EVIDENCE FOR THE EXISTENCE OF GEOGRAPHIC TRENDS IN UNIVERSITY WEB SITE INTERLINKING

Mike Thelwall[1]

m.thelwall@wlv.ac.uk

School of Computing and Information Technology, University of Wolverhampton, 35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK

The web is an important medium for scholarly communication of various types, perhaps eventually to entirely replace some traditional mechanisms such as print journals. Yet the web analogy of citations, hyperlinks, are much more varied in use and existing citation techniques are difficult to generalise to the new medium. In this context, one new challenging object of study is the modern multi-faceted, multi-genre, partly unregulated university web site. This paper develops a methodology to analyse the patterns of interlinking between university web sites and uses it to indicate that the degree of interlinking decreases with distance, at least in the UK. This is perhaps not in itself a surprising result, despite claims of a paradigm shift from the traditional virtual college towards collaboratories, but the methodology developed can also be used to refine existing web link metrics to produce more powerful tools for comparing groups of sites.

INTRODUCTION

The web and scholarly communication

The Internet and the Web have become increasingly central to scholarly communication. Perhaps the most visible aspect of this is the creation of numerous e-journals, some with the express purpose of replacing expensive print equivalents. But the web also plays host to collaborative workspaces and is increasingly seen as an important mechanism for disseminating results to peers, the tax paying public and consumers. Moreover, it often provides a public unrefereed creative space that is used for informal research, teaching and recreational information, for example in personal home pages. In this context it is natural to investigate whether it is possible to extract useful information from the web about the behaviour patterns of its users or its effectiveness in supporting scholarly activity and communicating findings. Although web server log files can give definitive information about the use of sites, these are not usually publicly available and so not a practical source of information about the popularity of a range of sites. Web hyperlinks, which are publicly visible, have attracted much interest by analogy with citations (Rousseau, 1997; Davenport & Cronin, 2000; Egghe, 2000; Cronin, 2001; Björneborn & Ingwersen, 2001) and

[1] To appear in the Journal of Documentation, 58(5), 2002.

probably also as a result of the success of Google in extracting implicit information from them. In particular, hyperlinks between different web sites appear to have a special significance. It has previously been demonstrated that indicators of institutional research quality can correlate significantly with counts of links to a university web site, showing that academic hyperlink structures can be successfully mined for patterns, (Thelwall, 2001a; Smith & Thelwall, 2002), but this paper presents a deeper investigation into the impact of geographic distance between universities on counts of links between them. Whilst of interest in itself as a study of an important factor in the research process, an underlying motive is the longer-term project to develop the methodology to a point where link metrics can be seen to give information that is reliable enough to be used to inform policy decisions.

It is important to flag at this early stage the issue of the multifaceted nature of university web sites. Each one contains a variety of material, with scholarly content probably forming a minority in almost all. Studies of undifferentiated academic hyperlinks have to contend with the problem that they originate from very different motivations and that a simple backlink count is necessarily a crude tool. Moreover, conclusions from such studies will be difficult to tie to underlying causative factors that do not apply to the majority of areas of web use.

Previous geographic and web link studies

Although previously widely accepted, the existence of a geographic component in research was first tested for by Katz (1994). This work compared the extent of co-authorship of academic papers between scholars in the UK, Canada and Australia and showed that in all cases there was a clear decrease in the total amount of collaboration with increasing distance between hosting institutions. In the era of the web, with 'collaboratories' of disparate researchers able to communicate instantly, some of the underlying forces behind local collaboration have surely lessened, but is this geographic pattern still visible in the new international medium of almost instantaneous communication, the web?

