International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

VOLUME 8 (2004): ISSUE 1. PAPER 2

A Statistical Analysis of UK Academic Web Links

Nigel Payne
Mike Thelwall

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

E-mail: n.c.payne@wlv.ac.uk
www.scit.wlv.ac.uk/~cm1993/mycv.html

Abstract

The analysis of web-based documents using quantitative techniques is a well-established area of research within the realm of information science. This paper builds on some of that work and presents the results of statistical analysis carried out on the web link structure text files of 111 UK universities. Summary statistics are produced using Alternative Document Models and the results of the statistical analysis are also graphically displayed, including trendline equations. Mathematical linear relationships were observed between certain bivariate data with subsequent Pearson correlation analysis revealing a number of very strong correlation relationships, particularly between site size and number of source / target directories and pages. This seems to support previous research by suggesting that the directory Alternative Document Model has some advantages over the domain Alternative Document Model.

Keywords

Academic, Hyperlink, Web link

Introduction

The Internet is expanding at an ever increasing rate and, as the amount of information available increases, computer scientists have realised that in order to develop new methods and techniques with which to analyse the patterns emerging from this complex network, it may prove beneficial to apply known theory from other disciplines, such as statistics or information science.

This paper hopes to show the relevance of using statistical methods to produce quantitative data in order to improve information systems use with reference to one of the original functions of the web, the interlinking of academic research. This paper uses metrics based on the university web links to help capture hidden statistical information about the relationships between the links and the domains/directories to which they refer, and may also

provide a valuable insight into the intrinsic patterns of academic link structures. At present, this type of study can provide summary statistics that can highlight topological and power law relationships within link structures.

The analysis of web-based documents using quantitative techniques is a well established area of research within the realm of information science. Much research has already been carried out, utilising informetric methods, following claims that search tools are not sufficiently developed to produce accurate, reliable results. Research has been conducted from both a bibliometrical (Rousseau, 1997) and a more mathematical (Broder et al., 2000) perspective, and the analogy between citations and hyperlinks has led to a great deal of study, including the creation of a metric, the Web Impact Factor based upon counts of inlinking pages (Ingwersen, 1998).

Against this background of research, it seems natural to investigate whether or not it is possible to extract useful information from the web regarding its effectiveness in supporting scholarly activity. There is now a considerable body of evidence to show that patterns of web linking between universities can be strongly associated with research productivity (Thelwall & Harries, 2003) and an association between links and geographic distance has also been demonstrated (Thelwall, 2002b).

The aim of this paper is to conduct a statistical analysis of UK academic web links with a view to identifying general mathematical patterns or relationships. The paper aims to make extensive use of the Alternative Document Models (ADMs) produced as a result of research carried out by Thelwall (2002a) and Thelwall & Wilkinson (2003).

PREVIOUS RELATED RESEARCH

Early web hyperlink research involving bibliometrical approaches includes Larson's exploratory analysis of the intellectual structure of cyberspace (Larson, 1996). This initial research has since been followed by other information science research showing that useful information about individual web pages and websites can be extracted directly from link structures (Kleinberg, 1999) and that indeed, the hyperlinks themselves can be studied as objects of interest in their own right (Broder et al., 2000). Björneborn & Ingwersen (2001) have gone on to propose new web methods based upon bibliometric methodologies and current graph theory.

It quickly became apparent that, although the study of web links was a relatively new area of research, it was one that had generated a large amount of interest and speculation. The very focused study of UK academic web links is part of a more general study of the Internet hyperlink structure as a whole, with comparisons being made with bibliometric citations.

Among the previously published studies dealing specifically with the hyperlink structure within UK universities, two in particular concentrate on the links between pairs of universities. The first, although not presenting clear evidence of an overall trend, did highlight some implications for the apparent geographic grouping of UK academic institutions, particularly with respect to the Scottish and Manchester universities (Thelwall, 2002c). The other study gave mathematical evidence to support the proposal 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, 2002d). In addition to this, Chen et al. (1998) concentrated their attention on counting links between computer science department websites in Scottish universities using pathfinder network diagrams. The end results produced did reflect the profiles of the individual universities, but the survey was limited by the small sample size and a lack of variety in the institutions under study.

This paper aims to extract significant mathematical patterns from the hyperlink structure of UK academic web pages and, fundamental to this objective is the definition of the web document, which should comprise the single indissoluble unit of coherent material. Until recently, all previous web link studies had used the web page as the primary source document for counting purposes. Thelwall (2002a) presented arguments in an attempt to explain why this is not necessarily ideal and why other alternatives, specifically ADMs, have the potential to produce better results. This is despite the fact that individual web pages are often the only choice if search engines are used for raw data, the logical choice of primary web documentation, and are by far the easiest basic web unit to identify and manage.

