The Connection between the Research of a University and Counts of Links to its Web Pages: An Investigation Based Upon a Classification of the Relationships of Pages to the Research of the Host University

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478
Gareth Harries
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: g.harries@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321485

Results from recent advances in link metrics have demonstrated that the hyperlink structure of national university systems can be strongly related to the research productivity of the individual institutions. This paper uses a page categorization in order to show that restricting the metrics to subsets more closely related to the research of the host university can produce even stronger associations. A partial overlap was also found between the effects of applying advanced document models and separating page types, but the best results were achieved through a combination of the two.

1. Introduction

1.1 Background to Web Link Research

Although Web authors could in theory create links to other pages at random, in reality their behavior patterns are strong enough to be successfully used in Web information retrieval algorithms (Brin & Page, 1998; Kleinberg, 1999; Flake *et al*., 2002). In information science, researchers identified several years ago that hyperlinks had the potential to reveal new types of information about both scholarly communication (Larson, 1996; Almind & Ingwersen, 1997; Rousseau, 1997; Cronin *et al*., 1998; Ingwersen, 1998) and the value of the Web pages themselves (Davenport & Cronin, 2000). In support of this, there is now a considerable body of research to show that patterns of Web linking between universities can be strongly associated with research productivity. Statistically significant correlations have been identified in the UK (Thelwall, 2001a; Thelwall, 2002a) Australia (Smith & Thelwall, 2002) and China (Tang & Thelwall, 2002). Associations have also been found between link counts and research-related figures for journals within a discipline (Vaughan & Hysen, 2002), for schools within a discipline (Chu *et al*., 2002) and for French Medical University complex websites (Douyère *et al*., 2002). An association between links and geographic distance has also been demonstrated for the UK using a research-related model (Thelwall, 2002d). In the light of these findings there is a pressing need to investigate further into why such associations are present so that meaningful inferences can be drawn from large-scale link counting exercises.

Two recent papers have developed advanced link counting methods in order to more successfully capture aspects of the underlying behavior of Web authors than the previously used variants of simple link page counts. These were created partly in response to concerns raised about the use of link data (Rousseau, 1999; Smith, 1999; Egghe, 2000; Thelwall, 2000b; Bar-Ilan, 2001; Björneborn & Ingwersen, 2001). The first models used different levels of aggregation of pages into "Web documents" in an attempt to remove the effect of spurious

---

[1] To appear in the Journal of the American Society for Information Science and Technology, 2002/3.

duplication of links, for example when a subsite contains an identical link on many pages (Thelwall, 2002b). The second approach used the same document models but ignored repeated links from one university to the same "document" at another. This produced a set of metrics at different levels of aggregation that all focused on the range of the connection to each target university (Thelwall & Wilkinson, 2002). All metrics are related since they use the same data and six out of the seven showed a significant and quite linear association with measures of British university research productivity. The models are described in detail below.

It is likely that only a small proportion of links actually target pages with scholarly content (Thelwall, 2001a) despite the significant research productivity associations found. The tentative explanation given for this was that links were perhaps related to the reputation of the target university, although this explanation was somewhat undermined by the discovery of evidence for a strong relationship with the research productivity of the source institution (Thelwall, 2001b). Another possible explanation is an increased general level of Web use in the more research active institutions. The uncertainty over the causes of general Web linking between university Web sites is a serious problem from the point of view of validity - interpreting the results. There are three interconnected issues related to this.

1. *Categorization* What kind of pages are linked to?
2. *Motivation* Why do scholars link to these pages at other universities?
3. *Host university research relationships* What is the cause of the relationship between the research conducted at a university and the propensity of others to link to its pages?

These questions are genuinely new ones because there is no *a priori* reason to believe that theories applying to off-line phenomena, such as journal citations and research productivity, will automatically transfer to the Web. In the context of general Web use, for example, although some studies have concluded that the Web mirrors society (e.g. Cresser *et al*., 2001) others emphasize the differences made possible by the technology (e.g. Kling & McKim, 2000; Sapienza, 1999; Stafford & Stafford, 2001; Teo, 2001).

