including information science (Oppenheim, 1995; Oppenheim, 1997). The definitive ratings for research in UK academic institutions are given by the five-yearly government Research Assessment Exercise (RAE). Ratings are decided upon by a series of subject-specific panels for 69 areas. Many aspects of research are judged with the quality of publications a major consideration. (See Elkin and Law (1997) for the scope of activities covered and a description of the exercise in the library and information science area.) The outcome of this process is a rating on a seven-point scale for each submission. The ratings given determine the destination of billions of dollars of direct government research funding over five years, an indication of the seriousness of the process (HEFCE, 1998). In the past, lower scores have not been rewarded whereas higher scores have

generally attracted progressively increasing funding, although the exact details vary between funding council. The task of assessing the RAE subject areas ('units of assessment' in RAE terminology) is very expensive both for the government and the universities that make submissions. In response to this, a proposal has been made by Oppenheim that an exercise relying mainly, but not exclusively, upon citation counting could prove much cheaper (Oppenheim, 1995; Oppenheim, 1997; Holmes & Oppenheim, 2001). This suggestion has been controversial (Warner, 2000a; Warner, 2000b) but there is agreement that a high citation count would seem to be at least a logical source of additional evidence for the importance of research publications. It has also been claimed that citations in the form of web links would also be a valid supporting, but not primary, indicator of the research regard of a department (Holmes & Oppenheim, 2001).

There have been several backlink studies of academic institutions and departments, some mirroring citation studies, but none have produced statistically verifiable conclusions. Chen *et al.* (1998) counted links between computer science department web sites in Scottish universities. The numbers produced did clearly reflect the profiles of the institutions concerned, but the survey was limited by the small sample size and a lack of variety in the institutions concerned. Ingwersen (1998) calculated institutional WIFs to assess the reliability of the results. Smith (1999) found the WIF a 'useful measure of the overall influence of the web space' for universities and research institutions. The recent study of Thomas and Willet (2000) was unable to find a significant correlation between backlink data and RAE rankings for library and information science departments, concluding that it was 'premature to use the [data] for evaluating the research performance of individual academic departments'. Finally, Thelwall (2001a) proposed methodological developments of the WIF from coverage of a small sample of UK universities.

Other Areas

WIF and other link-based calculations have also been applied to non-educational areas of the web, delivering interesting results and developing the methodologies, but not attempting to confirm results with respect to external measurements. Rousseau (1997) analysed the distribution of link pages across top level domains. Almind and Ingwersen (1998) used a range of calculations to discover that Denmark was relatively less visible on the web than other Nordic countries. Ingwersen's (1998) paper that developed Web Impact Factors confirmed the previous joint findings with Almind. Snyder and Rosenbaum (1999) calculated link page totals between the global Internet top-level domains, showing the unreliability of search engine results. Leydesdorff and Curran (2000) charted the progress between 1993 and 1998 of web pages in AltaVista's database relating to government, academic and commercial sectors of the web in the Netherlands, Brazil and in the global top-level domains. This study used various advanced search commands, including one for link counting, but did not attempt to compare the results with external sources of evidence about those relations.

Methodology

The British academic web was surveyed in order to obtain statistics about web pages that were the target of links from web pages on other British academic web sites. The targets of these links were then classified by the information that they

contained. These figures were then used to compile a selection of WIFs for each information type and institution. Each British university has a web site with a home page accessible through at least one domain of the form www.name.ac.uk, where 'name' is an abbreviation of the university name, for example www.man.ac.uk for Manchester University. The university sites were indexed by a crawler designed for accurate and comprehensive site coverage, including the identification of duplicate pages, where possible. Appendix one contains further information that will be of use to those wishing to replicate the experiment and Thelwall (2001c) describes the architecture of the crawler used. From the database created by the crawler were extracted lists of all pages in each institution that were linked to by at least one indexable page in another British university, together with a count for the number of such pages linking to them.

The Classification of Page Types

Each page that was the target of at least one link was classified by the author according to the type of information that it contained. The classification was initially into the groups of categories suggested by the literature on academic web site design (Middleton *et al*., 1999). This classification was then subdivided further, particularly in the area of research, to give finer grained detail of the actual content. The process of deciding upon the final classification was iterative: when pages were found that were troublesome to classify, the classification was revised and any potentially problematic pages previously classified were re-examined. Pages could, however, fall into more than one category, for example research into teaching pages could be classified as research or teaching-related material, or information pages that were the product of research groups. In these cases a subjective judgement had to be made as to the most appropriate category. At the end of this process a summary was compiled for each university of the total number of external British academic links pointing to pages of each classification type. Each site contained up to 2,311 web pages identified by the link study and, therefore, it was extremely time consuming to evaluate all of these pages. It was also necessary that they were all checked in a relatively short period of time to avoid creating bias due to links disappearing as the web pages were taken down or moved to a different location, and so a sample of 25 of the universities was used. Many pages had, in fact, disappeared already at the time of the study, and these were classified based upon the URL, if possible, otherwise classified as general information pages. The classification exercise took place between July and September, 2000.

