Three Target Document Range Metrics for University Web Sites[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
Tel: +44 1902 321470 Fax: +44 1902 321478
David Wilkinson
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: d.wilkinson@wlv.ac.uk
Tel: +44 1902 321452 Fax: +44 1902 321478

Three new metrics are introduced that measure the range of use of a university Web site by its peers through different heuristics for counting links targeted at its pages. All three give results that correlate significantly with the research productivity of the target institution. The directory range model, which is based upon summing the number of distinct directories targeted by each other university, produces the most promising results of any link metric yet. Based upon an analysis of changes between models, it is suggested that range models measure essentially the same quantity as their predecessors but are less susceptible to spurious causes of multiple links and are therefore more robust.

Introduction

Citations between scholarly articles have been used for a wide variety of purposes including assessing the impact of journals (Garfield, 1994, 1998), patterns of specialism in individual fields (Small, 1999), patterns of authorship within a discipline (White, & Griffith, 1982), geographic factors affecting research collaboration (Katz, 1993), research productivity for funding purposes and promotion decisions (Adam, 2002) and identifying gender discrimination (Wenneras & Wold, 1997). It has been suggested that Web links offer the potential for even wider applications (Davenport & Cronin, 2000), for example because their use extends to some artefacts of informal scholarly communication (Björneborn, 2001; Cronin, 2001). Metrics based upon hyperlinks could theoretically capture information about the more hidden aspects of the process of scientific endeavour such as the use of research results in teaching and by the general public. Doubt has been cast on the promise of Web link analysis, however, because of the problems associated with using search engines for raw data (Bar-Ilan, 1999; Rousseau, 1999; Thelwall, 2000b; Mettrop & Nieuwenhuysen, 2001) and other unreliability issues endemic to the Web (Egghe, 2000; Thelwall, 2000b; Bar-Ilan, 2001; Björneborn & Ingwersen, 2001). This has lead to the creation of new tools and methods for Web link mining (Thelwall, 2001a-c, 2002c,e), to which this paper is a further contribution.

Recent work has shown that counts of links to Web sites can correlate significantly with research related factors. This is true for links between universities in the UK (Thelwall, 2001a), Australia (Smith & Thelwall, 2002) and China (Tang & Thelwall, 2002), for counts of links to journal Web sites in librarianship and information science (Vaughan & Hysen, 2002) and in law (Vaughan & Thelwall,

[1] *This is a preprint of an article published in the Journal of the American Society for Information Science and Technology Vol 54 No. 6*, 489-496

2003), and for counts of links to Library schools (Chu *et al*., 2002). Deeper mining of Web link data has unearthed the hidden geographic trend in the UK university Web that closer institutions tend to interlink more extensively (Thelwall, 2002a). International patterns of Web linking have also been described for the UK, Australia and New Zealand (Smith & Thelwall, 2002) for the Nordic countries (Ingwersen, 1998) and for the European Union (Polanco *et al*., 2001). Most work has been based upon the implicit assumption that the appropriate way to count links to Web sites is to count either the number of Web pages containing the type of link under investigation (Larson, 1996; Rousseau, 1997; Ingwersen, 1998; Smith, 1999; Thelwall, 1999; Leydesdorff & Curran, 2000; Thelwall, 2002b) or the number of individual links (Thelwall, 2001a, 2002e). One substantially different approach has been suggested, however, with three alternatives derived by focusing on the concept of the 'Web document' and aggregating Web pages together into larger single units for counting purposes (Thelwall, 2002c). This was achieved with heuristics that collate at the URL directory and domain level as well as for complete universities. The domain and directory models were effective in reducing anomalies in the raw link count data that were caused by multiple pages linking to similar targets for a common underlying reason. This can occur, for example, when a Web site at one university carries links to the home page of another on each of its pages, perhaps to give credit for support or collaboration on a project.