1.2 Page Type Categorizations

The challenge of interpreting counts of links to university Web sites is greatly complicated by their heterogeneity. A single site is likely to contain information created by different types of authors (scholars, administrators, students), for different audiences (internal/external, prospective/current/past students, the public, other scholars) (Middleton *et al*., 1999) with differing content levels (academic papers, books, teaching notes, student assignments, job adverts, hobby pages, photos or videos of family members) in multiple recognizable and novel genres (lecture notes, link lists, frequently asked questions pages) (Cronin *et al*, 1998; Crowston & Williams, 2000; Haas & Grams, 2000; Harter & Ford, 2000; Cronin, 2001), and may even contain misinformation (Calvert, 2001).

Surveys based upon large approximately random samples of Web pages have tended not to perform any semantic analysis (Lawrence & Giles, 1999; Thelwall, 2000a; Koehler, 2002). There are, however, some sources of wide scale topic based classifications, such as from search engine directories and the output of programs that perform automatic classification or grouping (e.g. Jackson & Burden, 1999; Kleinberg, 1999). These could be used to help categorize pages, but would give topic-based results that could only partly address the question of relationship to the research of the host university. For example mirror sites and other copies of collections of pages presumably have the strongest connection to the research of the originating institution but have identical content in all locations.

The most common link targets are of particular interest and Thelwall (2001c) has looked at the 100 most linked to pages hosted by UK universities, counting only links from other UK universities. The list was dominated by university home pages, but there were also many home pages of groups associated with some form of electronic research, such as schools of computing or initiatives for the use of information technology in higher education. This paper used the term 'credit link' to refer to links created to home pages for the purpose of acknowledging a relationship rather than to guide the reader to useful information.

1.3 Motivations for Web Link Creation

There are no known specific studies of author motivation for link creation either for general or for university Web pages, although the issue is alluded to in various contexts. Wikgren (2001), for example, analyzed contributions in online media that involved a specific medical topic and found associations between the types of contributors and the types of resources that they cited.

Cronin *et al*. (1998) devised a scheme of eleven categories (each with a set of exemplars) to classify genres of invocation (not necessarily with links) of the names of a small set of leading academics, throwing into relief the extreme diversity of resources on the Web, even for this relatively narrow task This classification was genre-based rather than topic-based and used search engines to find the pages included. It was remarked, however, that the scheme did not attempt to take in to account "the contexts in which the invocation occurs", and that this was a logical future step for research.

A small body of work has cast some light on authors' motivations for the inclusion of URLs in the most scholarly of Web publications. Kim (2000) found that the reasons for citing URLs in e-journal articles extended those normally associated with print journal articles to include medium-specific motivations such as linking to illustrative multimedia. Goodrum *et al*. (2001) used the CiteSeer database of online papers from journals, conferences and other sources and discovered that computing conference papers, which are not in the database used for the Journal Citation Reports (Garfield, 1994) unless published subsequently in a journal, have a different citing spectrum. For example they tend to cite other conference articles more than journal articles do. Since conference articles are more likely to be freely available online than journal articles, at least in computing, this is evidence that even the most scholarly type of Web publication may follow a different average referencing pattern to print journal articles. Lawrence (2001), however, has given evidence that papers are more likely to be cited if they are made available online, which may in the longer term provide an impetus to increase the volume of online publishing and even out the difference with print journals.

1.4 Host University – Link Target Page Relationships

One previous study has classified links between university Web sites into broad categories according to one person's opinion of the content of each target page. These categories were genre rather than topic-based, being oriented on the type of information content of a page and included a dichotomy between research related pages and others. It was found that when link counts were restricted to the broadly research related target pages, the metric results associated more strongly with a measure of each target university's research productivity. This is the most direct indication yet that aggregates of link counts between universities can be used to reveal information about online research impact. Despite this, an association was also present for all categories of pages unrelated to research, albeit a weaker one (Thelwall, 2001a). In the light of the results from the document models (Thelwall, 2002b), however, it is not known whether the classification was successful because of a stronger widespread underlying relationship between links and research or because of the removal of outliers from the reduced data set.