The ADMs discussed aggregate source and target pages at the web page, directory and domain level, using the following definitions:

> **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. 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.
>
> **Directory**: All HTML files in the same directory are treated as a document. All 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**: As above except all HTML files with the same domain name are treated as a single document. This clusters together all pages hosted by a single subdomain of a university site. Domains are obtained by stripping any directory structure, file name, port number and password information from URLs, i.e. truncating each target URL just before the first slash it contained, if one was present.

It has been discovered that the domain and directory models were able to successfully reduce the impact of anomalous linking behaviour between pairs of websites, with the latter being the method of choice. Further to this, Thelwall & Wilkinson (2003) goes on to state 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 study of web links between university websites is at an early stage, with a great deal of research recently becoming available and even more currently in progress. It would appear to be an area of great interest and, based upon the variety of the content (Middleton et al., 1999) and suggestions of multiple underlying trends in the patterns of links between individual institutions (Thelwall, 2002c), there would appear to be a great deal of complexity and variety yet to be discovered. The result so far show that meaningful information can be extracted from large scale comparisons of web links between academic institutions, although the results are far from reliable at the individual university level, making the interpretation of link counts highly problematic and further research necessary.

RESEARCH QUESTION

This research is fundamentally important because a greater understanding of the mathematical patterns and relationships within the hyperlink structure of the Internet will develop an appreciation of the way the web is connected. It may be used by search engines in an attempt to increase their performance and accuracy, and may prove to be a useful tool in predicting future

development and evolution.

The themes that emerge are valuable in a number of respects. Analysis of the link structure of the web suggests that the on-going process of page creation and linkage, while very difficult to understand at a local level, results in structure that is considerably more orderly than is typically assumed. Thus, it gives a global understanding of the ways in which independent users build connections to one another, and it provides a basis for predicting the way in which on-line communities will develop as they become increasingly connected.

This paper carries out research by using a specially produced program, which calculates and processes statistics such as the number of different pages, directories and domains in the source universities. Descriptive statistical analysis and Pearson correlation analysis is then carried out on the results of the program with a view to answering the following research question:

Which mathematical models best characterise link structure relationships between UK university websites?

METHODS

This paper reports on a statistical analysis of UK academic web links, with a view to identifying mathematical patterns within the hyperlink structure. A specialist web crawler designed by Wolverhampton University's Statistical Cybermetrics Research Group initially collected the raw data (Thelwall, 2001), and produced text files of 111 UK universities as of July 2002, comprising a publicly available indexable database (Database 9) as part of the Wolverhampton University Academic Web Link Database Project (Thelwall, 2002/3). These text files contain the source and target links for each of the 111 UK Universities in the form of HTML-embedded URLs. They do not list URLs identified by embedded programs such as JavaScript, or non-HTML documents types with a hyperlinking feature such as PDF documents accessed over the web, but do include dynamically generated web pages e.g. ASP. Several papers have already been published based on results produced from research carried out using this information.

UK universities were chosen as the area of study for this research for the following reasons:

- There is a current, comprehensive data set provided by Database 9 of the Wolverhampton University Academic Weblink Database Project
- The academic web is relatively mature
- The size of the body of UK universities, at 111, is manageable, yet large enough for meaningful statistical measures to be taken
- Academic websites provide an opportunity for close comparison with academic research, if so required

One particular theme that cropped up time and time again was the proven existence of mathematical patterns within the structure of the web and the decision was taken to concentrate some effort towards identifying this phenomenon, if it existed, within the raw data produced by the Wolverhampton University's Statistical Cybermetrics Research Group web crawler.

With this in mind, the program deliverable was designed to perform two distinct functions using ADMs. Firstly, to process the raw data stored in the 111 UK university web links files and produce statistics on:

- The number of source domains in each university file

· The number of target domains in each university file
· The number of source directories in each university file
· The number of target directories in each university file
· The number of source pages in each university file
· The number of target pages in each university file
· The total number of pages (all URLs) in each university file

And secondly, the program would output a graphical display comparing the statistics derived above in a specific attempt to isolate any mathematical relational behaviour.

