Finding Similar Academic Web Sites with Links, Bibliometric Couplings and Colinks

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

David Wilkinson
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: d.wilkinson@wlv.ac.uk
Tel: +44 1902 *321452* Fax: +44 1902 321478

A common task in both Webmetrics and Web information retrieval is to identify a set of Web pages or sites that are similar in content. In this paper we assess the extent to which links, colinks and couplings can be used to identify similar Web sites. As an experiment, a random sample of 500 pairs of domains from the UK academic Web were taken and human assessments of site similarity, based upon content type, were compared against ratings for the three concepts. The results show that using a combination of all three gives the highest probability of identifying similar sites, but surprisingly this was only a marginal improvement over using links alone. Another unexpected result was that high values for either colink counts or couplings were associated with only a small increased likelihood of similarity. The principal advantage of using couplings and colinks was found to be greater coverage in terms of a much larger number of pairs of sites being connected by these measures, instead of increased probability of similarity. In information retrieval terminology, this is improved recall rather than improved precision.

Keywords: Document clustering, web metrics, web information retrieval.

Introduction

A growing area of research in information science is the analysis of Web based documents, often using quantitative techniques (Almind & Ingwersen, 1997; Aguillo, 1998; Cronin, 2001; Borgman & Furner, 2002), commonly known as Webmetrics. Much of the work in this area is focussed on Web links, motivated by both bibliometrics (Rousseau, 1997; Ingwersen, 1998) and computer science with graph theory (Broder *et al*., 2000; Björneborn, 2001b). Results have shown that university Web site links are influenced by a combination of geographic (Thelwall, 2002a) and research (Thelwall, 2001a) factors, showing that patterns can be mined from this kind of link data. Very simple mapping techniques have also been applied to visualise the flow of information between national educational systems (Thelwall & Smith, 2002) and other identifiable areas of the Web (Thelwall, 2001b). Colinks (see below) have also been used to map patterns of interlinking between universities in Europe (Polanco *et al*., 2001). One trend in academic Webmetrics research is to increasingly focus on smaller units of study based around a particular discipline in a

Information Processing & Management, to appear.

country (Chen *et al*., 1998; Thomas & Willett, 2000; Chu *et al*., 2002; Li *et al*., 2003; Tang & Thelwall, 2003). Such studies face an immediate problem: that of finding all the relevant pages from amongst the millions of potential candidates. Similar problems are faced by other researchers exploiting Web links (Garrido & Halavais, 2003; Park *et al*., 2002; Rogers, 2002). For example, if biology pages in UK universities were to be analysed, then a starting point for finding them would be to identify the domain names of all UK biology departments. But these may have multiple domain names and share domains with other departments, and members of the department may have separate sites for research groups, personal projects or personal home pages. Identifying all these would be extremely labour-intensive and almost certainly error-prone. There is, therefore, a need to develop automatic aids for this process.

The task of identifying similar pages on the Web is an important related issue in information retrieval. For example, many search engines have a "find related pages" link next to matching URLs (Arasu *et al*., 2001). Kleinberg's (1999) topic distillation algorithm is designed to identify groups of similar pages, as is the community identification algorithm of Flake *et al*. (2002). The underlying reason is that it is easier to estimate the importance of a page and the fit of its semantic content with an information request when it is interpreted in the context of its peers. The purely content based vector space model (Salton & McGill, 1983) has problems with issues such as semantic disambiguation that context can help to solve. Recent research in information science, however, has shown that the Web page is not necessarily the best unit for counting Web links, as used in all the above grouping techniques. For example, when counting links between UK university Web sites, grouping all pages in a single domain into a single conceptual unit gives results that best match an external evaluation criterion (Thelwall, 2002b). The trend to aggregate in domains or sites is also evident in search engines such as Google, which often return two results per site: the home page and a closely related other page inside the site. It is therefore natural to ask whether improved IR results can be obtained by clustering at the domain level in addition to the page level, a parallel motivation to that of Webmetrics for addressing the same issue.

This paper reports on an investigation into the efficacy of domain-based similarity clustering based upon two measures from bibliometrics: bibliometric coupling and colink counting. The (bibliometric) *coupling* count of a pair of Web domains is the number of domains that they both link to, i.e. there is at least one Web page in each of the two source domains that contains a link to at least one page in the target domain (not necessarily the same page). The *colink* count of a pair of Web domains is similarly the number of domains that link to both of them (Björneborn, 2001a). We conjecture that both metrics can be used to help differentiate between similar and dissimilar domains. Successful results would legitimate their exploitation in software applications for Webmetrics and Web information retrieval.