Web links are extensively studied as a tool for information retrieval (Brin & Page, 1998; Chakrabarti *et al*., 1999; Craswell *et al*., 2001) and as objects of interest in their own right (Broder *et al*., 2000; Björneborn & Ingewersen, 2001). More specifically, there have been many studies of web links to study the 'impact' of various areas of the web including countries (Ingwersen, 1998), national university systems (Thelwall, 2001a; Thelwall, 2001b; Smith & Thelwall, 2002) national departmental websites (Thomas & Willet, 2000; Chu *et al*., 2002), e-journals (Smith, 1999; Harter & Ford, 2000; Darmoni *et al*., 2000), and information web sites (Cui, 1999; Hernandez-Borges *et al*., 1999). Despite a faltering start with many initial negative results, Ingwersen's methodology has now given rise to many positive conclusions, in terms of significant correlations between link count metrics and accepted external metrics, such as institutional research quality for UK universities (Thelwall 2001a; Smith & Thelwall, 2002; Thelwall, 2002). It is worth emphasising, however, that there are no known positive findings, yet, for objects of size smaller than entire university web sites. This perhaps indicates the degree of averaging required to get significant results. There are two known previous studies of links between pairs of universities. The first found suggestions of geographic clusters of UK institutions, particularly for Scottish and for Manchester-based universities (Thelwall, 2001c). This used exploratory techniques and did not give clear evidence of an overall trend. The second gave evidence that link counts between universities

are approximately proportional to the quadruple product of the size (in academic staff numbers) and research quality of the source and target institution (Thelwall, 2001d).

The research question

The two interrelated questions to be addressed in this study are whether a geographic pattern can be ascertained in university web interlinking, and whether a methodology can be developed to illuminate such a relationship. The scope is UK university institutions because of the existence of a relatively definitive research rating for these, a crucial factor in the current state of development in web link metrics. A relatively homogeneous linguistic background is also an advantage in isolating the geographic factor. The many essential questions relating to the reliability of the techniques used will be addressed in the discussion after the main results.

METHOD

The university institutions chosen as the basis for this study were all the UK universities with the exception of the Open University, which is unusual in being a distance learning institution, and Cranfield University, which is primarily postgraduate. The large federal universities of London and Wales were treated differently by including their largest constituent colleges rather than themselves as single entities. This approach probably reflects public perceptions and certainly reflects the pattern of domain name usage, all having single (or multiple but equivalent) domain name roots. Also included are the largest non-university colleges, which are very similar in scope to legal universities. The list used corresponds to the well-known Times Higher Education Supplement list (Mayfield University Consultants, 2001), aimed at students choosing a university.

Information about links between universities was collected by a specialist information science web crawler (Thelwall, 2001e) and subsequently made publicly available (Thelwall, 2001f) with several papers already written using the data. Although the issues around collecting this kind of data are discussed in these papers, two key points will be highlighted. First, crawling a university web site necessarily produces results about only those pages that can be found by following links and excludes entire collections of pages, such as those hidden inside databases and in areas that exclude robot access. Second, the crawl has attempted to exclude mirror sites on the grounds that these are not created by the site owners. It is beholden upon the investigator, therefore, to show that this subset of the UK academic web is a meaningful object of study.

From the database of the link structure of the UK university institutions chosen, counts of links between sites were calculated using a specially written program. Links between different pages in the site of a single university were not counted. The identification of the target of each URL was string comparison of its domain name with a list of known root domain names for each university. For example, any link URL with domain name containing .wlv.ac.uk or .wolverhampton.ac.uk (both with initial dots) would be counted as pointing at Wolverhampton University. Links are also counted if the full domain names are equal to the two roots e.g. wolverhampton.ac.uk and wlv.ac.uk, in the above example. This arrangement was necessary to avoid false positives caused by one domain name being a sub-string of another, in particular qmced.ac.uk and ed.ac.uk.

Distances between universities were obtained from the co-ordinates of the official postcode of each university obtained from MultiMap.com. No correction was

made for the earth's curvature and no allowance was made for the Northern Irish universities being located on a different island. The distances are only approximations to the actual average distance between the constituent parts of universities, which span more than one postcode. This is particularly the case for multi-site institutions, but the approach used seems to be the best practical one. To protect against data entry and computational errors, the data set was processed with the multi-dimensional scaling program PROXSCAL in SPSS, which should ideally produce an accurate map of the UK locations. This was true, with the exception of the Robert Gordon University, which was placed on the wrong coast of Scotland but in a position that was algorithmically stable subject to small perturbations, therefore supporting the accuracy of the data overall.