RESULTS

Once the statistical analyser side of the program was finalised, the results for the 111 UK universities were calculated and stored. As the project was primarily concerned with identifying linear and power law relationships amongst this data, all charts were initially displayed using linear scales, with a linear trendline added. These results were then graphically displayed by the program using the following 6 charts:

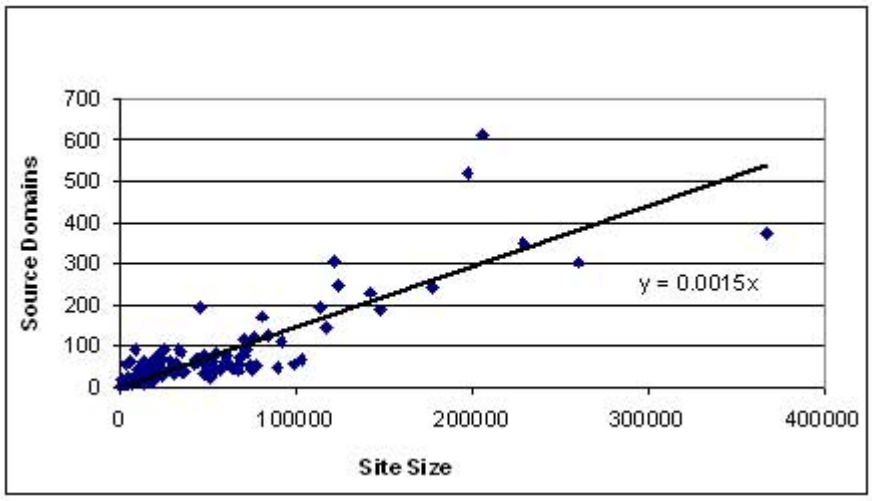

Figure 1: Linear Chart of Source Domains vs Site Size

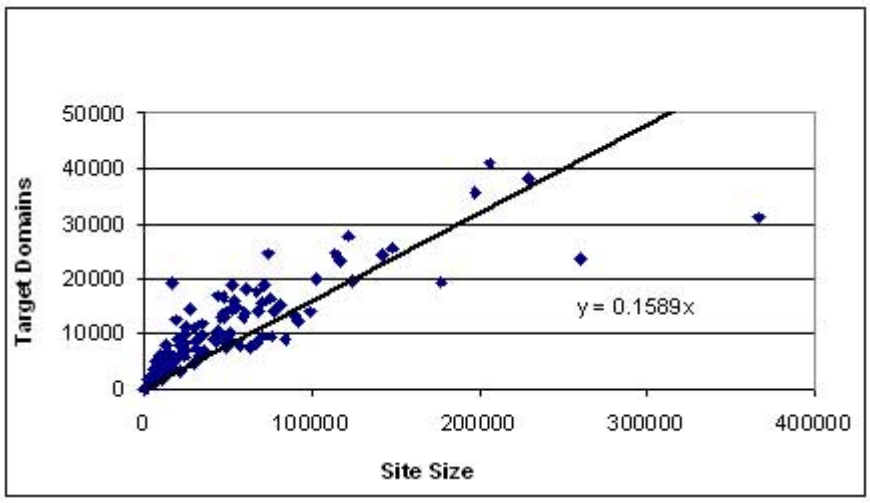

Figure 2: Linear Chart of Target Domains vs Site Size

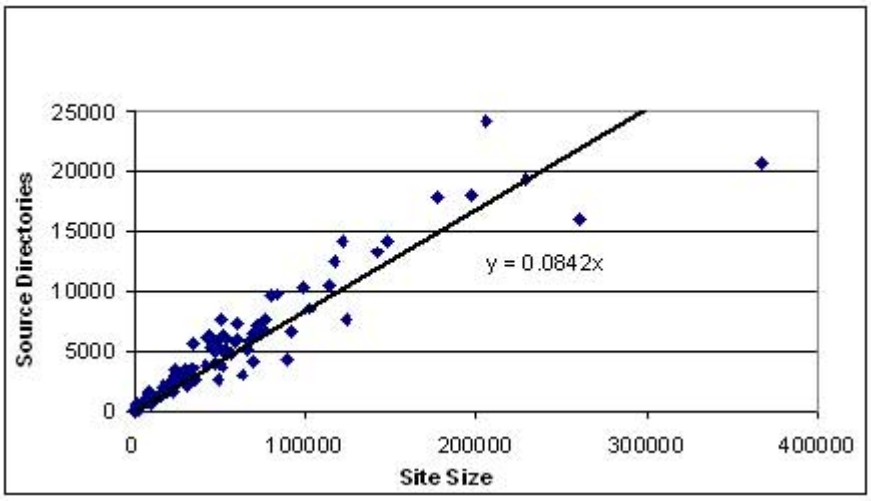

Figure 3: Linear Chart of Source Directories vs Site Size