It is fundamentally impossible to determine the reason for the link based upon the target page only. For example, a departmental home page could be cited as the source of good research information, an example of good or bad web design, or as the previous home of the source document author. It was decided that pages would be classified based upon their content alone, and not conditionally upon the context of the links, again a practical step. Some general rules were decided upon to aid the classification process, but for many pages it was still a subjective decision. Although many categories were used, the key distinction was between pages giving information about research conducted in the institution and those that did not. Home pages of faculty members were counted as research-related if they gave any information about the research profile of the person. Web pages that were created by research groups were normally counted, the exceptions being when they were pages of links to other sites, or were designed for teaching purposes. Pages with the primary purpose of

carrying links to external sites were not included in the figures for research-related material, even if maintained by a researcher or research group, because their function was not to describe indigenous research. It could be argued that such a page of links that attempted to be in some sense a definitive collection of links to high quality relevant artefacts elsewhere on the Internet was an original contribution to research and should be recognised, whereas an unfiltered one should perhaps not. It was, nevertheless, judged to be impractical to be able to make this distinction in the study. Home pages of departments were also classified as research related because of their discipline-based nature, but university home pages were not. It has been observed previously that many web pages are difficult to classify as being of a particular type and some could be seen as a built from a combination of basic types (Haas & Grams, 2000), but in this survey pages were only classified according to what appeared to be the dominant type. Electronic journal pages were not counted as research-related because they typically host all articles on a single server, and so most would be hosted away from the author's university web site.

Geographic and Federal Considerations

The structure of British higher education web sites is expected to result in a number of regional biases. An example of this is that there have been a number of regionally based initiatives that have included using the Internet to share teaching resources between neighbouring institutions (Thelwall, 1999). One result of this may well be increased interlinking between close universities. There are, however, some more formal links between institutions in London and in Wales, encapsulated in the federal structures known as the University of London and the University of Wales. To avoid the possibility that these will gain extra links from such a relationship, universities affiliated with these were not included in the second part of the study.

Results

The results of the survey were used to calculate WIFs, which were then tested for a significant relationship with a measure of institutional research quality through the well-known Pearson's correlation coefficient. It would also be useful here to assess whether the various correlations calculated are significantly different from each other. Care is needed here because if one data set correlates at a higher significance level than another, this does not then prove that it carries a stronger association. A separate (somewhat rare) test is needed to decide whether the difference in correlation coefficients is significant (Steiger, 1980).

The General WIF

The first WIF calculated was a general WIF, using as numerator a simple count of all links to the site, irrespective of whether the target page was research related or not. The denominator used was the number of full-time equivalent faculty members (Noble, 1999). The first graph shows for each of the 25 chosen institutions the general WIF from other British universities plotted against a research rating for the university derived from the official 1996 rating exercise (Mayfield University Consultants, 2000). The exact figures have been obtained by averaging the ratings of the different departments in an institution. Figure 1 shows a strong linear trend with a highly

significant correlation coefficient of 0.80, much higher than the critical value for this test of 0.61 for p = 0.001. The two largest anomalies on the graph are Wolverhampton (RAE Average: 0.5; WIF: 1.79) and Warwick (RAE Average: 5.4; WIF: 3.50). The high Wolverhampton factor is due to the very large number of links pointing to a single resource, the UK academic clickable map. Cambridge University also includes one dominant link to a web page counter resource, but this is less visible due to the large number of links to other pages in the site. Warwick's relatively high performance was due, in part, to its hosting of electronic journals, which accounted for 11% of links. Without this, Warwick's result would have been 3.12, although this is still a high score.

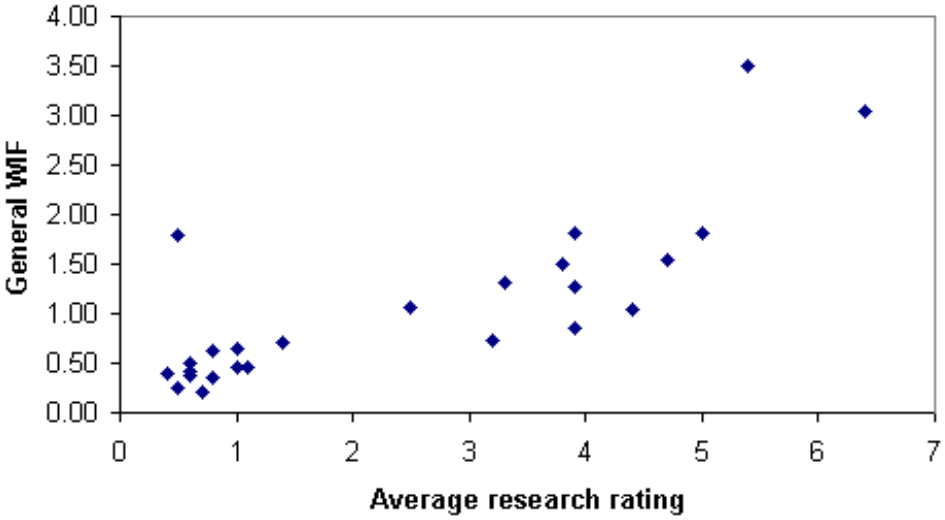

FIG. 1. General WIF plotted against research rating for 25 British universities

The Research WIF

The second set of calculations was for a research WIF, with numerator containing counts of links to pages judged to be research related. The second graph shows a plot of this against average research rating. This data also has a highly significant correlation coefficient, which is larger at 0.90, but the difference between the two correlations is not quite large enough to be statistically significant at the 5% level. If the non-research related links are separated from the research-related links, then they still correlate strongly with research ratings (r = 0.70), but this is a significantly lower value than the research based rating, at the 5% level. On this graph there are also anomalies, with Liverpool University being a low point at (RAE Average: 3.8; WIF: 0.19) and Aston high at (RAE Average: 3.3; WIF: 0.62). The reason for Liverpool's weak showing is possibly related to the fact that its main web site has used the official method to ban many crawlers from its site, although the main search engine crawlers, such as AltaVista's are allowed in. Liverpool does not fare as poorly in the general WIF calculation because of a longstanding chemistry links site with a huge number of links to it. Aston's site has a relatively high WIF, which comes mostly from a large number of links to a single research group. The relatively small size of Aston, in terms of faculty numbers, has allowed one group to exert a large influence on its result.