The average research quality of the institutions was taken from the Times Higher Education Supplement (Mayfield University Consultants, 2001), which averages the grades awarded to each one by the most recent government Research Assessment Exercise (RAE), in 1996. The RAE is conducted separately for a range of 69 subjects, with each submission being scored on a seven point scale. It is possible to average the results of all submissions of a university, taking into account the number of academic staff submitted in each case and the total in the institution in order to get an overall research quality score. This figure is subject to question because of, amongst other things, gamesmanship in the decision of how many staff to submit, but is probably on an international scale an unusually authoritative judgment. The research quality information is needed for effective mining of academic web link data.

RESULTS

As can be seen in Figure 1, the distribution of the 11,556 link counts between the 109 institutions (excluding links between different pages at the same university) is highly skew, with 18% zeros. The highest count is 33,228 from Warwick to University College London. This number is mainly due to a set of biochemistry pages on the Warwick server, "Comparison of Protein Structure Classifications" that contains tens of thousands of links to an UCL-hosted database of classification results (Hadley, 2001). This example illustrates that raw link counts are highly unreliable as indicators of the degree of web interlinking between the individuals at universities. A single web link may be the result of considerable deliberation by an eminent scholar or merely one of thousands created by an automatic process. This is a major problem when attempting to identify trends in the behaviour of the creating community. In order to partially overcome this problem the data to be used will be the minimum of the two link counts between universities, loosing the directionality of the data but gaining reliability. This then measures the degree of interconnectedness of the institutions and is based upon the assumption that anomalies will be sufficiently rare to be infrequently bi-directional.

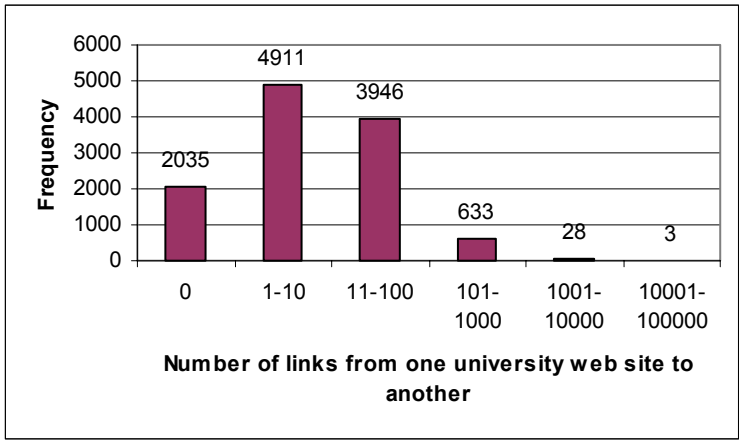

Figure 1. *The distribution of counts of links between the web sites of UK university institutions is highly skew.*

Following previous WIF studies (e.g. Thelwall, 2001a) it is to be expected that counts of links to a university should correlate with its research rating times the number of full-time academic staff. With the raw data for this paper the correlation is 0.546, but using the totals of all minimum link count figures associated with each institution instead, the equivalent correlation is 0.905, which tends to support the contention that the new data is more reliable for studying the behaviour of the creators. Figure 2 shows the trend for the latter case. The research rating figure used here is the one derived from the official government Research Assessment Exercise.

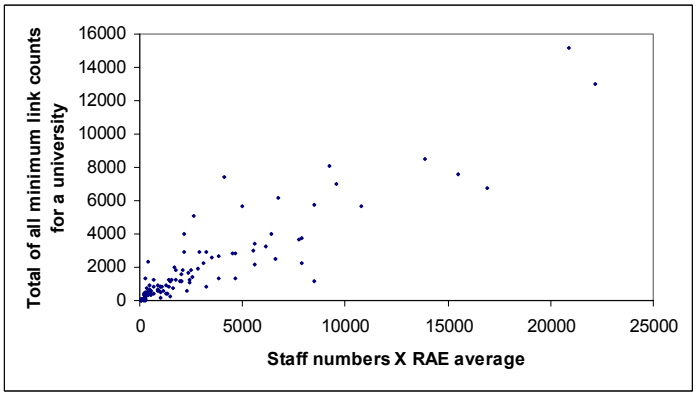

Figure 2. *The total of all minimum link counts associated with an institution shows a significant correlation with its average research rating times the number of full-time equivalent academic staff.*