In this paper, three new methods of counting links between Web sites will be described and assessed, ones that attempt to approximate the range of Internet collaboration rather than its intensity. The assumption is that the range of different URLs at one institution linked to by another maybe a useful indicator for such a relationship. Behind this is the belief that it may be often the case that if a single URL is linked to more than once by a university site then this will be a result of multiple pages from the same author or internal knowledge sharing in the source site. Thus the number of different link targets is potentially a useful indicator of the extent of Web knowledge. In some models of link counting it may also be the case that the same link is repeatedly used for the same reason, undermining the validity of link counts. An analogy can be made with citations on several levels. A comparable citation range measure could be developed to answer questions such as: "how many different articles/authors/journals has this paper/author/journal/research group cited?" In fact standard citation statistics could be viewed as already being range based (from a Web perspective) since they are based upon the number of citations per manuscript, ignoring repeated citations. On the Web, links are normally counted per page, with an important question being how to aggregate pages into documents for link counting purposes. The opposite of range can be conceived as depth: multiple links to the same target or citations to the same paper may be evidence of a stronger relationship between the two.

The analysis will focus on two issues: the reliability of the metrics and whether they are measuring something genuinely different to the existing ones. The methods proposed will extend the heuristics developed in Thelwall (2002c) and will be tested on the same data, the UK academic Web.

Range Models for Link Counting

The range models described here count how many *different* URLs or Web documents are linked to. A count on this basis may be more representative of the extent of the spread of collaboration, formal and informal, across the universities. The term collaboration here is used in a wide sense, including the use of material posted

on the Web by one scholar and anonymously referenced or used by another. The range concept can be applied to each of the Web document models (Thelwall, 2002c) and its interpretation would be different in each case. Table 1 summarises the new models produced by applying this approach to the first three document models. The fourth one, the university model, would be unchanged and is not included. The original document models will be termed here the standard models to avoid confusion. Essentially the difference is that with the standard models the total number of links between each source and target page/directory/domain are counted whereas with the range models the total number of *different* target pages/directories/domains targeted by each source institution is calculated.

TABLE 1. Hybrid models of Web content.

Model	Counting methodology
Web page/file range	The number of different Web pages linked to in the target site is to be counted. Any target URL is only counted once and subsequent links to the same URL are ignored. A Web page in this context is identified with its URL. Any URL starting with http:// is allowed and URLs will be truncated before any internal target designator symbol '#' to avoid multiple links to different parts of the same page. The focus in this definition is on the URL rather than the page and is not assumed that the target page actually exists at the time of testing. The reason for the lack of checking is that the intention to link is viewed as more important than whether there was a typo in the URL or if the target had disappeared.
Web directory range	The number of different Web directories linked to in the target site is to be counted. Any target directory is only counted once and subsequent links to the same directory are ignored. Directories are defined by truncating URLs just before the last slash that they contain, if one is present. URLs with different "port numbers" (Thelwall, 2002d) than the default 80 are assumed to be associated with different directories.
Web domain range	The number of different Web domains linked to in the target site is to be counted. Any target domain is only counted once and subsequent links to the same directory are ignored. Domains are obtained by stripping any directory structure, file name, port number and password information from URLs.

Note that, unlike in the original document model heuristics, aggregation is only necessary on *target* URLs to implement the calculation. Aggregation at source level only serves to reduce counts to the same document, which is unnecessary here since only the number of different target documents is being counted. Further, the range counting operates separately for each source university but the results will be totalled for each target institution across all sources to give a 'range count' measurement. The choice of aggregating at the source institution level is based upon the hypothesis that the likelihood of overlap in link creation motivation between links from separate institutions is qualitatively different to that within a single university. It is accepted, however, that a simple count of the number of different targets at any given university, irrespective of link source, would also be a plausible type of range measurement.

In the individual Web page range model it will be impossible for a small number of highly targeted Web pages to dominate the link count for a pair of universities because an URL will only count once, even if it is linked to many times.