It should be noted that very few of these pages contained significant research content, echoing the findings of Bar-Ilan (2000) and reinforcing the conclusions of Kling and McKim (1999) that publication in established journals is more favourable for authors than web publication. An author could, however, post a paper on their own university web server as well as in an electronic or print journal. The lack of evidence for such dual publication may be due to the existence of factors mitigating against it both before and after the submission and acceptance of a paper in a journal. The belief that prior electronic publication may prejudice editors against acceptance is a disincentive to pre-publish on the web (Harter & Taemin, 2000). After publication, copyright restrictions can also stop a paper from being published on the author's university web site. In the context of the papers found in this study, none of which were in e-journals (which were excluded as described above), the only advantages of 'publication' on the author's university web server instead of in a refereed journal would be ease of access and the availability of web resources, such as multimedia. The lack of trustworthiness, in terms of peer review, and publicity, in terms of being formally announced to scholars, would clearly be major disadvantages.

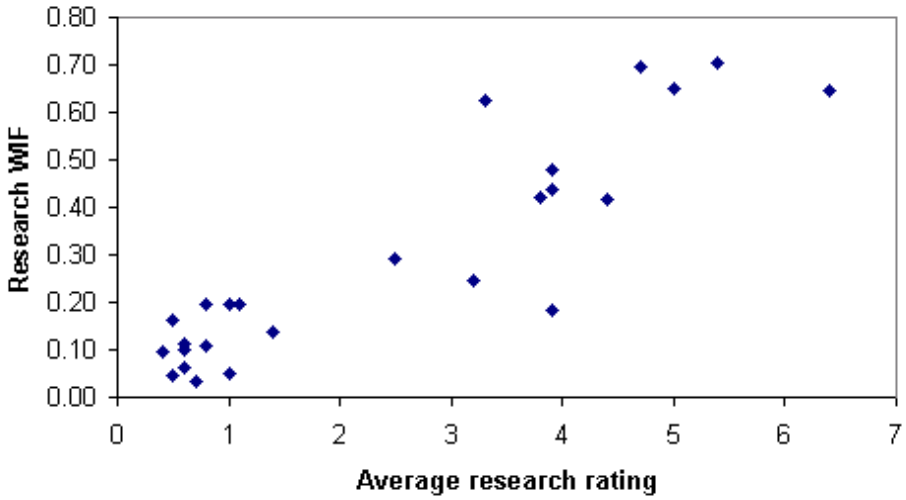

FIG. 2. Research WIF plotted against research rating for 25 British universities

In order to validate the classification of the web pages, a second checker was given a list of a random selection of pages from the sites surveyed together with a description of the categories, and asked to classify them without seeing the original results. The results were then compared with the original classification. Of the 110 pages that were accessible, only 46 received an identical classification, indicating real problems with the detailed classification process. However, nearly all of the pages, 106, were classified by both researchers on the same side of the binary divide between research related and non-research related pages. This gives some confidence in the freedom from individual arbitrariness of the main results from this paper. Due to the unproven reliability of the individual subcategories employed, these results will not be discussed except to record that all correlated positively with academic rating. This was even true of the category of non-academic links, which consisted mainly of student society pages, but also included religious pages, hobby pages and personal pages without any academic content.

Comparison with AltaVista

Using Search Engines for Raw Data

The results from the almost standard publicly indexable crawls described above will be compared to ones obtained by using AltaVista's advanced query syntax, the method used by two previous papers (Ingwersen, 1998; Smith, 1999). A series of queries were used to count the number of indexed pages in the ac.uk domain with links to a given domain. The logical form of the query is shown here for one university, Sussex, which owns domain names that end in both susx.ac.uk and sussex.ac.uk, for example www.sussex.ac.uk and www.cogs.susx.ac.uk.

> host:ac.uk AND (link:sussex.ac.uk OR link:susx.ac.uk) AND NOT (host:sussex.ac.uk OR host:susx.ac.uk).

This survey is thus similar to one previously reported (Kelly, 2000) which used both AltaVista and Infoseek to count external links pointing to UK higher education institutions, but did not perform calculations, analyse the data or take into account alternative domain names. The request shown is for the number of web pages in the UK academic domain that contain a link to a page with a domain name ending in either of the standard Sussex endings, but excludes pages that themselves have this ending. In practice, because it is known that queries can 'time out' (Smith, 1999), these queries were split into a logically disjoint set and the results totalled. The figures resulting from this calculation were in principle similar to the general WIF calculation described above, but did have a few important differences.

- The database is AltaVista's, and, since its design parameters are a trade secret, the crawling method is an unknown element.
- The count is of pages rather than links and so a lower overall total would result from pages with links to different pages in the same university.
- The domain chosen is the whole ac.uk domain, including non-university organisations such as further and higher education colleges, education companies, education and research websites and organisations.
- AltaVista includes links to resources other than web pages, for example clickable email links.

The AltaVista general WIF

The AltaVista general WIF calculated with these figures for numerators produced a correlation of 0.78 with research ratings, not a statistically significant difference with the general WIF calculated in the previous section. This was, however, a statistically significant difference at the 5% level between this correlation and the value of 0.90 for the research WIF.