Figure 3 shows an ambiguous relationship between geographic distance and average link counts. There is generally a fall-off with distance but there is also a large amount of interconnection between the top English and top Scottish universities, producing the second peak. This shows that the relationship is greatly affected by non-geographic factors. The uneven spread of top institutions in the UK can be

factored out by drawing upon the previous discovery that link counts are related to the research quality and academic staff size of both source and target institution, with the quadruple product of these being a useful predictor of link counts (Thelwall, 2001d). Dividing the actual minimum link counts by this quadruple product is a logical step towards the elimination of the research and size-related trends. In the new data, however, the relative size of fluctuations in the smallest counts are greatly magnified and so a weighted average is needed to leverage the greater reliability of the larger counts. The quadruple product itself is the logical choice for relative weight sizes, being proportional to the expected link count. Figure 4 shows averages for the results of these new calculations. A clear downward trend is now evident, with a dramatic fall-off from the closest institutions.

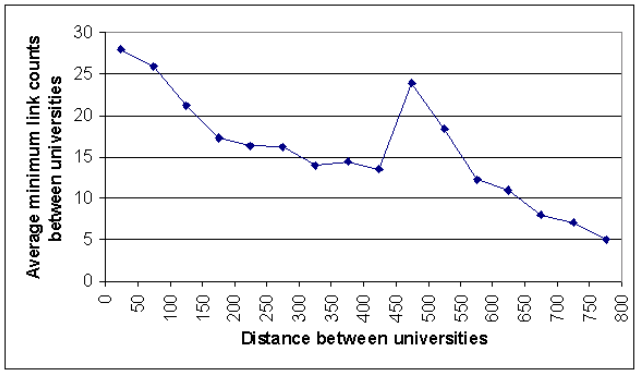

Figure 3. *Average minimum link counts between institutions shows an uneven fall off with distance apart, the large jump caused by interlinking with the top English and Scottish universities.*

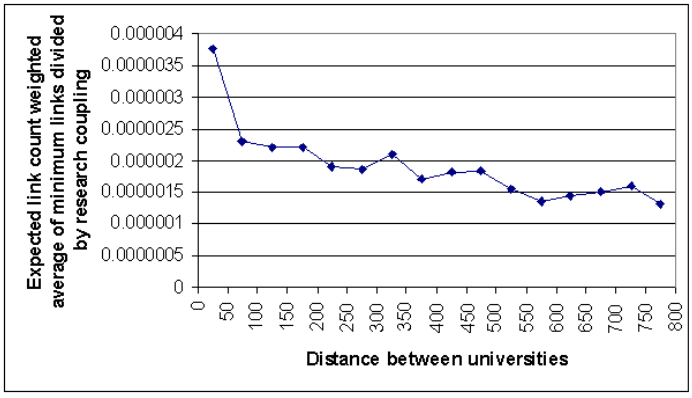

Figure 4. *Average minimum link counts between universities show a clear decline with distance, after factoring out the research and size related components.*

DISCUSSION

Critical commentary on the results

Although Figure 4 gives a seductively believable impression of a clear outcome, there are many reasons to tread carefully about drawing general conclusions. One important consideration is that the figures come from only one country, and that it has a far from uniformly scattered group of universities. It is possible that the graph reflects to some extent a core and periphery relationship centred on clusters of research-based institutions in England and Scotland, for example. The larger, more research-oriented universities necessarily dominate the calculations because the weighting in the average. Future results confirming the trend for other countries would certainly strengthen the findings. Another possibility, which could be implemented if the degree of interlinking of all universities increased in the future, would be to carry out the same exercise with a subset of institutions, excluding the largest.

At the level of raw data collection, additional problems are present. The source is a crawl of UK university websites that involved human intervention to identify mirror sites and other anomalies to exclude, an unwanted but necessary source of potential errors. More of a problem, however, is the fact that link targets are unvetted and that the counts used will include those targeted at mirror sites. The use of the minimum link count calculation was an attempt to minimise the impact of this, but it is clearly unsatisfactory whenever two sites both contain 'additional' links to each other. The other principal source of concern is the extent of coverage of sites. This is an issue because the decisions of individuals can have a big impact on the number of pages indexable on a site. Examples of this are given below.