This should be effective at minimising the impact of repeated 'credit links' – links on many pages of a site to the name page of an affiliated university. In fact the list of the 100 most highly targeted URLs in UK universities includes almost half (45) university home pages (Thelwall, 2002f), many of which may be there as a result of repeated credit links. It is possible, however, that a single cause will produce multiple target pages, which is a weakness for this model. This can happen, for example, when a library site links to many different journal description pages hosted by another university (Tang & Thelwall, 2002). It is also the case that an individual attempting to compile a list of online resources may link to, say, many different online papers from a single author, rather than linking to them indirectly via the papers' author's home page. In this latter case it is perhaps debatable whether a range-oriented counting exercise should count the links separately or not. In the directory model, presuming that the files are hosted in the same directory, only one would be counted. This could be seen as a more human-oriented model because it steps back slightly from the structure in which the information is stored. For this model the design decision whether to link to all pages in a small target site rather than just its home page will tend to no longer influence the calculations. For the domain range model the question perhaps becomes whether anyone at the source institution has heard of the entity owning each domain name in the target institution and believes it significant enough to create a link to one of its pages. Table 2 gives some very simplistic assumptions for these models and presents an interpretation of them based on these.

TABLE 2 Simplistic assumptions for the models and descriptions based upon them.

Model	Web page range	Directory range	Domain range
Simplistic assumption	A Web page is a self contained individual resource	A directory contains files created by an individual	A domain contains files created by an identifiable scholarly unit, such as a large research group
Interpretation of range count from university A to university B	The number of resources hosted by B that are considered useful by someone in A	The number of individuals in B with work considered useful by someone in A	The number of scholarly units in B with work considered useful by someone in A

The following are examples of why the assumptions in Table 2 cannot be held to be generally the case, but it may still be useful as a first approximation to interpretation of the data as long as it is not taken to be a definitive description.

- A single Web directory can contain the work of multiple authors, for example in a digital library.
- A single prolific Web author may have pages in tens or hundreds of directories.
- A domain can host a variety of objects. Perhaps in the computing departments of more research oriented universities there are a similar number of domain names as members of staff, but at the other extreme a large humanities faculty in a less research oriented university could share one domain name, or only have a directory tree on the main Web server.

Methodology

The three range models were implemented on a publicly available database of UK university Web sites (Thelwall, 2001c) hosted at http://cybermetrics.wlv.ac.uk and obtained by a specialist information science Web crawler (Thelwall, 2001b). This crawler covers Web sites accurately in the sense of comprehensively testing for and eliminating duplicates but the results cannot claim to be complete because of the robots.txt convention denying access to some sites and the crawler only being able to find pages by following links from other known pages, normally starting from the home page. The database chosen for this study was of 110 UK universities from July to August 2001, although the two small institutions from this database that were not included in the Education Guardian Tables (2001) were dropped because of the lack of data about their research activities. The database does not include any identified areas of Web sites that are mirror copies of documentation produced elsewhere, although this process is error-prone due to its reliance upon human intervention.

In order to implement the range heuristics, a computer program was written and applied to the revised databases constructed according to the three document models. For each university, its backlink count is based on totalling the links to it from each other university used, using the following heuristics for the different models.

- *Web page/file range model* The link count from institution A to institution B is a simple count of the number of different URLs in the database of the link structure of university A that were from B. URLs in the database had already been truncated to remove internal targets.
- *Web directory range model* The link count from institution A to institution B is a count of the number of different URLs in the database of the link structure of university A that were from B, after truncating each target URL just before the last slash it contained, if one was present.
- *Web domain range model* The link count from institution A to institution B is a count of the number of different URLs in the database of the link structure of university A that were from B, after truncating each target URL just before the first slash it contained, if one was present, then truncating at any ':' character after the main domain name (port number) and removing anything leading up to a '@' character before the domain name (an optional URL feature sometimes containing password information).

Evaluating the results of a metric designed to measure something that has not been measured before (the range of Web connectivity) is clearly not straightforward. Two approaches will be used: quantitative and qualitative. The quantitative approach is to correlate the results with accepted measures of research output in a way that has proven successful in previous papers (Thelwall 2001a, 2002a,e). The rationale is that link counts have been shown to strongly correlate with research productivity and so if the new measures do not then this would be a cause for concern. The qualitative approach, pioneered in a related context for the original document model, is to analyse the reasons for the greatest drops in link counts for a university through converting from one of the document models to the range equivalent. The results of this investigation would particularly aid the *interpretation* of the metric values. An estimate of the total research productivity over the period 1996-2000 for each institution was obtained by taking the official UK government 2001 peer-review driven Research Assessment Exercise set of ratings and totalling the product of each rating with the total number of staff submitted to the category (Education Guardian, 2001; Thelwall, 2002c). This is an estimate of total research productivity rather than

average research capability, so that the highest scores will go to the institutions with the most staff submitted for rating and the highest average ratings. Staff not submitted do not figure anywhere in this calculation and are ignored. This process is does not in any sense give a definitive measure of total research performance, but it is thought to be the most effective statistic of its kind on an international scale because the numbers come from a highly organised segmented peer-review process that is used to decide upon the allocation of UK government research funding (www.rae.ac.uk).