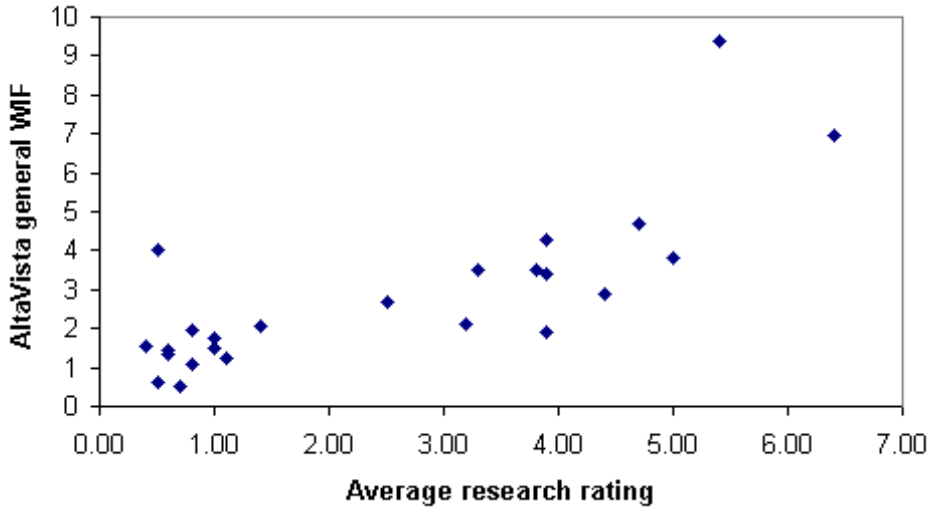

FIG. 3. AltaVista general WIF plotted against research rating for 25 British universities

Although the correlations matched, the individual results did vary in proportion between institutions. In all cases AltaVista found more link pages than the publicly indexable crawl found links, with the proportion of links varying from 26% (Paisley) to 44% (Cambridge). It is known that search engines cannot index the entire web and must be selective about which pages to index (Brin & Page, 1998; Lawrence & Giles, 1999) and so the fact that AltaVista found over twice as many link pages as the crawl found links was unexpected. A further analysis was undertaken to uncover the root cause of this phenomenon.

A Comparison of the Link Pages found by AltaVista and the Publicly Indexable Crawl

The number of pages indexed on sites by AltaVista was compared with the crawl totals and it was found that AltaVista indexed a small percentage more, 11%. The results differed greatly between sites, with AltaVista indexing 55% less pages for Lampeter and 21% more for Glasgow. There were also some extreme results such as Liverpool, which had used the robots exclusion protocol to allow AltaVista to index its site but exclude research crawlers. In this case AltaVista claimed 66,780 pages compared to 0. In order to probe further the reasons for the link count discrepancies, individual sets of links were checked. One set of links is described here to illustrate the method and results.

Table 1 shows a breakdown of the discrepancy between the two methods in pages found on the Glasgow University site that point to pages at Wolverhampton University. This count was obtained by the following AltaVista query.

> (link:wolverhampton.ac.uk OR link:wlv.ac.uk) AND (host:gla.ac.uk OR host:glasgow.ac.uk)

Results from this were compared with pages extracted from the crawl database. AltaVista found many more link pages: 60 compared to 27, with 21 of the backlink pages being common to both methods. Of the five unique pages found by the crawler but not AltaVista, one no longer carried the link to Wolverhampton, making it possible that AltaVista had recently visited it. One of the other four had disappeared

since the crawl, so a similar explanation is possible here. Six pages retrieved by AltaVista and one by the crawler were identical copies of pages at different URLs. A search engine normally attempts to identify and eliminate duplication in its index but for speed this process is expected to be less than 100% efficient (Heydon & Najork, 1999). Of the remaining pages, four had gone and one contained only an email address link, leaving 28 unaccounted for. The AltaVista link: advanced search command was used to find the indexed pages that linked to these, for example

link:www2.gla.ac.uk/

This showed that all of these pages were either not linked to by any other Glasgow University web pages or were only linked to by pages that were not linked to by the rest of the web site. This confirmed that they were not in the publicly indexable set. Of these, six were linked to pages elsewhere on the web but 17 were not. AltaVista must, therefore, have learned about the existence of these pages from a source other than a recent crawl of the web, or, as seems less likely, has discarded pages that link to these. A clue was found in the number of old web server addresses in the links. Glasgow University, like many other universities, has restructured its domain names, but has left the old versions active for existing links. This means that AltaVista may well have a memory for old known URLs and so will find pages that have been linked to in the past. A search engine, when recrawling a site, can either crawl it from scratch or refresh its existing database by rechecking known pages (Choo & Garcia-Molina, 2000). The evidence from this small sample is that AltaVista either uses the second approach or, if it uses the first, seeds it with a list of known URLs. There are two other possible sources of new URLs: external links and owner registration. In either case the pages have received extra human effort to publicise them and may possibly be more likely to contain external links, which could explain why AltaVista has given so many more links. The authors of ten of the seventeen anomalous pages were identified and emailed to query the results. Two replies were received, both stating that their pages had not been registered by them in AltaVista, but had been linked to by other pages in the past. One of the pages was, in fact, linked to by a page indexed by AltaVista and the page was revisited by AltaVista during the study, but the link was not registered by it. This appears to be an anomaly in the indexing software.