- A strong centralised content management system could hide much content in a web-accessible database. This is part of the "invisible web" problem.
- Universities with an index page automatically linking to all staff and student home pages will tend to be better covered.
- Universities banning robots from large areas of their site will tend to be less well covered.
- Web management decisions such as how much storage space to allocate users, whether to allow them to have publicly accessible pages and how vigorously to delete old material can have an impact on actual web site size.

Because of all these problems, which are endemic to web link count studies, the raw data is much more subject to arbitrary fluctuations which makes it harder to identify patterns even when they are actually present in the behaviour of the web page creators. This explains the amount of averaging that was required to get a clear trend, but also provides a caution that it could be caused by factors unrelated to the behaviour of the majority of people associated with the institutions. The evidence presented here is, however, supportive of the hypothesis that geographic distance is a factor in academic web link creation, but cannot be taken as a definitive result or assigned a statistical level of significance.

Does the choice of universities impact on the results? Probably the most contentious part is including individual colleges rather than the entire university for the big University of London (a point of difference with Katz). The colleges of the University of London are geographically close together but tend to be unusual in their degree of specialisation, with perhaps only University College London being fully

multidisciplinary. This may well be a cause of less subject-based interlinking between them and thus lower than expected overall counts. If all of these colleges except UCL are excluded from the data then the trend in Figure 4 is even more pronounced for low distances, but the tail of the graph is more irregular.

Does the choice of interval impact upon the results? For smaller intervals, the trend is still visible, but much more ragged. This indicates that the variability of the data requires the averaging of a relatively large number of points in order to bring out a pattern.

Interpretation of results

Given that subject to appropriate averaging, there is some evidence of a geographic trend in university web site interlinking, what could this signify? Unfortunately this is a question that cannot be given a clear answer as yet because of the variety of reasons for creating links (Cronin *et al*., 1998; Thelwall, 2001a) and the absence of a definitive study to categorise the reasons for hyperlink creation and to ascertain the proportion in each category. What is known, however, is that although link counts correlate strongly with research, link targets are very rarely academic papers, although many more links to academic papers are present in postscript and PDF documents on the web (Goodrum *et al*., 2001), mediums not covered in this survey. Common scholarly targets are researchers' home pages, research group and departmental sites. Other frequent targets are the university home page, general information pages and leisure activity sites. A count of links represents, therefore, a metricised pot-pouri of motivations, and it would be presumptuous to interpret the results as a finding about any particular aspect of scholarly communication, such as the geographic spread of collaboratories. The connection with research is, however, clear in the data although the mechanism through which it is realised, if any, is not.

One feature of the results is, however, striking. The comparison between figure 3 and the equivalent graph from Katz's (1994) paper (covering publications between 1981 and 1990), which does not display a double hump, gives an opportunity for a retrospective reanalysis. Comparison with the earlier graph is a little problematic because it reports total collaborations rather than average collaborations per pair of universities within the distance interval apart and does not, therefore, take into account the number or size of institutions within each distance band. It also presumably covers the much smaller number of institutions that were universities before 1990. Nevertheless it can be inferred from the graph shape that the top universities in England did not collaborate extensively with the top universities in Scotland. The results presented here raise the possibility that the electronic era has changed this and made such collaboration more common. In order to verify this speculation Katz's study would have to be repeated using current data on co-authorship of scientific papers.

Potential applications of web link research

In the context of the ongoing effort to develop more powerful web link metrics, started by Ingwersen (1998), this paper presents one technique for dealing with the low quality of the data, the minimum link count, and some evidence of a geographic trend that can be used to build new metrics that factor out this and research trends. For example, if the expected degree of interlinking can be predicted for a pair of universities then this can be compared with observed counts and used to provide supporting evidence for the efficacy of individual university policy decisions

or categories thereof. Some examples of hypotheses that could be investigated are given below. All are of the type that could not be effectively attacked using bibliometric techniques with the Institute for Scientific Information citation databases.