The analysis of model differences is conducted as follows. For the universities with the highest drop in value between two models under comparison, summary URL count files for links to the university in question are consulted to identify whether the cause can be traced to a small number of source universities, or whether it is a widespread phenomenon. In the former case, links to the target university under question from each of the identified source universities are identified with the objective of seeking a pattern. Such a pattern may be obvious from the context of the URLs, otherwise the source pages are visited to continue the investigation.

Results

Spearman correlations are presented in Table 3, alongside the previously discovered correlations for the standard models (Thelwall, 2002c). Figure 1 shows the results for the highest correlating model, and Figures 2-3 show the ratio of the URL counts to standard counts for the different models. The first of these illustrates the different distributions of ratios and the second illustrates the variation of ratios for individual universities.

TABLE 3. Spearman correlations between the results of the different counting models against research productivity. All correlations are significant at the 0.1% level.

Model	Correlation
Standard Web page	0.920
Web page range	0.936
Standard directory	0.925
Directory range	0.940
Standard domain	0.923
Domain range	0.886

From Table 3 it can be seen that the directory range model produces results that correlate best with research productivity. In fact, this metric gives a better correlation than any of the previous standard link counting models, as can be seen from the table, despite the already very high correlation values previously found. From figure 1 it can be seen that the trend is very linear, although there are still outliers in the data in the sense of values that would not fit with an assumption of a normal distribution for the counting results.

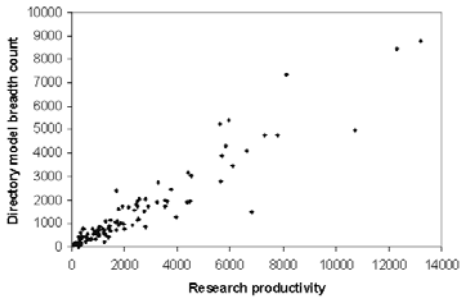

FIG. 1. The number of links to each university based upon the directory range model plotted against its research productivity

From Figure 2 it can be seen that they all the range models display very similar behaviour with respect to the standard models upon which they are based. It is perhaps surprising that the lines are not ordered in height terms according to the degree of generality concerned, however. It might be expected, for example, that the file model line should be above the directory model line. This indicates that there are two competing processes at work: with the standard domain and directory models reducing the counts through aggregation in competition with the range models reducing totals through not counting multiple links to the same URL or partial URL. From Figure 3 it can be seen that for each individual university the results are not highly predictable for one ratio compared to the others. However, a Spearman test found a significant correlation between these values, which is probably a result of the statistics not being independent.

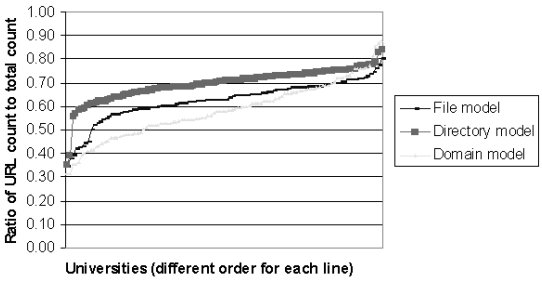

FIG. 2. The ratio of the number of links to each university based upon the range model to the number based upon standard models. The universities are in ascending order of ratio for each line

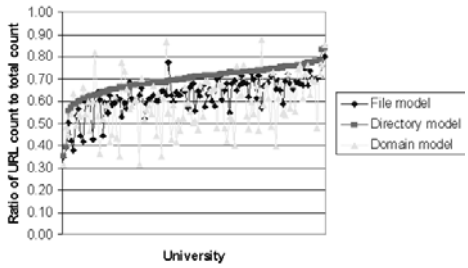

FIG. 3. The ratio of the number of links to each university based upon the range model to the number based upon standard models. The universities are in ascending order of ratio for the directory model line.