The question of the extent to which seven Web link metrics produce different results when restricted to a particular type of resource for link targets is addressed in this study. A classification of the most commonly linked to URLs from UK universities was made based upon the relationship to host university research and then each of the metrics evaluated on subsets of the data. Two specific questions were used to drive the investigation.

- Do the metric results associate more strongly with research productivity when restricted to host university research-related targets than when applied to the full data set?
- Do the advanced document models completely subsume the categorization process in the sense of producing the same overall effect, or are the two complementary?

2. The Classification Scheme

A classification of (link target) pages as to their relationship with the research of the hosting institution is given in Table 1. It is hypothesized that the different categories of pages stand in qualitatively different relationships to the scholars at the hosting institution from the perspective of attracting links from other institutions. The first illustrates this point most clearly: a link to a mirror copy of a site is hypothesized to carry different information from another to a locally created and owned page. Similarly, since the focus of this study is on research productivity, pages unrelated to this logically fall into a separate group. The third type reflects the hypothesis that certain types of information attract online attention disproportionate to their overall value as a result of their primary use being online (see Kling & McKim, 2000; Thelwall, 2001c). An example of this would be a list of all US university home pages in contrast to a list of all US university postal addresses. The Table 1 list is applied in order, so that a page is classified according to the first criterion that it meets.

TABLE 1. Classification of host university research relationship

Type of page	Description and exemplars
1. Non-locally created	Mirror sites or copies of pages produced by other organizations. Web sites for external establishments that are hosted on a university's servers (e.g. a national professional society).
2. Non-academic content (or academic content produced by non-faculty)	Student union pages. All Web sites created by student societies, even if they have an academic flavor (e.g. an astronomy society). Web pages that have a recreational purpose, even if written by faculty. Pages giving general information about the geographic area of the university. Careers information, central administrative pages, online prospectuses and information for prospective students. In this category were also included library and museum Web sites, catalogues and email lists. These do have academic content but it was speculated that these are not strongly associated with the direct work of researchers.
3. High profile pages and subsites	Pages or collections of pages that attract particularly high numbers of links because of the *type* of resource they are. These are primarily online databases, gateway sites or external links lists. The following additional types of resources are also included: e-journals and conference Web sites; multiple-university collaborative project pages; projects funded by a national learning technology initiative; and pages with direct pedagogical content.
4. Other pages	All pages not alluded to above. For example: university, departmental and research group home pages; online scholarly articles created by the host institution; home pages of researchers; student created course-related pages.

3. Advanced Web Document Models

When counting hyperlinks the need arises to identify repeated links that it is hypothesized should only be counted once. For example if one Web page contains two identical links, perhaps one at the top of the page and one at the bottom, then it is likely that both have been created for the same reason and so it seems reasonable to treat them differently by counting only one of them (Thelwall, 2001a). The advanced Web document models generalize this to aggregate source and target pages also at the directory, (sub)domain and site level, as described below. The descriptions are taken almost verbatim from Thelwall (2002b).

- *Individual Web page* Each separate HTML file is treated as a document for the purposes of extracting links. Each unique link URL is treated as pointing to a separate document

for the purposes of finding link targets. URLs are truncated before any internal target marker '#' character found, however, to avoid multiple references to different parts of the same page.

- *Directory* All HTML files in the same directory are treated as a document. All target URLs are automatically shortened to the position of the last slash, and links from multiple pages in the same directory are combined and duplicates eliminated.
- *Domain name* As above except all HTML files with the same domain name are treated as a single document for both link sources and link targets. In particular, this clusters together all pages hosted by a single subdomain of a university site.
- *University* As above except that all pages belonging to a university are treated as a single document for both link sources and link targets.

As an example, when using the directory model two links to different pages in a common target directory would be counted as only one if they came from a common *source* directory. Both the domain and directory models have been found to be very effective in terms of producing high correlations with university research productivity. The range metrics use the same set of document models but count only the number of *different* target URLs, therefore ignoring multiple links from one university to the same document at another, irrespective of originating document (Thelwall & Wilkinson, 2002). They are equivalent to the above four definitions except with the university link model applied to document sources in all cases. From this it can be seen that the university range model is the same as the original university model and so there are seven different metrics in total. The results of these metrics are an improvement over those from the standard page based document models through the reduction of outliers in the data.