Related Research

Computer Science

Google's seminal PageRank (Brin & Page, 1998) algorithm introduced Web links as an important component of search engine ranking and since then they have become a standard tool, used in many different ways to improve both precision and recall in search results (Pirolli *et al*., 1996; Haveliwala *et al*., 2000; Arasu *et al*., 2001; Gao *et al*., 2001). PageRank does not work by grouping pages together, instead

it employs the principle that highly linked to pages are likely to be useful, especially if pages linking to them are also highly linked to. Kleinberg's (1999) algorithm uses links and page text in order to group semantically related pages and identify both pages that link to many relevant pages (hubs) and those that are linked to by relevant pages (authorities). This algorithm is unstable with respect to the removal of a small number of links (Ng *et al*., 2001a) an important consideration for clustering algorithms, but a stable variation has been proposed, borrowing from PageRank (Ng *et al*., 2001b). The fundamental problem with attempting to separate the Web into clusters based upon its link structure and any kind of prescriptive definition of a cluster is that the size of the Web means that any standard approach based upon testing combinations would be computationally infeasible. Flake *et al*. (2002) circumvented this problem by developing a heuristic-based algorithm that works by iteratively extending any seed community of pages into a (possibly) larger one that is reasonably well interconnected. The advantage of this over the Kleinberg approach is its independence of page contents. For information retrieval purposes semantics can then be used in an additional filtering step. Another fundamental problem is that the Web is extremely well interconnected (Broder *et al*., 2000) and although links tend to be between pages with similar contents, a proportion of links jump to unrelated material (Watts & Strogatz, 1999; Chakrabarti *et al*., 2002), complicating the topic identification process unless algorithms are used that allow for a small number of irrelevant links. The Flake approach has been adapted for the purpose of automatically identifying communities from academic Webs (Thelwall, 2003) and this represents an alternative feasible approach for the problem addressed in this paper.

Bibliometrics

In bibliometrics, a range of metrics have been used as devices to automatically assess the similarity of documents or collections of documents. Proximity mapping and clustering are common tasks in information science in order to map a discipline (White & Griffith, 1982) a set of journals (Cawkell, 2000) countries (Glänzel & Schubert, 2001) or even the whole of science (Small, 1999). Various techniques have been used, mainly to estimate the relative proximities of all pairs of entities involved. Some common techniques for this are multi-dimensional scaling, triangulated drawing (Small, 1999), Salton's measure (Glänzel & Schubert, 2001), pathfinder network scaling (Schvaneveldt *et al.*, 1989) and variations (Chen, 1999). Alternatively, clustering has been achieved with factor analysis, cluster analysis and simple thresholding (Small, 1997). There are also several sources for the raw similarity data. Citation counting appears to be the original, presumably tracking the flow of information through the literature. Co-citations were proposed as an alternative measure (White, 1973) under the assumption that two authors or articles that are frequently cited together are likely to have something important in common. In the complementary measure of bibliographic coupling, number of documents cited by both is counted. Small (1997) claims improved clustering performance through the use of a combination of these, also adding indirect referencing through an intermediate journal article. Although the purpose of most of these techniques is to provide an easily assimilated visualisation of a complex area of scientific endeavour through its literature, they have also been explicitly cited as motivation for computer science link based methods.

The Research Question

The question to be addressed is whether colinks and bibliometric coupling can be used as a more effective way to identify similar academic Web sites than either random choices or direct links. We will use unique domain names as a proxy for academic sites. This is an oversimplification, but a necessary one in the absence of an effective automated method for identifying coherent sites on Web. In order to assess this question we choose a discipline-based notion of similarity. Clearly there are other potential candidates as a basis for assessing similarity, such as by genre type or owner type, but this one directly addresses the motivation for the study. We will also go further than simply assessing whether colinks and couplings are useful and attempt to find the optimal combination to be used.

The specific questions are as follows, concerning identifying similar academic domains by discipline type.

- Is the existence of a bibliometric coupling or co-link between a pair of sites a more reliable indicator of similarity or identity than a link?
- What is the optimal combination of co-links, links and couplings to identify similar and identical sites?
- Are substantially larger numbers of pairs of similar sites related by colinks and/or couplings than by direct links?
- Are higher absolute or relative coupling or co-link values better indicators of similarity?