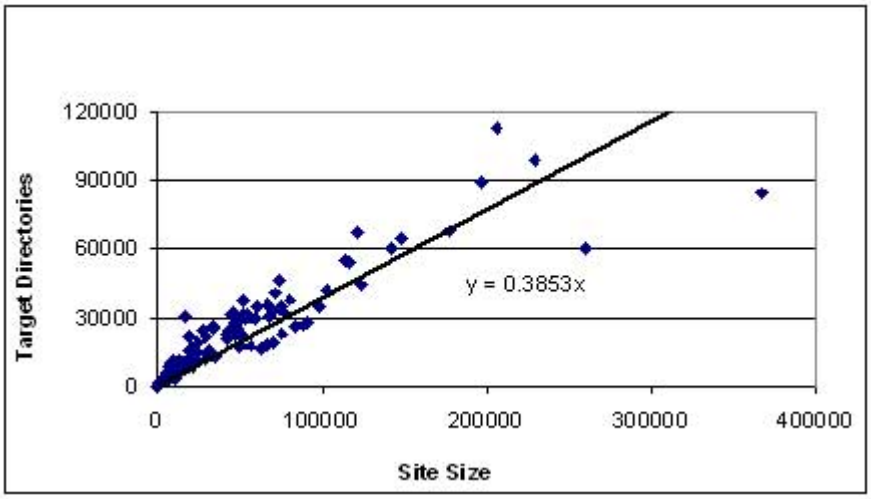

Figure 4: Linear Chart of Target Directories vs Site Size

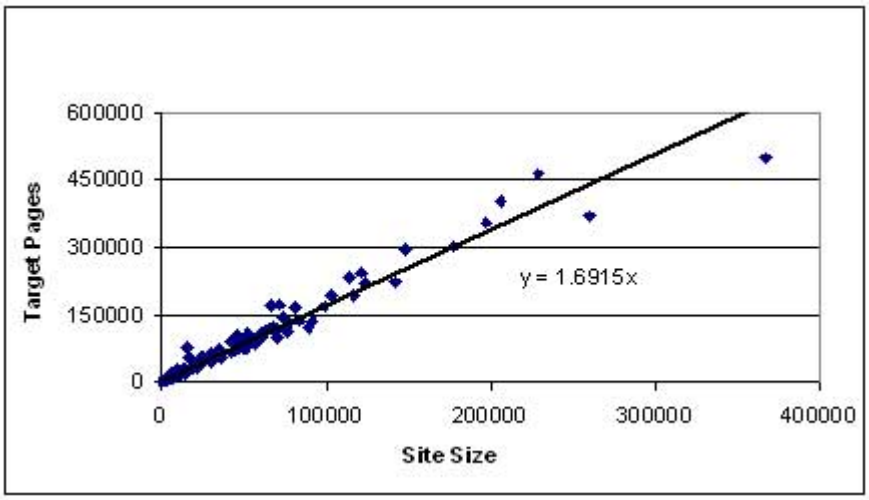

Figure 5: Linear Chart of Target Pages vs Site Size

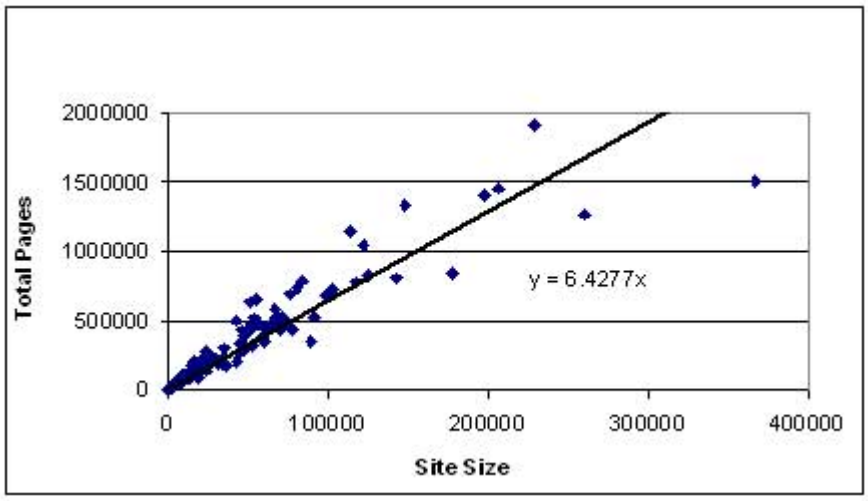

Figure 6: Linear Chart of Total Pages vs Site Size

In creating the above charts, site size is defined to be the number of source pages with all duplicates removed. (Duplicates were also removed for the number of Target Pages, but Total Pages is taken to be all URLs contained within the text file). Further to this, it is assumed that the y-intercept is 0 i.e. when site size is zero, the corresponding number of pages, directories and domains is zero.