- Do highly centralised content management systems impede or facilitate web use by individual academics?
- Do technical aspects of site design, such as dynamic site front ends, make it easier or harder for other academics to cite through hyperlinks desired content pages?
- Are there significantly different patterns of web use between disciplines?
- Are there regional web-based collaboration trends, or research similarity collaboration trends, in addition to average trends across a country?

It must be emphasised that at the current level of development of web link research, the methodologies described here should be used to support other approaches, rather than to be used as a primary source of evidence. As is evident in the discussion above concerning Katz's graph, however, it can also be used for exploratory research. It is hoped that future discoveries and web use trends will improve the reliability of the data and increase confidence in the results, eventually to the extent that they can be used with confidence as a primary source of evidence.

One of the problems of web link analysis is that the degree of averaging necessary to get stable results currently necessitates a degree of generalisation than makes explorations of specific web phenomena, such as disciplinary web use, difficult. This is a particularly important point in the light of the body of social informatics research that demonstrates the fallacy of technological determinism: that which is made theoretically possible by technology interacts in different ways with existing cultures and motivations to generate different outcomes, even in subfields of the same discipline (Kling & McKim, 2000). The research presented here is refuting a technological deterministic viewpoint, however, that geography is irrelevant on the web. Moreover, the above list of illustrative research questions show that the averaging necessary for stable results does not necessitate a determinist approach.

CONCLUSIONS AND FURTHER WORK

The positive results both strengthen the case for using web link mining as a tool with the potential to reveal underlying trends in academic web site interlinking, and provide a method to extract both research and geographic trends from the raw data in order to be able to analyse other factors more successfully. One further clear outcome from the discussion is the need to refine the methodology in order to be able to tie link count results to the average behaviour of those associated with the universities: to exorcise the spectre of potential causation by anomalies.

It is acknowledged that the degree of effort required to extract information from web links currently far outweighs its value for anyone for whom web links are not their primary object of study. In order to facilitate the use of these techniques more widely it is intended to publish on the web the raw data from web crawls together with tools to analyse it (http://cybermetrics.wlv.ac.uk/database/). The methodology will also be published so that users can critique it and interpret the results in an appropriate context.

In terms of expanding the results to cover additional countries, the main barrier is expected to be small web sites in developing countries and a lack of definitive research ratings in many others. In a large country such as the USA, it may

be possible, however, to work with several subsets of institutions that are judged to be of a similar research level.

Returning, finally, to the main focus of this paper, the extent of the geographic trend is perhaps more surprising than its existence. It is ironic, nevertheless, that with this key global communications technology, universities are still most likely to be linked to their neighbours.

REFERENCES

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.

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

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

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S. R., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999). Mining the web's link structure. *IEEE Computer*, 32(8), 60-67.

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

Craswell, N., Hawking, D. & Robertson, S. (2001). Effective site finding using link anchor information. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)* (pp. 250-257). New York: ACM Press

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

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

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

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

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

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

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

Hadley, C. (2001). Comparison of Protein Structure Classifications Available: http://globin.bio.warwick.ac.uk/~hadley/db/, visited November19, 2001.

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

Hernandez-Borges, A. A., Macias-Cervi P. & Gaspar Guadardo M. A. (1999). Can examination of WWW usage statistics and other indirect quality indicators help to distinguish the relative quality of medical Web sites? *Journal of Medical Internet Research*, 1(1), e1. Available: http://www.jmir.org/1999/1/e1/index.htm

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.

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

Mayfield University Consultants (2001). League Tables 2001. *The Times Higher Education Supplement* May 18, T2-T3.

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

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

Smith, A. & Thelwall, M. (2002). Web Impact Factors for Australasian Universities, *Scientometrics*, 54(1-2), 363-380.

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

Thelwall, M. (2001c, to appear). An initial exploration of the link relationship between UK university web sites. *ASLIB Proceedings*.

Thelwall, M. (2001d, to appear). A Research and Institutional Size Based Model for National University Web Site Interlinking. *Journal of Documentation*.

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

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

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

Thomas, O. & Willet, P. (2000). Webometric analysis of departments of Librarianship and information science. *Journal of Information Science*, 26(6), 421-428.