In the following discussion page/directory/domain/site outlink refers to a link in a Web page that is targeted at a different page/directory/domain/site. Similarly, a site page/directory/domain/site inlink is a link in a Web page in a page/directory/domain/site outside the one in question that targets it. Since each inlink is also an outlink, the difference is one of perspective (Björneborn, 2001b).

The case for file, university and domain (standard and range) counting models is illustrated in Table 2, based upon the Björneborn diagram in Figure 1 (Björneborn, 2001a). In the diagram, all pages are assumed to link to all other pages, including themselves. Site self-links are always ignored, whichever metric is being calculated. The directory model follows the same principles but is not shown for reasons of space.

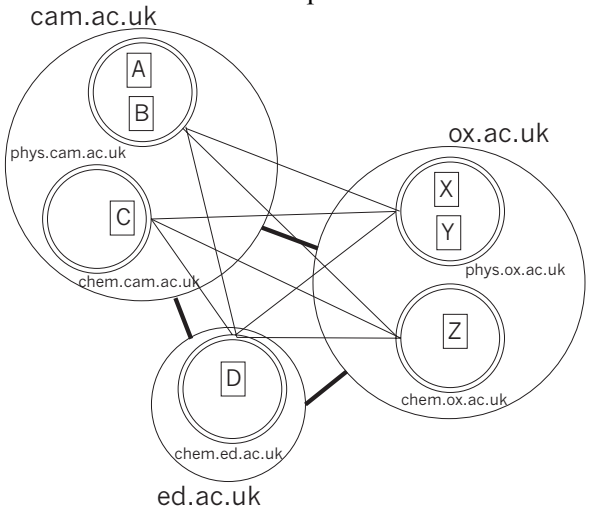

FIG 1. A Björneborn simplified diagram of three universities in which all pages are assumed to connect to all other pages (links not shown). Domain model links are in fine lines and site model links in thick lines. All links are bi-directional.

TABLE 2. Link counts resulting from the Fig. 1 configuration, where all pages link to all other pages.

Model	Links from cam to ox	Links from ed to ox	Total inlinks to ox
Standard file	9	3	12
File range	3	3	6
Standard domain	4	2	6
Domain range	2	2	4
Standard university	1	1	2

4. Methodology

UK universities were chosen for the study because of the existence of an unusually authoritative source of assessment for the research outputs (Education Guardian, 2001). The gross research productivity of each institution is estimated to be the average government research rating times the total faculty. The link structure of the UK universities was obtained from version 2 of a publicly available database of the link structures of 107 major university institutions from July, 2001 (Thelwall, 2001d) created by a specialist information science Web crawler (Thelwall, 2001e) (http://cybermetrics.wlv.ac.uk/data) in order to avoid the known problems of reliance upon commercial search engines for data (Rousseau, 1999; Mettrop & Nieuwenhuysen, 2001). All crawlers only cover a proportion of the Web (Lawrence & Giles, 1999; Thelwall, 2002c), but based upon previous results, this fraction is nevertheless considered to be a meaningful object of study. A program was written to process the link database and produce for each university a summary file of all pages in its Web site (past or present) that were the targets of links from other universities in the database, together with the frequency of these inlinks. No check was performed to see whether the link targets actually existed under the assumption that it was the intention to link that was important, irrespective of typos or the subsequent disappearance of the target page.

For each university the first author then classified all target pages that had at least 5% as many links as the home page according to the categorization above. The 5% figure was imposed for practical purposes in order to avoid having to classify the tens of thousands of pages with just one or two links to them. Making the bar sensitive to the size of the site through the use of the home page site inlink count was a simple method of giving equal attention to pages with similar *relative* impact on the overall site inlink count. The unclassified pages in this schema are 'noise' in the data set. Under the hypothesis that these are not biased in content, improved overall results from this partial categorization would suggest that more strongly significant results could be obtained from a fuller categorization. In addition, an attempt was made to avoid a systematic bias in the low inlink count pages. For each university, all target URLs were briefly scanned in the summary file, again by the first author, and then any common patterns of file naming identified were classified. For example there were hundreds of pages starting with pinkerton.bham.ac.uk that had low link counts, and were classified *en masse* as 'non-locally created'. The classifications were revisited a second time at the end of the process in order to try to maintain intra-indexer consistency.