From these graphs, we can see that there is strong evidence of a linear relationship, particularly between the site size and page / directory models, and a decision was made to perform further analysis on these charts, and to concentrate on power law identification and levels of correlation.

With this in mind, the statistical output on two quantitative variables of the program deliverable were plotted on a scatter plot using log-log axis in an attempt to determine whether any power law relationship existed between the bivariate data. A power type trendline and the Coefficient of Determination, R2 was added, and the results were then graphically displayed using the following charts:

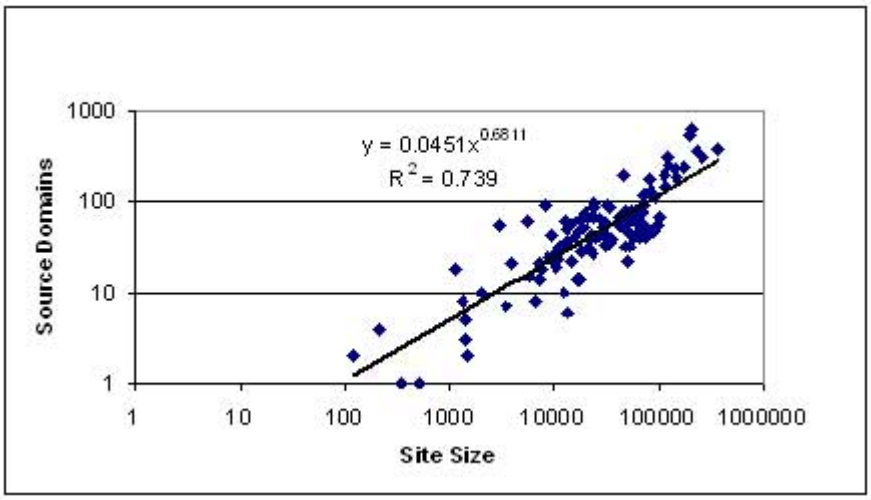

Figure 7: Logarithmic Chart of Source Domains vs Site Size

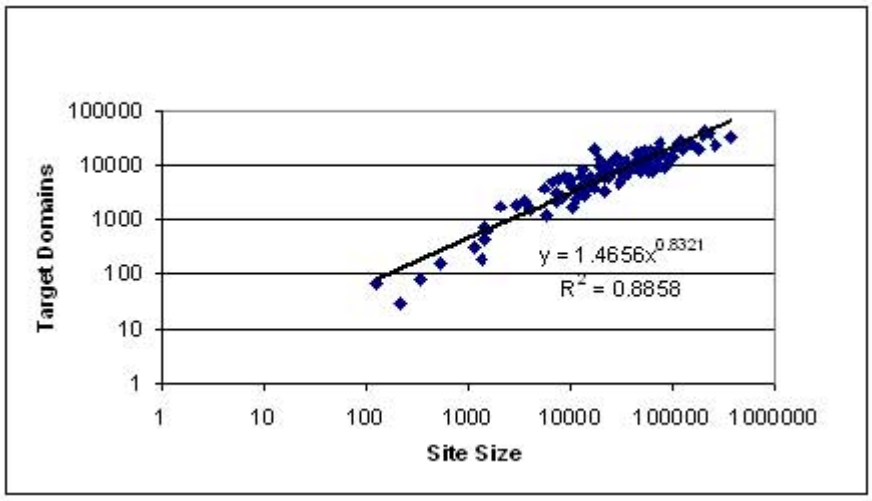

Figure 8: Logarithmic Chart of Target Domains vs Site Size

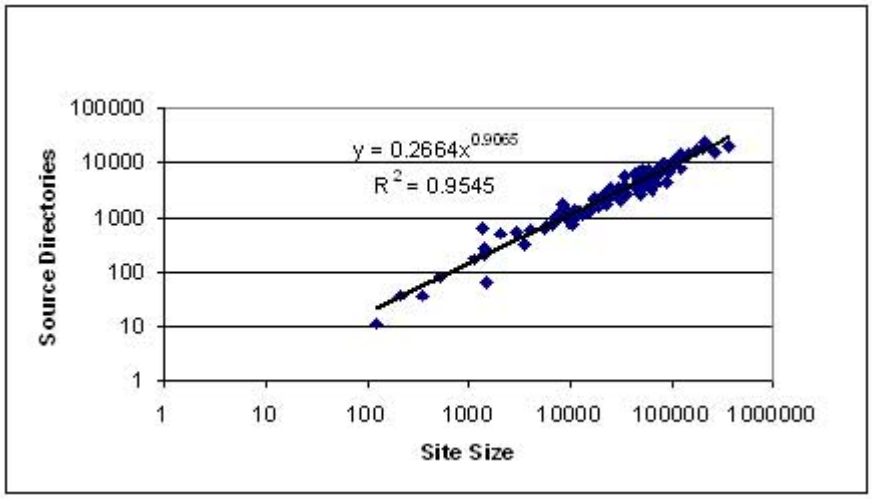

Figure 9: Logarithmic Chart of Source Directories vs Site Size

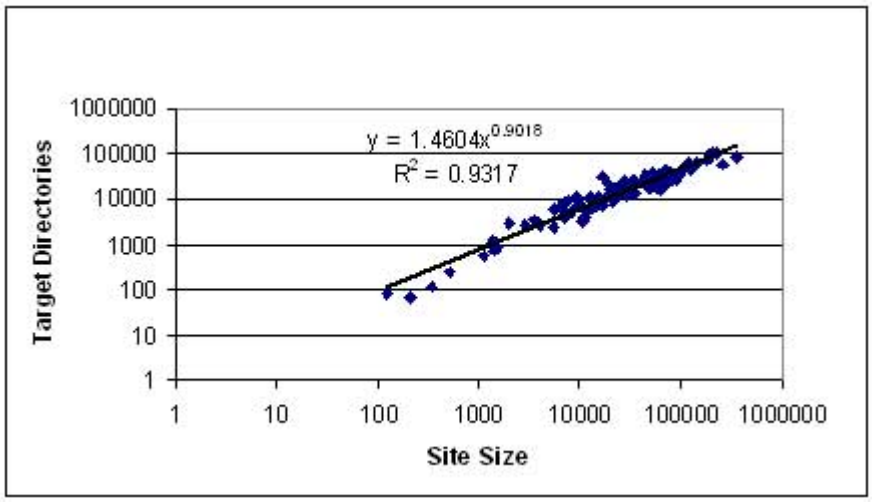

Figure 10: Logarithmic Chart of Target Directories vs Site Size

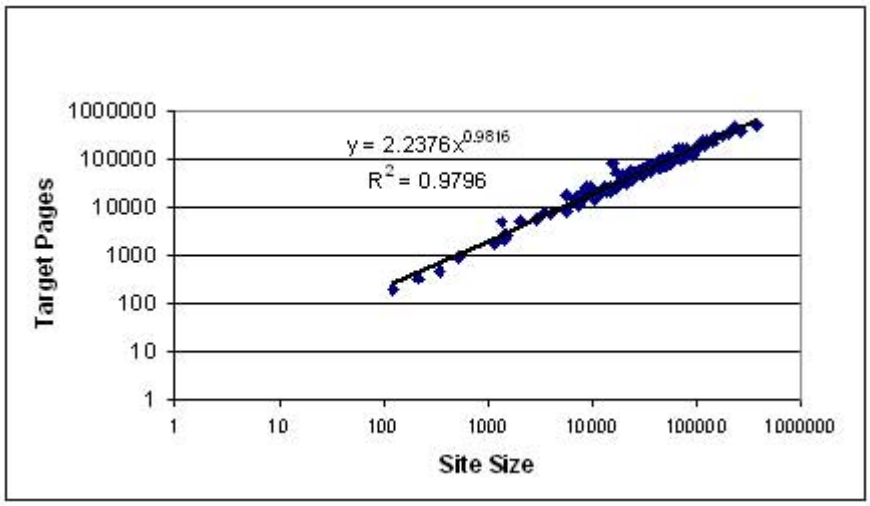

Figure 11: Logarithmic Chart of Target Pages vs Site Size

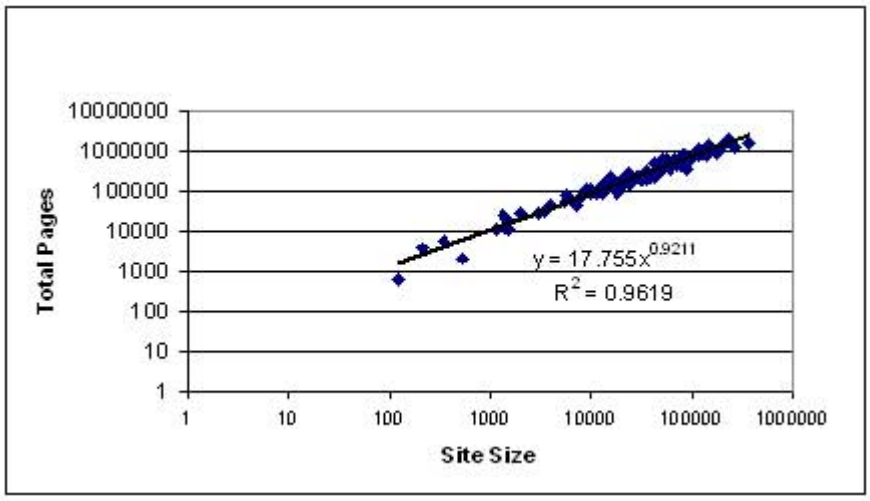

Figure 12: Logarithmic Chart of Total Pages vs Site Size

The equation of the power type trendline was displayed in the form y = cxa. That is, 'y' is a power law in 'x' with a power or index of 'a' multiplied by a number or normalisation constant 'c'.

Now, as we are using logarithmic scales, we can manipulate this equation using the Laws of Logarithms to give us the straight line equation of our trendline as follows:

$$y = cx^a$$
$$\therefore \log y = \log (cx^a)$$
$$= \log c + \log x^a$$
$$= a\log x + \log c$$

Thus our power equation has been reduced to linear form whereby the graph of log y versus log x gives us a straight line that has a slope identical to the power or index of the power law and y-intercept 'c'.