	Found by crawler	**Found by AltaVista**
Pages appearing more than once in the index with identical HTML but different URLs	1	6
Pages gone at the time of checking and only found by one crawler	1	4
Pages containing an email link to Wolverhampton University but not a standard link	0	1
Pages linked to by pages on external sites, but not by indexed pages on the Glasgow University site	0	6
Pages linked to by indexed pages on the Glasgow University site.	4	5
Pages without any indexed pages at all linking to them	0	17
Total pages reported by one crawler but not the other	**6**	**39**

Table 1. Pages at Glasgow University containing a link to Wolverhampton University that were found by only one of the two methods

The Reliability of AltaVista

During the extensive use of AltaVista in the analysis, it was found to be, in general, consistent and reliable in its results. This is in agreement with a recent survey (Thelwall, 2001b), but in complete contrast to the findings of Snyder and Rosenbaum (1999) who found its performance to be unreliable and demonstrably incorrect. They cited the search 'host:osu AND link:edu' as retrieving only four pages out of a site of 1,408, despite the first 20 pages indexed by AltaVista containing a link to an edu host. This query was repeated in order to study the cause of the different conclusions, and the same command returned 'About 2,364' pages out of 'About 5,494' indexed, a qualitatively different result. This was compared with a search for the logical opposite 'host:osu.edu AND NOT link:edu', which returned 'About 3,133', a very small discrepancy of only 3. The first 20 links on the latter search were tested and it was discovered that most of them did contain links to an edu site, but that these links did not appear in the HTML of the page. This is possible because all the links in question were relative links most giving just the page name of the page linked to, a possibility identified by Smith (1999). Relative links are allowed in HTML when the target page is in the same location (often, but not always, the same directory or folder) as the source page. It is clear, then, that AltaVista counts links that are explicitly in the text of the page, absolute links. This means that its results for *internal* links on any web site will be greatly unreliable, but external link counts are not effected since they cannot use relative links. It must also be concluded that AltaVista has improved its algorithm in the new version launched (Rousseau, 1999) since the research of Snyder and Rosenbaum, although the general concerns that they raised about the opaqueness of commercial search engine retrieval processes are still valid.

The AltaVista Original General WIF

One final set of Impact Factors was calculated, using the same numerators as the previous calculations but with the page counts from AltaVista for denominators, as originally proposed by Ingwersen. These AltaVista original general WIFs correlated strongly with research ratings, but with a much lower correlation coefficient of 0.57 (significant at the 1% level for correlations since above the critical value of $r = 0.51$ for this test). The difference between this correlation value and the higher crawler results correlation was, however, not quite significant at the 5% level. This result means that although the results are suggestive of the improvement in the calculation from the use of faculty numbers for the WIF denominator, they do not provide statistical evidence of it.

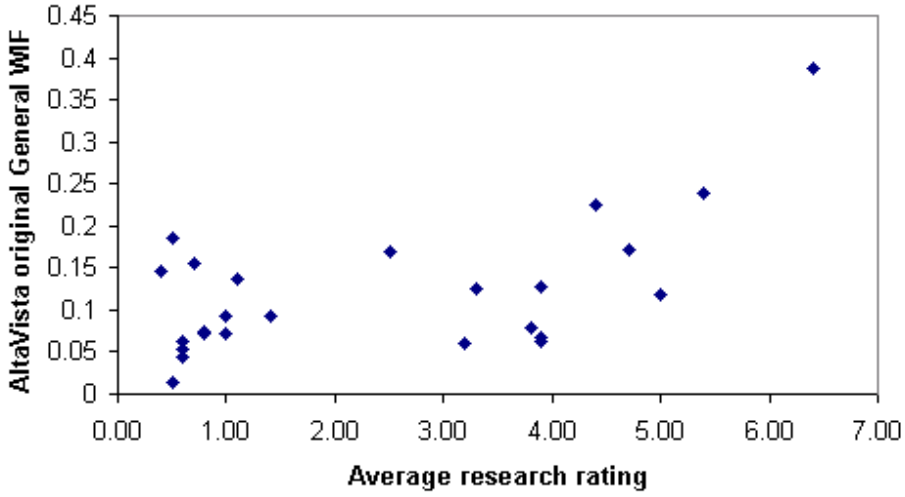

FIG. 4. AltaVista original general WIF plotted against research rating for 25 British universities

Discussion

The use of Search Engines for Web Link Analysis

The comparison of results with those from AltaVista suggest that it would give similar results to a publicly indexable crawl in the undifferentiated case of all web pages. For undifferentiated searches AltaVista has, therefore, demonstrated a useful capability. The real problem with this for research is that the underlying software is subject to minor and major changes without notice. The results of this survey do not therefore give a licence to use it as a replacement for an academic crawler for web indexing research without using methodological safeguards. Rousseau (1999) proposed one methodology for coping with variations in results over time, although this seems less necessary for AltaVista now (Thelwall, 1001b).

What do WIFs measure?

The high correlation between the research WIF and RAE scores gives credence to the notion that some aspect of research is being measured by it. The detailed categorisation of the pages classified as research-related shows that what is being measured is far more complex than print or e-journal citations. In particular, the near absence of links to identifiable research publications meant that the 'research' backlink counts were dominated by pages with more general intentions than citing the findings of others. Given the number of links to departmental home pages, research group pages, and the home pages of individual researchers it may be the case that one aspect of what it being measured is the informal electronic artefacts of invisible colleges. In this context the survey of Cronin *et al*. (1998) that categorised types of web pages which included professors' names gave a large variety of types and subtypes, one of these being a category for the home page of another person or organisation. The existence of scholarly digital communities has not gone unnoticed (Cronin & McKim, 1996) but many of the types of communication used would not be recognised by the methodology used here. For example items posted on a mailing list service would not 'count', even if postings were publicly available as web pages,

because they would not be on the server of the university of the person that wrote them. This is an echo of the problem that caused e-journal articles to be excluded from the calculation.