Methodology

The question will be addressed with respect to one academic area: the UK in 2002. This is a relatively mature Web-using area. The link structure was obtained from a publicly available database (Thelwall, 2001a) created by a specialist information science Web crawler (Thelwall, 2001b). It contains information on all Web pages that can be found by following HTML links from the university home page. This is a limitation but a necessary one for Web research (Thelwall, 2002a).

A previously written program, also publicly available from the same source as the link database, was used to convert the links to being based upon domain names alone, and to exclude links between domains within the same university. The resultant structure is a collection of 6,754 domains and a record of all the other UK university domains that each domain links to. A link is counted between two domains if both of the following hold.

- The domains are hosted by different universities.
- At least one page from the first domain contains at least one link to any page in the second domain.

In line with the document model concept (Thelwall, 2002b), no record was kept of how many links there are between domains because multiple links are frequently caused for spurious reasons (Thelwall, 2003). A new program was then written to calculate co-links and bibliometric couplings for this structure, $(6{,}754^2 - 6{,}754)/2 = 22{,}804{,}881$ values for each one. This excludes a site matched with itself and identical pairs in reverse order. A random number generator was used to select ordered pairs of domains, selecting 300 from each of eight categories comprising all combinations of:

- Coupling count > 0 or coupling count = 0
- Colink count > 0 or colink count = 0
- Link between domains in either direction or no link between domains

These were then entered into a spreadsheet and their order randomised. The spreadsheet was then set up so that the three key values for each pair of domains were hidden, but not the domains themselves so that the author could assess the similarity of the domains without knowledge of their linking categories. The first 500 were categorised.

A pair of domains were classed as similar if they belonged to the same UK Research Assessment Exercise (RAE, http://www.rae.ac.uk) subject area and identical if they appeared to concern precisely the same topic, or were the main domains of organisations with the same or heavily overlapping RAE category subject coverage. Almost all of the latter matching pairs were departmental Web sites. Some examples of Web sites judged identical are given below. Note that none of them are precisely the same, perhaps reflecting government policy for diversity in higher education (Dearing, 1997).

- Department of Mathematical Sciences / Mathematics and Statistics Department
- Department of Electronics and Computer Science / School of Electrical and Electronic Engineering
- School of the Biological Sciences / School of Biochemistry and Molecular Biology

Some examples of similar Web sites are given below.

- Division of Informatics & Artificial Intelligence / The School of Mathematics
- Biochemistry Department / Crystallography and Bioinformatics groups
- Kids on the Net / SchoolsNet UK
- Department of Mathematics / Faculty of Maths and Computing
- Computing Department / Department of Computer Science unofficial pages.
- Careers site / transferable skills site.

Some examples of pairs of sites that were not judged similar.

- Computer Science Department/ Electronics Research Group, Engineering Department.
- University of Oxford Clinical School / Biochemistry Department

The new program was also adapted to produce descriptive statistics on the eight categories and the range of co-link coupling values.

Results

Summary Statistics

From Table 1 it can be seen that most pairs of domains, over 93%, were not associated together in any of the three ways by links. In fact only a very small percentage of domains were directly connected by links in either direction, 0.07% in total. Much higher percentages were joined by colinks or couplings. Figure 1 shows that the distribution of the raw colink count values follows an approximate power law. The graph is almost linear due to the logarithmic scales on the axes, but it shows that the number of sites with a given colink count rapidly decreases as the count increases. In fact many Web link related phenomena follow power laws (Broder *et al*., 2000; Pennock *et al*., 2002). A similar pattern was also present for coupling. As a consequence, relatively few pairs of sites had colink or coupling values bigger than 1.

Figure 1. The distribution of colink values for pairs of domains in the data set.

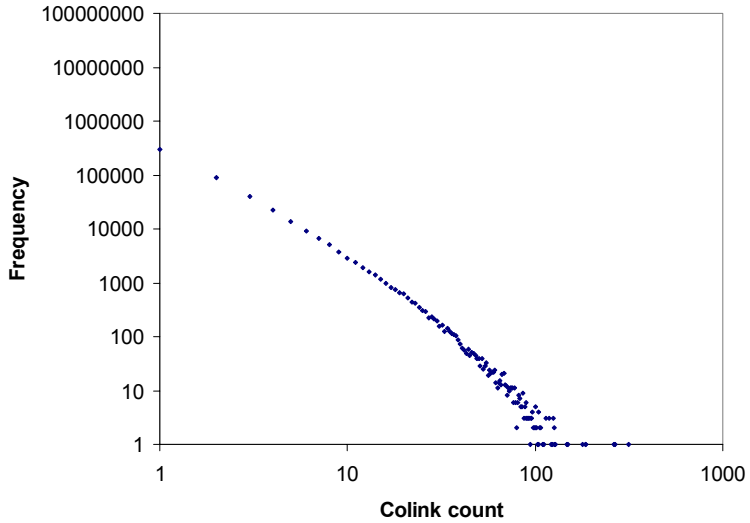


Table 1. The number of pairs of domains falling into each combination of linking categories.