The classification decisions are claimed to represent something of an 'expert' categorization on the basis of a similar large scale survey undertaken by the same person (Thelwall, 2001a). Classification decisions are known to be problematic, even for professionals such as librarians dealing with relatively consistent genre types such as books. In fact even inter-indexer consistency has been claimed to be an unreliable indicator of accuracy (Cooper, 1969). In this light the classification of the very complex genre-fluid Web pages is presented as a *facet* of their Web positioning rather than an attempt to divine what they 'really are'. This is, however, a doubly unscientific methodology in that it uses an individual's judgment, and moreover that individual is involved in the research project and may therefore be accused of conscious or subconscious manipulation of the results. In partial response to this, the full test of classification results are available online for inspection at

http://cybermetrics.wlv.ac.uk/database/uk_2001_page_classification.txt. This includes the subclasses used to facilitate the categorization process. Successful results from this kind of study would give permission to undertake the much bigger project of repeating the exercise with a scientific classification methodology. Based upon the difficulties encountered in other surveys, this is expected to be a complex and time-consuming task (Cronin *et al*., 1998; Crowston & Williams, 2000; Thelwall, 2001a) and perhaps somewhat of a labor of Sisyphus in the long term. One practical problem, however, is that may not be easy to find people that are able to evaluate university Web pages without being aware of the approximate identity of their owners and the reputation of the hosting universities.

After the classifications were complete a program was written to produce new link structure databases for each category used and the results were again processed by programs previously written for the document models and standard link counting programs applied to the new data sets. This provided 56 lists of link counts, which were then entered separately into SPSSPC for the calculation of correlation coefficients. Figure 2 illustrates the whole process

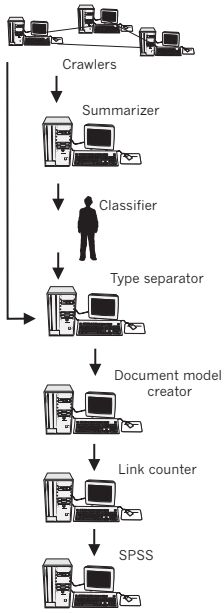

FIG. 2. A diagram to show the sequence of information processing operations after the raw data had been obtained from the crawler database.

5. Results

Correlations between research productivity and link counts are given in tables 3 to 6. The data set did not include a crawl of one of the UK universities, Cranfield, but it was included in the 108 for inlink counts in the first two tables. In the last two tables links from Cranfield are excluded as a result, making the total count 11,449 = 108 X 107 - 107 after excluding counts from a university to itself. The rows of the tables are ordered in terms of decreasing overall link count, and the total numbers are given in Table 7.

TABLE 3. Spearman correlations between university Web site inlink counts and research productivity for 108 UK university institutions. All correlations are significant at the 1% level.

Model	All pages: types 1-4	Own pages: types 2-4	Own academic pages: types 3-4	Own academic standard pages: Type 4
Standard file	0.920	0.916	0.931	0.942
Standard directory	0.925	0.924	0.937	0.946
Range file	0.936	0.937	0.940	0.946
Range directory	0.940	0.940	0.943	0.949
Standard domain	0.923	0.923	0.933	0.935
Range domain	0.886	0.884	0.895	0.883
Standard university	0.807	0.805	0.856	0.869

TABLE 4. Pearson correlations between inlink counts and research productivity for 108 UK university institutions. These are presented as purely descriptive statistics, and comparisons between values are inappropriate due to the non-normality of the data.

Model	All pages: types 1-4	Own pages: types 2-4	Own academic pages: types 3-4	Own academic standard pages: Type 4
Standard file	0.848	0.828	0.820	0.915
Standard directory	0.934	0.936	0.934	0.922
Range file	0.901	0.892	0.881	0.932
Range directory	0.941	0.948	0.947	0.940
Standard domain	0.928	0.929	0.931	0.932
Range domain	0.902	0.902	0.904	0.903
Standard University	0.466	0.463	0.490	0.508

TABLE 5. Spearman correlations for inter-university link counts for 11,449 pairs of different UK institutions from the list of 108 and the product of the institutions' research productivities. All correlations are significant at the 1% level.