A fundamental difference between citation analyses and research backlink counts (as defined here) is that the former are based directly upon the works of an author (even if author reputation has been a motivating factor for including a citation), whereas the latter is based predominantly upon authors, research groups and departments. It may be, then, that backlink counts for individual authors would be based more upon their reputations as individuals than their oeuvres. The inclusion of non-research link source pages, such as teaching material also provides a potential avenue for measuring the absorption of a scholar's work from the frontier to the core of their subject, an important development to be able to record, but one not easily reflected in citation analysis. It would be interesting to test the reputation hypothesis by comparing counts for researchers with similar profiles and citation counts, but differing external indicators of recognition, such as journal editorial board memberships. Extending the argument to university research WIFs, these appear to be measuring the web projection of the reputation of their departments, research groups and individual faculty members.

For the other three hypotheses, the positive results indicate either that an aspect of research is visible through the noise of non-research backlinks or that research and backlinks are associated for some other reason, although not necessarily in a causal way. The significant correlation between RAE averages and the non-research related WIFs suggests that the latter is indeed an ingredient. The combination of types of pages found that were the target of links indicates that, in addition to the research factor discussed above, the utility of the information provided by the university is also being measured. This information would be of a wide variety of types, perhaps a list of all university web site addresses on one page, or lecture notes on another. Other components are also likely to be present, however. At the most general level, some are perhaps just an expression of commonality, such as the inclusion of a link to someone's page about: their favourite soccer team; their religious beliefs; their hobby. Ingwersen's "Web Impact" seems an apposite term for the entity that a general WIF measures for a university.

What can WIFs be used for?

The most that can be claimed from the results demonstrated here is that an estimate of the research ability of a UK university based upon any of the WIF calculations would be probabilistically better than assuming that they are all the same, despite the confounding factor of the unevenness of research quality across an institution. The research WIF would, theoretically and in terms of results, be the best choice for association with research, but is the least practical of the four to be used because it involves the manual classification of pages. In the future of the web, however, this may be a simple task, and forms of research WIF may even replace citation analysis if there is a further academic publishing paradigm shift, as suggested by Berners-Lee and Hendler, (2001). For the UK, using WIFs at the moment to substitute for RAE scores is plainly ridiculous. But this country has been claimed to be exceptional. "Under the Thatcher Government, British academics became one of the most assessed groups in the world" (Anderson, 1991). Yet there is a need for some to assess the scholarly potential of others operating in different environments. Higher education in Europe, for example, is far from simple, with there being more systems

than countries (Knudsen *et al*., 1999). Yet, with the funding driven promotion of interdisciplinary collaboration between multiple different European Union associated countries (Europa, 2001), European researchers can find themselves looking for partners in unknown subjects from institutions in unfamiliar countries. For example, a consortium of computer scientists from Western Europe may find themselves trying to identify a sociologist from Estonia, say, to perform the role of assessing the social significance of their project, perhaps struggling to ascertain whether there is significant status difference between an Ülikool and a Rakenduskörgkool. Whether there is any utility in the relatively crude estimates given by the easily calculated general WIFs, for example, depends upon whether the importance of the information justifies the extra time spent finding more reliable sources, or whether they can provide a useful starting point for such further investigations. In this context, it is not essential that general WIFs do not actually measure research, as long as they correlate significantly with it, a point controversially made by Oppenheim for citation counting in the rather more important context of university research funding assessment (Warner, 2000a).

On a macroscopic level, backlink analysis can be better used to compare groups of organisations. In a country like Britain with an established research assessment exercise, it would be possible to analyse the data for evidence that a class of universities were not being discriminated against. As a hypothetical example of this, if it were to be found that the new universities followed a different relationship between research rating and WIF than the old, then this may be a cause for concern. Further analysis might, however, reveal an innocent explanation, such as a different distribution of computing departments with subsequent heavier use of the web. More generally, the association established here between backlinks and research ratings for one country opens the door to attempts at exploratory analysis of research relationships between groups of academic institutions based upon link structures. The situation for backlinks is different in several important respects to citations, however, which should make those using it much more cautious in their interpretation of the data.

- Unless the data is from e-journals or electronic versions of print journals, the quality and reliability of the data is likely to be lower.
- Technical issues, including reliability of information retrieval tools, naming for documents and multiple copies of pages or collections of pages can skew the results.
- For university web sites, a problem for general WIFs is not whether the 'valid' data is 'corrupted' by a few links that have not been created based upon an assessment of the research value of the target pages, but whether a pattern can be distinguished from a mass of general links.
- Results for individual institutions are very unreliable, particularly for small ones, which could have results dominated by a single high profile group.

It should also be pointed out that should backlink analyses be used for studies that have a real impact upon universities, they would be extremely vulnerable to manipulation. This stems from the unrefereed nature of the medium and from the ease by which large numbers of web pages can be generated automatically by those with programming skills.