Type of link connection	Number of pairs of domains	Percentage
No link connection	21,311,733	93.45%
Direct link only	1,922	0.01%
Colinked only	357,283	1.57%
Colinked and direct link	2,519	0.01%
Coupled only	977,971	4.29%
Coupled and link	1,577	0.01%
Coupled and colinked	142,517	0.62%
Coupled, colinked and direct link	9,359	0.04%
Total	22,804,881	100.00%

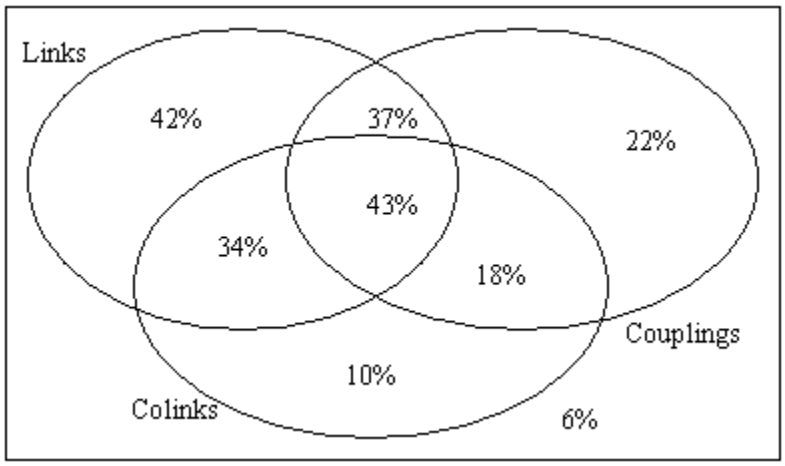
Related Pages

Table 2 and Figure 2 show the percentage of site pairs in each category that were deemed similar and identical respectively. Each category contained approximately 60 pages and so the percentages are all rough estimates for the whole population. The information in Figure 2 overlaps with that in the table, but it is included as a helpful visualisation.

Table 2. The percentage of pairs of domains in each category that were categorised as similar or identical. See Table 4 for total numbers in each category.

Type of link connection	Percentage categorised as identical	Percentage categorised as similar
No link connection	0%	6%
Direct link only	4%	38%
Colinked only	1%	9%
Colinked and direct link	3%	31%
Coupled only	0%	22%
Coupled and link	9%	28%
Coupled and colinked	2%	17%
Coupled, colinked and direct link	11%	32%

Figure 2. The percentage of pages classified as similar or identical in each category. For example 37% of pairs of pages that have links and couplings but not colinks are similar or identical.



A different perspective on the same data is given in Table 3 by estimating the total number of pairs of sites that fall into each category. Note that the totals are highly unreliable because of the low percentage for the largest category, no link connection. In particular the total percentage of identical pairs of domains is unrealistically low, assuming that there are a large number of departments for each of the RAE categories and that most pairs of these would be classed as identical.

Table 3. The predicted number of pairs of domains falling into each combination of linking categories. Predictions are based upon multiplying the percentage of pairs in each category (Table2) by the number of pairs in the category (Table 1).

Type of link connection	Predicted number of identical pairs	Percentage of all identical pairs	Predicted number of similar pairs	Percentage of all similar pairs
No link connection	0	0%	1,291,620	81%
Direct link only	77	1%	807	0%
Colinked only	5,178	58%	36,246	2%
Colinked and direct link	75	1%	865	0%
Coupled only	0	0%	219,200	14%
Coupled and link	141	2%	588	0%
Coupled and colinked	2,315	27%	26,128	2%
Coupled, colinked and direct link	1,040	12%	4,011	0%
Total	**8,886**	**100%**	**1,579,466**	**100%**

Revised Coupling and Colink Thresholds

The data was reworked with thresholds of 2, 3, 4, 5 and 10 for the colink and coupling values, but this traded off small increases in the percentage of matching pairs for large decreases the total number of pairs falling into the category, as can be seen from Table 4. This approach is only recommended for applications where precision outweighs recall.