In this case, it was discovered that the value of 'a' was very close to 1 for all six graphs, indicating amongst other things, that the two variables were positively linearly related, with the trendline gradient of almost 1.

It was then decided to use the Pearson Product Moment Correlation Coefficient, given by the formula:

$$R = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

and, more specifically, the Coefficient of Determination, R2.

R2 close to 1 would imply that the model is explaining most of the variation in the dependant variable (site size) and may prove to be a very useful model. R2 close to 0 would imply that the model is explaining little of the variation in the dependant variable and may not be a very useful model.

The correlation coefficients are summarised as follows:

CHART	CORRELATION COEFFICIENT
Site Size / Source Domains	0.739
Site Size / Target Domains	0.8858
Site Size / Source Directories	0.9545
Site Size / Target Directories	0.9317
Site Size / Target Pages	0.9796
Site Size / Total Pages	0.9619

We can see that, although the source and target domain charts display a high level of correlation (0.7 to 0.9), the directory and page charts exhibit very strong correlation (0.9 to 1). The very strong correlation between site size and target pages and site size and total pages may be expected as site size is defined to be the number of source pages with all duplicates removed but the results seem to suggest a definite relationship between the size of the site and the number of source directories, target directories, target pages and total pages.

However, the higher level of directory correlation compared to the level of domain correlation, in both the source and target links, may support the evidence gained by Thelwall (2002a) and Thelwall & Wilkinson (2003) in that the directory model seems to be have some advantage over the domain model from the perspective of counting links. However, it is vital to remember that a correlation, even a very strong one, does not mean that we should immediately jump to conclusions about causation. We should always be aware that the correlation in itself is no proof of assertion.