Conclusion: Extracting Macroscopic Information from Web Links

Smith (1999) found for Australasian universities no significant correlation between research output, measured in terms of publication counts, and the equivalent of the AltaVista original general WIF calculated above. He concluded that 'at this stage in the web's development, an institution's web output is significantly different in character from its research output.' The correlation found here, even for the same WIF calculation, could be due to the different set of universities used or a more reliable research output indicator to correlate with, but it is also possible that this is a reflection of the maturing of academic web use over time.

Web link metric, then, can yield results that correlate highly with non-electronic phenomena, despite all of the acknowledged problems. Specifically, an association between research ratings and university web site WIFs has been demonstrated, with research WIFs appearing to be measuring the reputation of a university and its faculty rather than the quality of their output. It is suggested that a combination of web based measures with page content parsing can be even more powerful than simple domain-based page counts, as can combinations with non-web based information. It is worth emphasising that on a microscopic level the results are still unreliable, a finding also controversially claimed by Warner (2000a) for the use of citation counts to measure departmental research ratings. These findings open the door to further studies of other, newer areas of the web and for longitudinal studies to chart the changing nature of the way in which communities use the Internet. The same conclusion can now be made for backlink analysis as has been made for traditional citation analysis: that those who rely upon it risk making serious errors, but those who fail to use it 'may be bypassing a valuable source of information' (Biddle, 1996).

Acknowledgements

The quality, scope and accuracy of this paper was substantially improved by the detailed and insightful comments of the reviewers.

References

Almind, T. C. & Ingwersen, P. (1998). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4) 404-426.

Amento, B., Hil, W., Terveen, L., Hix, D. & Ju, P. (1999). An empirical evaluation of User Interfaces for Topic Management of Web Sites. *CHI 99 Conference Proceedings*, pp. 552-559, Addison Wesley, New York.

Anderson, A. (1991). No citation analyses please, we're British. *Science*, 252, 639.

Bar-Ilan, J. (1999). Search Engine Results over Time - A Case Study on Search Engine Stability. *Cybermetrics*, 2/3. Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

Bar-Ilan, J. (2000) The Web as an information source on Informetrics? A content analysis. *Journal of the American Society for Information Science*, 51(5), 432-443.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes - A review and analysis. *Scientometrics*, 50(1), 7-32.

Berners-Lee, T. & Hendler, J. (2001). Scientific publishing on the 'semantic web'. *Nature*, 410, 1023-1024.

Biddle, J. (1996). A citation analysis of the sources and extent of Wesley Mitchell's reputation. *History of Political Economy*, 28(2), 137-169.

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

Borgman, C. L. (2000). Digital libraries and the continuum of scholarly communication. *Journal of Documentation*, 56(4) 412-430.

Brin, S. and Page, L. (1998). The Anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 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.

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.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S. R., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999). Hypersearching the web. *Scientific American*, June 54-60.

Chen, C., Newman, J., Newnam, R. & Rada, R. (1998). How did university departments interweave the Web: A study of connectivity and underlying factors. *Interacting with computers*, 10, 353-373.

Chen, C. (1997). Structuring and visualising the World-Wide Web with Generalised Similarity Analysis. *Proceedings of the 8th ACM Conference on Hypertext (Hypertext '97)*. April, 1997. Southampton, UK. Available: http://www.brunel.ac.uk/~cssrccc2/papers/ht97.pdf

Choo, J. & Garcia-Molina, H. (2000). The Evolution of the Web and Implications for an Incremental Crawler, *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, pp. 200-209.

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

Cronin, B. (2001b). Hyperauthorship: a postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science & Technology*, 52(7).

Cronin, B. & McKim, G. (1996). Science and scholarship on the world wide web: a North American perspective. *Journal of Documentation*, 52(2), 163-171.

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.

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.

Davenport, E. & Cronin, B. (in press). Who dunnit? Metatags and hyperauthorship *Journal of the American Society for Information Science & Technology*.

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

Elkin, J. & Law, D. (1997). The 1996 Research Assessment Exercise: the Library and Information Management Panel. *Journal of Librarianship and Information Science*, 29(3), 131-141.

Europa, (2001). Fifth framework programme. Available: http://europa.eu.int/comm/research/fp5.html, Accessed 16 February, 2001.

Fosmire, M. & Yu, S. (2000). Free Scholarly Electronic Journals: How Good Are They?. *Issues in Science and Technology Librarianship*, Summer 2000. Available: http://www.library.ucsb.edu/istl/00-summer/refereed.html

Garfield, E. (1979). Citation indexing: its theory and applications in science, technology and the humanities. New York: Wiley Interscience.

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

Gibson, D., Kleinberg, J. & Raghavan, P. (1998). Inferring web communities from link topology. *Hypertext 98: Ninth ACM Conference on Hypertext and Hypermedia*. ACM, New York, NY, USA.

Gowrishankar, J., Divakar, P., Baylis, M., Gravenor, M. & Kao, R. (1999). Sprucing up one's Impact Factor (two letters to the editor). *Nature*, 401, 321-322.

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 for Information Science*, 51(13), 1159-1176.

Harter, S. P. & Taemin, K. P. (2000). Impact of prior electronic publication on manuscript consideration policies of scholarly journals. *Journal of the American Society for Information Science*, 51(10), 940-948.

HEFCE (1998). An introduction to the work of the Higher Education Funding Council for England Available: http://www.hefce.ac.uk/Pubs/HEFCE/1998/98_16.htm.