Figure 4 gives a different perspective of the data, plotting the results of the directory model for each pair of universities. From this it can be seen that there are still clear outliers in terms of individual pairs. The biggest outlier, 557 links from South Bank University to different directories at Reading University, mainly due to a teacher linking to the Web directories of current and previous students at Reading University. For example, the page at archive.museophile.sbu.ac.uk/cs/people/jpb/teaching/tutees.html contained links to 15 Reading student home directories and 15 further student project directories. As can be seen from the graph, this kind of anomalous linking to different directories is relatively rare, but not unique.

Spearman correlations for the range and standard models between the variables in Figure 4 give 0.779 and 0.766 respectively, supporting the conclusion that the range model is a slight improvement over the standard model, even at the between-university links level.

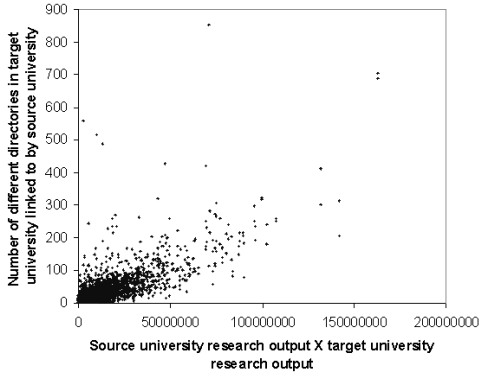

FIG. 4. The directory range link count plotted against the product of the source and target university research productivities

Discussion and Conclusions

The URL counting models all show highly significant Spearman correlations with target university research productivity, with the directory based results displaying the highest reported correlation of any known link metric, an important finding. In order to follow up this result, an investigation was conducted into whether the range metrics were genuinely measuring something different to the standard document model metrics. The causes of reduction in values for the directory model between the standard and range metrics were ascertained through inspection of the original link structure files. Appendix 1 reports the results of this investigation for twenty sets of links between universities. The first ten are cases where the reduction is highest and the second ten represent the average case, where the reduction is at the median level. For 45% of the counts from one university to another there was no change at all. Many (17%) of these had no links recorded by either counting method. The ten top link reduction cases could be tracked down to a single highly targeted directory or set of directories. The ten median cases mostly were a result of one directory being targeted by links from two sources. In summary, then, there was little evidence of multiple link counts representing a deeper connection between the two departments or the hosting of high value research information on the target site. An exception was the case of the University College London biochemistry database, but even this was clearly an artefact of the format in which the information was chosen to

be stored in the source and target databases. The very high link count attributed to this cause is much too big to be treated than anything other than an anomaly. In summary, the range model cannot be claimed to be measuring something genuinely different from the standard model, and so the principle outcome of this study is that the directory-based URL counting model appears to be a better model for analysing interlinking between universities than any of the standard models.

The results presented here concern only one national university system, crawled at one time, and covers only the publicly indexable pages on the sites covered. This is clearly a drawback that should encourage caution in the interpretation of the conclusions in other contexts. The very high correlation found for the directory range model does, however, encourage the belief that it may well be robust enough to stand transportation to other countries. It is known, however, that national variations in Web use do occur (Thelwall *et al*., 2002, Thelwall, 2000a). More problematic is the issue of exactly what the metrics are measuring. Based upon the change analysis, admittedly only for the directory model, and similar analyses of the standard models (Thelwall, 2002c), the following claims are made.

- The range metrics all measure something that correlates with research productivity.
- The different link metrics are measuring a combination factors including the use value of information on the target site and relationships with other universities.
- The metrics with highest correlations with research productivity are more meaningful in the sense of being less dominated by multiple links created for a single underlying cause.

References

Adam, D. (2002). The counting house, Nature, 415, 726-729.

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

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

Bremner, D. H. (1999). The Preparation of Thieno[2,3-b]pyridines. No longer online but copy available at: http://web.archive.org/web/19991104140549/http://scinet.tay.ac.uk/~orgres/paper/index.html, accessed 10 February, 2002.