Model	All pages: types 1-4	Own pages: types 2-4	Own academic pages: types 3-4	Own academic standard pages: Type 4
Standard file	0.750	0.748	0.754	0.752
Standard directory	0.766	0.764	0.767	0.764
Range file	0.773	0.771	0.772	0.769
Range directory	0.779	0.778	0.777	0.773
Standard domain	0.771	0.769	0.771	0.764
Range domain	0.735	0.734	0.736	0.727
Standard University	0.424	0.427	0.441	0.457

TABLE 6. Pearson correlations for inter-university link counts for 11,449 pairs of different UK institutions from the list of 108 and the product of the institutions' research productivities. These are presented as purely descriptive statistics, and comparisons between values are inappropriate due to the non-normality of the data.

Model	All pages: types 1-4	Own pages: types 2-4	Own academic pages: types 3-4	Own academic standard pages: Type 4
Standard file	0.223	0.208	0.193	0.572
Standard directory	0.739	0.737	0.739	0.746
Range file	0.315	0.291	0.274	0.689
Range directory	0.785	0.794	0.791	0.807
Standard domain	0.833	0.832	0.837	0.834
Range domain	0.786	0.786	0.789	0.788
Standard University	0.186	0.188	0.197	0.208

TABLE 7. Total inter-university link counts for each model across all universities.

Model	All pages: types 1-4	Own pages: types 2-4	Own academic pages: types 3-4	Own academic standard pages: Type 4
Standard file	382,096	355,546	321,206	226,660
Standard directory	235,266	222,107	201,445	172,796
Range file	215,234	198,738	181,698	141,465
Range directory	153,846	146,232	132,980	118,199
Standard domain	99,160	96,442	89,304	80,505
Range domain	54,991	53,854	50,641	46,765
Standard University	9,790	9,747	9,576	9,356

6. Discussion

6.1 Comparison of Models

From tables 3 and 5 it can be seen that in all cases the directory range model produces the highest correlation consistently across the different types of page. The Pearson values in tables 4 and 6 are included as descriptive statistics. It is not valid to draw conclusions from comparisons between Pearson correlations because the data is highly skewed due to both domination by low research productivity values and the existence of link count outliers. It is interesting that the Pearson values show much bigger differences than those from the Spearman formula, however. This is due to the reduction in outliers in the data because even a single large outlier can have a big effect on Pearson values whereas the magnitude of the outlier has much less effect on the rank dependant Spearman.

The general trend in Spearman values is increasing from the file model to the directory range model and then decreasing down to the university model. This can be explained in terms of two competing trends, both caused by increasing aggregation. Firstly, greater aggregation reduces outliers by increasing the level of generality so that most causes of outlier behavior, normally sites with high link or backlink counts, have their multiple links aggregated into one. The competing trend is that with less links the averaging is statistically less effective and subject to increases in variance. The directory range model seems to strike the optimal balance between these two extremes for all types of pages, echoing the findings of Thelwall & Wilkinson (2002) for the first column of data.

6.2 Comparison of Types of Page.

The effect of restricting the counts to the different types of page is not clear cut. For the domain range model there is apparently a contradiction: values increase on the more restricted sets of pages when counting university inlinks but decrease when counting inter-

university links for pairs of institutions, although the changes in both cases are relatively small. This is possibly due to the greater variability of the smaller counts in the latter case than in the former. It is clear, however, that the directory range model on the type 4 pages only is significantly better than the unrestricted case, although both values are very high. There is evidence, then, that a detailed consideration of types of pages can improve link metrics, but it appears that choosing the appropriate document model is much more important.

Examining the entire set of values in tables 3 and 5, it can be seen that the categorization of pages appears to be more effective for the models involving less aggregation (i.e. with lower total link counts). This suggests that the classification procedure captures many cases where there are multiple links between the same directories and domains.