Outliers

The study of outliers is important because the accurate identification of anomalies is a precondition to more successful analysis of this kind of data.

The following universities were noted as outliers for the corresponding charts:

Site Size / Source Domains	Paisley, Chichester, Abertay, Newport, London Guildhall
Site Size / Target Domains	Luton, Glasgow, Birmingham, Keele
Site Size / Source Directories	Gloucestershire, Luton
Site Size / Target Directories	Chichester, Keele, Birmingham, Glasgow
Site Size / Target Pages	Luton, Hull

Site Size / Total Pages	Paisley, Harper-Adams, Chichester, Surrey Institute of Art & Design

The reoccurring nature of several university names warranted further investigation. In particular:

University of Luton. This university has a large number of target domains / pages and source directories for a relatively small site size, due partly to the variation in the domain structure within the university (e.g. conservatoire.uce.ac.uk and students-union.uce.ac.uk).

University of Chichester. Although not the smallest university in terms of site size, it does have the smallest number of both target domains and directories. The large number of internal links is explained by the design of its web page menu structure.

University of Paisley. This is not only the smallest university in terms of site size but also has the smallest number of source directories, source and target pages. It appears as the furthest left point on the corresponding charts.

University of Glasgow. This is at the opposite end of the site size scale to Paisley, (appearing to the right-hand side of the charts), being the second largest site size behind Edinburgh. The fact that the two largest sites are both Scottish universities may warrant further investigation.