Chu, H., He, S. & Thelwall, M. (2002). Library and information science schools in Canada and USA: A webometric perspective. Journal of Education for Library and Information Science, 43(2), 110-125.

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

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.

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

Garfield, E. (1998). From citation indexes to informetrics: is the tail now wagging the dog? Libri, 48(2), 67-80

Education Guardian (2001). About the tables, http://education.guardian.co.uk, Accessed 17 December, 2001.

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

Katz, J. S. (1994). Geographical proximity in scientific collaboration. Scientometrics, 31, 31-43.

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

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

Polanco, X, Boudourides, M. A., Besagni, D. & Roche, I. (2001). Clustering and mapping Web sites for displaying implicit associations and visualising networks. University of Patras.

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

Small, H. (1999). Visualising science through citation mapping, Journal of the American Society for Information Science, 50(9), 799-812.

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). Web Impact Factors for Australasian Universities, Scientometrics, 54(3), 363-380.

Tang, R. & Thelwall, M. (2002, to appear). Exploring the pattern of links between Chinese university Web sites, Proceedings of the 65th ASIST Annual Meeting Volume 39 (ASIST 2002), pp. 417-424.

Thelwall, M. (2000a). Commercial Web sites: Lost in Cyberspace?, Internet Research, 10(2), 150-159.

Thelwall, M. (2000b). 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). A web crawler design for data mining, Journal of Information Science, 27(5), 319-325.

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

of Wolverhampton. Available: http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf.

Thelwall, M. (2002a). Evidence for the existence of geographic trends in university web site interlinking. Journal of Documentation, 58(5), 563-574.

Thelwall, M. (2002b). A comparison of sources of Links for academic Web Impact Factor calculations. Journal of Documentation, 58(1), 60-72.

Thelwall, M. (2002c). Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university Web sites, Journal of the American Society of Information Science and Technology, 53(12), 995-1005.

Thelwall, M. (2002d). Methodologies for crawler based Web surveys, Internet Research: Electronic Networking and Applications, 12(2), 124-138.

Thelwall, M. (2002e). A research and institutional size based model for national university Web site interlinking, Journal of Documentation, 58(6), 683-694.

Thelwall, M. (2002f). The top 100 linked pages on UK university Web sites: high backlink counts are not usually directly associated with quality scholarly content, Journal of Information Science, 28(6), 485-493.

Thelwall, M. Binns, R. Harries, G. Page-Kennedy, T. Price E. and Wilkinson, D. (2002). European Union Associated University Websites, Scientometrics, 53(1), 95-111.

Vaughan, L. Q. & Hysen, K. (2002, in press). Do Web Link Counts Resemble Citation Counts: An Empirical Examination, ASLIB Proceedings.

Vaughan, L. & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? Journal of the American Society for Information Science and Technology, 54(1), 29-38.

Wenneras, C. & Wold, A, (1997). Nepotism and sexism in peer-review. Nature 387 341-343.

White, H. D. & Griffith, B. C. (1982). Author co-citation: a literature measure of intellectual structure. Journal of the American Society for Information Science, 32(3), 163-172.

Appendix 1. Tables of link counts and change analyses for the directory model

Note that all names given are derived from the university Web domain name, which can be visited to ascertain the identity of any institution described, should it be considered necessary.

TABLE 4. The ten universities with the lowest URL count to overall count ratios.

From	To	Count	URL count	Ratio
shu.ac.uk	anglia.ac.uk	100	10	0.100
city.ac.uk	leeds.ac.uk	132	15	0.114
hud.ac.uk	leeds.ac.uk	321	37	0.115
ox.ac.uk	wlv.ac.uk	115	15	0.130
ncl.ac.uk	wlv.ac.uk	41	6	0.146
unn.ac.uk	tees.ac.uk	41	6	0.146
cam.ac.uk	ucl.ac.uk	1974	301	0.152
st-and.ac.uk	leeds.ac.uk	291	47	0.162
ic.ac.uk	tay.ac.uk	6	1	0.167
ic.ac.uk	luton.ac.uk	6	1	0.167

TABLE 5. Analysis of the reasons for the above results.