Hernández-Borges, A. A., Macías-Cervi, P., Gaspar-Guardado, M. A., Torres-Álvarez de Arcaya, M. L., Ruiz-Rabaza, A. & Jiménez-Sosa, A. (1999). Can Examination of WWW Usage Statistics and other Indirect Quality Indicators Distinguish the Relative Quality of Medical Web Sites? *Journal of Medical Internet Research*, 1(1). Available: http://www.jmir.org/1999/1/e1/index.htm

Heydon, A. & Najork, M. (1999). Mercator: A scalable, extensible Web crawler. *World Wide Web*, 2, 219-229.

Holmes, A. and Oppenheim, C. (2001). Use of citation analysis to predict the outcome of the 2001 RAE for Unit of Assessment 61: Library and Information Management. *Information Research*, 6 (2).
Available: http://www.shef.ac.uk/~is/publications/infres/6-2/paper103.html.

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

Kelly, B. (2000). WebWatch: A survey of links to UK university web sites. *Ariadne*, 23. Available: http://www.ariadne.ac.uk/issue23/web-watch/

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.

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

Kling, R. & McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. *Journal of the American Society for Information Science*, 50(10), 890-906.

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. Nature, 400, 107-109.

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

Knudsen, I., Haug, G. and Kirstein, J. (1999). Trends in Learning Structures in Higher Education. Available: http://www.rks.dk/trends1.htm Accessed: 7 March 2001.

Mayfield University Consultants, (2000). League Tables 2000, *The Times Higher Education Supplement*, April 14, II-III.

McDonald, S. & Stevenson, R. J. (1998). Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing and information retrieval in hypertext. *Interacting with Computers*, 10(2), 129-142.

Middleton, I., McConnell, M. & Davidson, G. (1999). Presenting a model for the structure and content of a university World Wide Web site, *Journal of Information Science*, 25(3), 219-227.

Oppenheim, C. (1995). The correlation between citation counts and the 1992 research assessment exercises ratings for British library and information science departments, *Journal of Documentation*, 51, 18-27.

Oppenheim, C. (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology, *Journal of Documentation*, 53, 477-487.

Oppenheim, C. (2000). Do patent citations count? 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, 405-432.

The Noble Publishing Company (1999). *Noble's Higher Education Financial Yearbook 1999*, Noble.

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.

Snyder, H. & Rosenbaum, H. (1999). Can search engines be used for web-link analysis? A critical review. *Journal of Documentation*, 55(4), 375-384.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2), 245-251.

Thelwall, M. (1999). Will MANs and SuperJANET dominate educational technology in the UK?, *International Journal of Educational Technology*, 1(1). Available: http://www.amstat.org/publications/jse/

Thelwall, M. (2000). Web Impact Factors and search engine coverage, *Journal of Documentation*, 56(2), 185-189.

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

Thelwall, M. (2001b). The responsiveness of search engine indexes, *Cybermetrics*, 5(1). Available: http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

Thelwall, M. (2001c, to appear). A Web Crawler Design for Data Mining, *Journal of Information Science*.
Thelwall, M. (2001d). Applying Multivariate Statistical Analysis to University Web Links, University of Wolverhampton.
Thomas, O. and Willet, P. (2000). Webometric analysis of departments of Librarianship and information science. *Journal of Information Science* 26(6), 421-428.
Warner, J. (2000a). A critical review of the application of citation studies to the Research Assessment Exercises, *Journal of Information Science,* 26(6), 453-460. (Includes comment by Oppenheim.)
Warner, J. (2000b). Research assessment and citation analysis, *The Scientist* 14(21), 39. Available http://www.the-scientist.com/yr2000/oct/opin_001030.html
World Wide Web Consortium (1999), Performance, implementation, and design notes. Accessed February 27, 2001. Available: http://www.w3.org/TR/html4/appendix/notes.html#h-B.4.1.1

Appendix 1. The Extraction of the Link Structure of University Web Sites

The web site crawler used in the research started from each home page and followed all links to pages on the same site. In most cases the publicly indexable (Lawrence & Giles, 1999) part of the web site was covered, meaning all pages that could be found from the home page by following links. Some web sites did not have any links on the home page, using instead an HTML form for selecting pages, a feature that is ignored in crawls of publicly indexable pages. For these sites, in order to allow a crawl to take place, an alternative page was found as the starting point for the crawl, usually a page of standard links to the departmental home pages. A page was judged to be on the same site if the same three dot-separated words as the home page, .man.ac.uk in the above example. This allows sub-domains such as lib.man.ac.uk and www.maths.man.ac.uk to be included. The larger universities had over a hundred alternative domain names of this form, often one for each department and one for many individual research groups. Some universities also have non-derivative domain names for separate sites, for example www.mcc.ac.uk for Manchester University's computer centre. These sites, when identified, were also crawled. Sites such as this were found from the appearance of unidentified British academic domain names in the database of links from known sites compiled from earlier crawls. It is almost certain, however, that some secondary sites will have been missed because there were no external links to them or because they were registered in a non-academic domain. For example some universities have registered commercial domain names for industry-related projects or areas of the university site, and there were even some web sites for individual academics, such as stephenhawking.org.uk. These were all ignored because it was not practical to check the identity of the owner of each non-academic web site. The web crawler obeyed the convention of ignoring pages specified in the robots.txt file (World Wide Web Consortium, 1999). This lead to some sites not being crawled at all whilst others were crawled in entirety. Most, however, had a limited number of banned areas. The sites that were not crawled were those where the robots.txt file specified that automatic web crawlers were not allowed to download any web pages at all, or that only a list of named web crawlers were allowed to crawl the site.