It is interesting that omitting pages hosted by a university but not created by it produces worse results in many cases, indicating that such pages tend to be hosted by institutions with higher research productivity. This is quite a broad category, including some sites unrelated to research. For example a hosted online publisher's site would presumably be unrelated to research but that of a national professional society probably entails some kind of scholarly connection. There appears to be a trend for professional societies to be no longer hosted on subdomains of university sites, but to have their own independent domain name, even if invisibly hosted on the same server. It was noticed during the classification exercise that such renaming had occurred repeatedly, with the target page of many links being a redirection page giving the new domain name. For mirrors of computer documentation this trend is not evident, however, with the reverse being a logical possibility. The increasingly cheap cost of computer storage and increased general knowledgeability about the Internet must be lowering the real cost of hosting large copies of other sites. However, such sites are unlikely to attract extensive links from other universities, since they are not the official source. An exception to this is Imperial College's *Sun SITE Northern Europe* (src.doc.ic.ac.uk), which is an official archive (mirror) site. It is cheaper and presumably quicker to download from Imperial than the USA, at least for UK universities, which may be a factor if it hosts heavily used resources.

Removing non-academic pages produced increases in all cases in Table 3 and at least small increases in all cases except one in Table 5. This is to be expected and corroborates previous indications that links to research related pages correlate more highly with research than those to other pages (Thelwall, 2001a).

Removing high profile pages increases correlations with university inlink counts in all cases except one but decreases between university correlations in all cases except one. Investigations into the raw data failed to unearth an identifiable cause for this, so it is assumed that this is an effect of reduced numbers for the latter case.

7. Conclusions

The results of the categorization process indicates that it is capable of producing small increases in correlations and therefore presumably in the overall reliability of the results. There is also evidence of an overlap between the effects of applying categorization and advanced document models. It can be seen that one does not subsume the other because the best results occur with a combination of the two. Under the assumption that the advanced document models are successful in eliminating outliers, it can be concluded that counts of links to pages with a closer relationship to the research of the host university genuinely associate more strongly with its research productivity than general pages do. Despite this finding, it must be recalled that the type classification exercise was very labor intensive even though it was conducted in an experimental fashion. Much more time would be needed to verify the conclusions in a more scientific manner and to include all link targets, not just the highest inlinked pages and areas. It is estimated that correlations may rise to perhaps between 0.951 and 0.955 if such an exercise were to be undertaken, with the residual variability being attributed to genuine differences in the Web profile of universities' research and their Web use policies.

The future Web researcher now has a variety of choices of link counting methods to call upon depending upon both the type of information that they are hoping to extract and the accuracy that is needed from the source data. The conclusion is not, therefore, that scholars must use the directory range model and classify all pages, but that this is the logical choice when clear-cut results are unlikely to be extracted from anything but the purest Web data available. Such techniques may be required to extract information for smaller sites, or when differential Web use by disciplines (Kling & McKim, 2000) and unusual patterns of online information sharing (van Raan, 2001) would otherwise dominate the calculations.

The approach used in this paper represents a new paradigm in information science research that has been made possible by technological progress over the past ten years and is an example of a genre for which Web-based informal collaboration is extremely useful. The approach can be summed up as one of using enormous computing resources to mine information from a huge repository of low quality (in the sense of fitness for purpose) sources. The original crawling of Web sites used rooms full of idle computers in an intensive (over a month) exercise and had been placed free online because it would be an inefficient use of scholarly resources and Internet bandwidth for every researcher to run their own crawler (Bar-Ilan, 2001). The follow-up work involved further extensive use of computers over extended periods of time in order to repeatedly process the link files to produce the two lots of 28 data sets of 108 or 11,449 values used for the tables above, each one of which was the result of processing one or 107/108 text files. A less visible, but nonetheless crucial aspect of the work was that the availability of cheap fast computing time has allowed this paper to be information science research, focusing on the meaning of the data, whereas ten years ago the processing of similar quantities of data would have been a computer science project focusing on the development of sufficiently efficient software. Including full test runs but excluding the Web crawls and programming time, the total continuous running time to process the data presented in this paper was approximately 400 hours on a 1.3GHz 512 DDR computer.