From	To	Main cause
shu	anglia	Links from the web pages of a large SHU-based web site resource to its predecessor on the Anglia server. (The Schools Online Project).
city	leeds	107 links to the home page of LaTeX2HTML creator Nikos Drakos. The use of this package was widespread in the City School of Informatics and the way it was typically used was to convert a single LaTeX document into a set of interlinked HTML pages, all stored in a specially created directory. This arrangement accounts for a particularly high score in the directory model.
hud	leeds	274 LaTeX2HTML links (see above).
ox	wlv	92 links to the UK clickable map of universities that allows the user to jump to the home page of all UK universities and major HE colleges.
ncl	wlv	36 links to the UK clickable map (see above).
unn	tees	32 links to the Online Public Access Catalog of a neighbouring institution from various subject based-directories of the Learning Resources Centre student guides.
cam	ucl	1,435 directories linking to two directories related to an online biochemistry database from the CAMbridge database of Protein Alignments organised as Structural Superfamilies. Each superfamily has its own directory, many with links to UCL for related information, hence the high directory link count.
st-and	leeds	210 LaTeX2HTML links (see above).
ic	tay	6 links to one directory containing a Chemistry paper presented at a conference hosted by IC. For some reason this paper was not hosted on the IC server but on the author's own university server. The reason was probably that the page had been set up in a complex way for an unusual viewing arrangement with four frames designed to "allow more than one part to be viewed simultaneously" (Bremner, 1999).
ic	luton	6 links to the directory of a Web computing manual (server-side includes) from different directories owned by a theoretical physics lecturer.

TABLE 6. Ten universities with average (approximately median) URL count to overall count ratios.

From	To	Count	URL count	Ratio
unn.ac.uk	york.ac.uk	8	7	0.875
uwe.ac.uk	port.ac.uk	8	7	0.875
uwe.ac.uk	shu.ac.uk	8	7	0.875
warwick.ac.uk	king.ac.uk	8	7	0.875
warwick.ac.uk	sunderland.ac.uk	16	14	0.875
wlv.ac.uk	essex.ac.uk	24	21	0.875
wmin.ac.uk	ncl.ac.uk	8	7	0.875
york.ac.uk	bangor.ac.uk	16	14	0.875
wlv.ac.uk	cam.ac.uk	90	79	0.878
shef.ac.uk	bham.ac.uk	213	187	0.878

TABLE 7. Analysis of the reasons for the above results.

From	To	Main cause
unn	york	Two links to the site of a National Health Service information site, one from a library subject resource guide and the other from the "General health links" page of the Faculty of Health, Social Work and Education.
uwe	port	Two links to a robotics conference home page, one from the pages of a lecturer and the others from the pages of a European network for robotics research.
uwe	shu	Two links to a schools online science project site, one from the library educational resources page, and one from a subsite designed to provide useful links as a service to teachers in schools and colleges.
warwick	king	Two links to the university home page, one from a subject network page and one from a large list of UK academic web sites.
warwick	sunderland	Two links to a site for chemistry tests, one from the links page of an online educational technology journal and one from the links page of the Educational Technology Service, producers of the journal.
wlv	essex	Three links to the home page, one from the UK clickable map, one from an automatically classified list of web pages, and one linking to it as the host of a useful data archive. Two links to the press information area of the Institute for Social and Economic Research, one from the home page of a lecturer and the other from the teaching notes of a module associated with them.
wmin	ncl	Two links to the Transport Operations Research Group, one from a list of UK university transport studies groups, part of a larger set of transport-related links pages, and one from a list of all the links in the site, in text format and in a different directory.
york	bangor	Two links to each of two directories. The first, a social sciences school home directory was linked to by two separate lists of links to UK social science departments, one maintained by an individual lecturer and one the official departmental list. The second directory targeted twice was a list of psychology departments, linked to once by the official psychology links list and once by a list of psychology links in the Centre for Reading and Language.
wlv	cam	Eleven directories with two links. For example there are two links to the law department, one from an official list of useful law-related links and the other from a similar list in a different directory.
shef	bham	Many pages with multiple links, including 8 to the home directory. Two of these links came from large lists of UK universities, two from different copies of the same page and the rest came from different sources.