CONCLUSIONS

This paper is seen to have fulfilled the original aim of conducting a statistical analysis of UK academic web links with a view to identifying general mathematical patterns or relationships. It has used raw data produced by the Wolverhampton University's Statistical Cybermetrics Research Group, and built on previous work carried out by researchers using ADMs on academic websites.

The statistical analyser program processed the link structure text files of 111 UK universities and produced a graphical display based on the results. It was apparent from the graphical display that linear relationships were clearly in evidence in the link structure of the university websites. Additionally, there did appear to be very strong levels of correlation when using the Coefficient of Determination, R2, particularly among the directory and page models.

Also, based upon the theoretical analysis and the quantitative study of the statistics, particularly for the counts of links between site size and source/target domains and directories, there seems to be evidence that the directory ADM is considerably better than the domain ADM from the perspective of analysing web link structure and this appears to support research carried out by Thelwall (2002a) and Thelwall & Wilkinson (2003). A weakness in the methodology, however, is that 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.

Although it seems likely that these results would be generally applicable, it is not inconceivable that there would be countries to which it would not apply, for example if a URL, directory or domain structure was commonly used that was substantially different to that used in the UK. That said, the very high correlation found for the directory model does encourage the belief that it may well be robust enough to stand transportation to other countries.

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, 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.

It is evident then, that the reality of hyperlink analysis is that it is a complex and problematical tool. Although patterns can be extracted from hyperlinks, it is still the case that they are a largely unregulated phenomenon. As a result great care must be taken to validate data when conducting hyperlink analyses to avoid drawing false conclusions because of data unreliability, although in our case the expected mathematical patterns defining a relationship between site size and pages, directories and domains appeared evident. Nevertheless, the importance of the inherent properties of web links means that the type of research described within this paper looks set to have a promising future. In addition, these positive results strengthen the case for using web link analysis as a tool with the potential to reveal underlying trends in academic website interlinking.

References

BJORNEBORN, L. & INGWERSEN, P. (2001) Perspectives of Webometrics. Scientometrics, 50(1), pp. 65-82.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A. & WIENER, J. (2000) Graph Structure in the Web. Journal of Computer Networks, 33(1-6), pp. 309-320.

CHEN, C., NEWMAN, J., NEWMAN, R. & RADA, R. (1998) How did University Departments Interweave the Web: A Study of Connectivity and Underlying Factors. Interacting with Computers, 10, pp. 353-373.

INGWERSEN, P. (1998) The Calculation of Web Impact Factors. Journal of Documentation, 54(2), pp. 236-243.

KLEINBERG, J. (1999) Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5), pp. 604-632.

LARSON, R. (1996) Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace. ASIS 96. <**http://sherlock.berkeley.edu/ asis96/asis96.html**> (updated on 29 July 1996, accessed on 19 April 2003).

MIDDLETON, I., McCONNELL, M, & DAVIDSON, G. (1999) Presenting a Model for the Structure and Content of a University World Wide Website. Journal of Information Science, 25(3), pp.219-227.

ROUSSEAU, R. (1997) Sitations: an Exploratory Study. Cybermetrics, 1(1), paper 1. <**http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html**> (updated on 20 November 1997, accessed on 13 August 2003).

THELWALL, M. (2001). A Web Crawler Design for Data Mining. Journal of Information Science, 27(5), pp. 319-325.

THELWALL, M. (2002a) Conceptualising Documentation on the Web: An Evaluation of Different Heuristic-Based Models for Counting Links Between University Websites. Journal of the American Society for Information Science and Technology, 53(12), pp. 995-1005.

THELWALL, M. (2002b) Evidence for the Existence of Geographic Trends in University Website Interlinking. Journal of Documentation, 58(5), pp. 563-574.

THELWALL, M. (2002c) An Initial Exploration of the Link Relationship Between UK University Websites. ASLIB Proceedings, 54(2), pp. 118-126.

THELWALL, M. (2002d) A Research and Institutional Size Based Model for National University Website Interlinking. Journal of Documentation, 58(6), pp. 683-694.

THELWALL, M. (2002/3). A free database of university web links: data collection issues. Cybermetrics, 6/7(1), paper 2. <**http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.html**>

THELWALL, M. & HARRIES, G. (2003) 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. Journal of the American Society for Information Science and Technology. 54(7), pp. 594-602.

THELWALL, M. & WILKINSON, D. (2003) Three Target Document Range Metrics for University Websites. Journal of the American Society for Information Science and Technology. 54(6), pp. 489-496.

Received 22/April/2004
Accepted 29/June/2004