Discussion

In answer to the first research question, links are clearly the most reliable indicator of similarity and of being identical. The hypothesis that identical domains were more likely to be colinked or coupled than to directly link is not supported by the evidence.

The combination of indicators most likely to identify similar domains is all three, but this is not substantially different from just using links irrespective of the values of the others. Similarly, the most likely combination for finding identical domains is to restrict attention to pages with all three indicators or links and bibliometric coupling. Colinks appear to be relatively weak for this purpose.

The same question can also be answered from a different perspective: the need to identify the majority of identical or similar pairs of domains. For this, the actual numbers of pages in each category must be taken into account. From Table 3 it can be seen that the advantage of using couplings and colinks is not that they are more powerful than links but that they cover a much greater range of domain pairs.

The final question addressed the issue of whether more highly colinked or coupled pairs of domains were more likely to be similar or identical. This was indeed the case, but the improvements were not dramatic. However, for the purpose of designing an algorithm to identify pairs of similar domains the threshold can be raised to increase the likelihood that pairs are similar or lowered to increase the total number of pairs in the category but with only marginal changes in similarity likelihood to be expected.

There are several limitations of this study. Firstly, the area covered is only the UK academic area. Results may be different for other academic areas, particularly less well developed ones. There is also a possibility that customs for link creation or domain organisation are different in one or more other countries. Secondly, the domain is not always identified with a coherent site: sites may span multiple domains or there may be several domains on one site. Unfortunately the lack of tools or even a theory for identifying Web sites automatically over large areas of the Web means that this is currently a necessary limitation.

The potential extension of the results to non-academic areas will be of particular interest to search engine algorithm designers and to those whose Web based research includes this area. The academic Web seems to be particularly well interlinked so it is likely that the results would be substantially different if tested on the whole Web, at least in terms of the absolute percentages reported.

Conclusion

Colinks and bibliometric couplings between domains can yield some useful information about similarity but their primary use is in casting a wider net than direct links, which are relatively much scarcer. They do have the potential to be used in algorithms to find related sites, therefore, and threshold values can be set to adapt an algorithm to the required balance of similarity likelihood (precision) and category size (recall), based upon the degree of colinking or coupling, but this is likely to have only a marginal effect.

As a final point, from a Webmetrics perspective, it is perhaps discouraging that the percentages reported in all figures are so low. In fact a majority of pairs of domains that trigger all indicators in Figure 2 are still not closely related. A follow-up paper will investigate this phenomenon in an attempt to provide an explanation.

References

Aguillo, I. F. (1998). STM information on the Web and the development of new Internet R&D databases and indicators. In: *Online Information 98: Proceedings***.** Learned Information, 1998. 239-243.

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

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A. & Raghavan, S. (2001). Searching the Web, *ACM Transactions on Internet Technology*, 1(1), 2-43.

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.

Björneborn, L. (2001a). 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. (2001b). 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.

Borgman, C & Furner, J. (2002). Scholarly communication and bibliometrics. In: Cronin, B. (ed.), Annual Review of Information Science and Technology 36, pp. 3-72, Medford, NJ: Information Today Inc.

Broder, A. Kumar, R, Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph dtructure in the Web, Journal of Computer Networks, 33(1-6), 309-320.

Cawkell, T. (2000). Visualizing citation connections. In: B. Cronin, and H.B. Atkins, (eds.), *The Web of knowledge: a festschrift in honor of Eugene Garfield*. Metford, NJ: Information Today Inc. ASIS Monograph Series, pp. 177-194.

Chakrabarti, S., Joshi, M. M., Punera, K. & Pennock, D. M. (2002). The structure of broad topics on the Web, WWW2002, Available: http://www2002.org/CDROM/refereed/338/

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(4), 353-373.

Chen, C. (1999). *Information visualisation and virtual environments.* London: Springer.

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

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

Dearing, R. (1997). Report of the national committee for enquiry into higher education. Available: http://www.leeds.ac.uk/educol/ncihe/

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

Gao, J. Walker, S., Robertson, S., Cao, G., He, H., Zhang, M. & Nie, J-Y (2001). TREC-10 Web Track Experiments at MSRA 384-392. TREC 2001. Available: http://trec.nist.gov/pubs/trec10/t10_proceedings.html.

Garrido, M. & Halavais, A. (2003). Mapping networks of support for the Zapatista movement: Applying Social Network Analysis to study contemporary social movements. *In*: M. McCaughey & M. Ayers (Eds). *Cyberactivism: Online Activism in Theory and Practice*. London: Routledge.