It is envisaged that future work will include the production of a simple Web interface to the link data sets that will allow any future researcher to test hypotheses about university interlinking, such as geographic ones (Thelwall, 2002d), based upon advanced document models and established type classifications. The ultimate goal is to reduce the technical cost of entry to effective analysis of this kind of data to approximately that which is required for the Institute for Scientific Information's Web of Science.

8. References

Almind, T. C. & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. Journal of Documentation, 53(4) 404-426.

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

Björneborn, L. (2001a). Shared outlinks in webometric co-linkage analysis: a pilot study of bibliographic couplings on researchers' bookmark lists on the Web. Royal School of Library and Information Science.

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

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

Brin, S. & Page, L. (1998). The Anatomy of a large scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1-7), 107-117.

Calvert, P. J. (2001). Scholarly misconduct and misinformation on the World Wide Web, Electronic Library, 19(4), 232-240.

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.

Cooper, W. S. (1969). Is interindexer consistency a hobgoblin? American Documentation, 20, 268-278.

Cresser, F., Gunn, L. & Balme, H. (2001). Women's experiences of on-line e-zine publication, Media culture and society, 23(4), 457-473.

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

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

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

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.

Douyère, M., Soualmia, L.F., Le Duff, F., Thelwall, M. & Darmoni, S.J. (2002, to appear). Web Impact Factor : un outil bibliométrique appliqué aux sites Web des facultés de médecine et des CHU français, Neuvièmes Journées Francophones d'Informatique Médicale. 6-7 mai 2002, Québec-Canada.

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

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

Flake, G. W., Lawrence, S., Giles, C. L. & Coetzee, F. M. (2002). Self-organization and identification of web communities, IEEE Computer, 35, 66-71.

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

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

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

Harter, S. & Ford, C. (2000). Web-based Analysis of E-journal Impact: Approaches, Problems, and Issues, Journal of the American Society for Information Science, 51(13), 1159-76.

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

Jackson, M. S. & Burden, J. P. H. (1999). WWLib-TNG - New Directions in Search Engine Technology. IEE Informatics Colloquium: Lost in the Web - Navigation on the Internet, November 1999, (pp. 10/1-10/8).

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.

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.

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

Koehler, W. (2002). Web page change and persistence - A four-year longitudinal study, Journal of the American Society for Information Science, 53(2), 162-171.

Larson, R.R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. ASIS 96. Available: http://sherlock.berkeley.edu/asis96/asis96.html.

Lawrence, S. L. (2001) Online or invisible? Nature, 411(6837) 521.

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

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

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.

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

Sapienza, F. A. (1999). Communal ethos on a Russian Émigré web site. Javnost, VI(4), 39-52.

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

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

Stafford, T.F. & Stafford, M. R. (2001). Identifying motivations for the use of commercial Web sites. Information Resources Management Journal, 14(1), 22-30.

Tang, R. and Thelwall, M. (2002, to appear). Exploring the pattern of links between Chinese university Web sites, Proceedings of the ASIST Annual Meeting Volume 39 (ASIST 2002).

Teo, T. S. H. (2001). Demographic and motivation variables associated with Internet usage activities. Internet Research: Electronic Networking Applications and Policy, 11(2), 125-137.

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, to appear). A Research and Institutional Size Based Model for National University Web Site Interlinking, Journal of Documentation.

Thelwall, M. (2001c). The top 100 linked pages on UK university Web sites: high backlink counts are not usually directly associated with quality scholarly content, University of Wolverhampton.

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

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

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

Thelwall, M. (2002b, to appear) Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites, Journal of the American Society for Information Science and Technology.

Thelwall, M. (2002c). Methodologies for Crawler Based Web Surveys, Internet Research: Electronic Networking and Applications, 12(2), 124-138.

Thelwall, M. (2002d, to appear). Evidence for the existence of geographic trends in university web site interlinking. Journal of Documentation.

Thelwall, M. & Wilkinson, D. (2002). Three Target Document Range Metrics for University Web Sites, University of Wolverhampton.

van Raan, A. F. J. (2001). Bibliometrics and Internet: Some Observations and Expectations, Scientometrics, 50(1), 59-63.

Vaughan, L. Q. & Hysen, K. (2001). Do Web Link Counts Resemble Citation Counts: An Empirical Examination, University of Western Ontario.

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