Glänzel, W. & Schubert, A. (2001). Double effort = double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, 50(2), 199-214.

Haveliwala, T. H., Gionis, A. & Indyk P. (2000). Scalable techniques for clustering the Web. WebDB 2000. Available: http://www.research.att.com/conf/Webdb2000/PAPERS/8c.ps.

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

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

Li, X., Thelwall, M., Musgrove, P. & Wilkinson, D. (2003, in press). The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE (Research Assessment Exercise) ranking in 2001, *Scientometrics*, 57(2).

Ng, A. Y., Zheng, A. X. & Jordan, M. I. (2001a). Link analysis, eigenvectors and stability. In: *Proceedings of the 17th Joint International Conference on Artificial Intelligence,* Seattle, WA, August 2001. Morgan Kaufmann. pp. 903-910.

Ng, A. Y., Zheng, A. X. & Jordan, M. I. (2001b). Stable algorithms for link analysis. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001),* New York: ACM Press, pp. 258-266.

Park, H. W., Barnett, G. A. & Nam, I. (2002). Hyperlink-affiliation network structure of top Web sites: Examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science,* 53(7), 592-601.

Pennock, D., Flake, G., Lawrence, S., Glover, E. & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the Web, Proceedings of the National Academy of Sciences, 99(8), 5207-5211.

Pirolli, P., Pitkow, J. & Rao R. (1996). Silk from a sow's ear: extracting usable structures from the Web. Conference proceedings on Human factors in computing systems, ACM Press New York, NY, USA, pp 118-125.

Polanco, X, Boudourides, M. A., Besagni, D. & Roche, I. (2001). Clustering and mapping Web sites for displaying implicit associations and visualising networks. University of Patras. Available: http://www.math.upatras.gr/~mboudour/articles/Web_clustering&mapping.pdf

Rogers, R. (2002). Operating issue networks on the Web, Science as Culture, 11(2), 191-214.

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

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. NY: McGraw-Hill.

Schvaneveldt, R.W., Durso F.T. & Dearholt, D.W. (1989). Network structures in proximity data. In: G. Bower (ed.), *The psychology of learning and motivation* 24, Academic Press pp. 249-284.

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, 38(2), 275-293.

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

Tang, R. & Thelwall, M. (2003, in press). Disciplinary differences in US academic departmental web site interlinking, *Library and Information Science Research*.

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). Exploring the link structure of the Web with network diagrams, *Journal of Information Science* 27(6) 393-402.

Thelwall, M. (2001c). A publicly accessible database of UK university Website links and a discussion of the need for human intervention in Web crawling, University of Wolverhampton. Available: http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf

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

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

Thelwall, M. (2002b). 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*, 53(12), 995-1005.

Thelwall, M. (2003, in press). A layered approach for investigating the topological structure of communities in the Web, *Journal of Documentation*, 59(3).

Thelwall, M. & Smith, A. (2002). A study of the interlinking between Asia-Pacific university Web sites, *Scientometrics* 55(3), 335-348.

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

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks, *Nature*, 393, 440-442.

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

Table 4. The results of varying the minimum number of colinks or couplings k to trigger category membership. The data is for the combined Identical and Similar categories.

	k = 1			k = 2			k = 3			k = 4			k = 5			k = 10		
	Cases	Matches	%	Cases	Matches	%	Cases	Matches	%	Cases	Matches	%	Cases	Matches	%	Cases	Matches	%
No link connection	66	4	6%	162	20	12%	194	26	13%	211	28	13%	224	29	13%	243	32	13%
Direct link only	50	21	42%	89	32	36%	123	45	37%	140	50	20%	155	55	35%	193	71	37%
Colinked ≥k only	69	7	10%	41	4	10%	28	2	7%	23	2	36%	16	2	13%	7	1	14%
Colinked ≥k and direct link	67	23	34%	54	19	35%	45	15	33%	39	14	44%	33	13	39%	19	9	47%
Coupled ≥k only	58	13	22%	26	4	15%	19	4	21%	10	2	9%	6	2	33%	2	1	50%
Coupled ≥k and link	67	25	37%	60	26	43%	44	20	45%	39	17	33%	35	14	40%	21	81	38%
Coupled ≥k and colinked ≥k	60	11	18%	24	7	29%	12	3	25%	9	3	36%	7	2	29%	1	1	100%
Coupled ≥k, colinked ≥k and direct link	63	27	43%	44	19	43%	35	16	46%	29	15	52%	24	14	58%	14	8	57%