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

VOLUME 10 (2006): ISSUE 1. PAPER 3

Operationalising "Websites": lexically, semantically or topologically?

Viv Cothey*
Isidro Aguillo & Natalia Arroyo**
* School of Computing and IT
University of Wolverhampton
Wolverhampton, WV1 1EQ, United Kingdom
E-mail: viv.cothey@wlv.ac.uk
** Internet Lab
CINDOC-CSIC
Joaquin Costa, 22
28002 Madrid, Spain
E-mail: isidro@cindoc.csic.es ; narroyo@cindoc.csic.es

Abstract

Methods to investigate the structure of the Web graph in order to better understand its properties are of interest to many researchers. The scale and complexity of the Web-page digraph is typically managed by aggregating together or clustering individual Web-pages in order to form "Websites". It is the properties of these Websites which then become the focus of research. The most popular Web-page clustering technique is "lexical" and uses the url syntax in order to assign Web-pages to "Websites". Semantic clustering, that is clustering Web-pages according to the similarity of their content has also been proposed. In this paper we consider a third approach to Web-page clustering which is based on the topological properties of the Web-page within the Web-page digraph. We present the technique and report the results of an experiment to compare the use of url-lexically and topologically determined Websites in two sub-domains, one within the Spanish country level domain and the other within the UK country level domain of the Web.

Keywords

World Wide Web, Web graph, Website, topology, cliques, clusters

1. Introduction

The discovery of possible patterns and the emergence of regularities within the Web-graph is of interest, for example to inform the design of information discovery systems (Bharat et al., 2001; Menczer, 2004). Analogies may be drawn between the Web-graph and other networks of information such as citation networks although this is problematic (Egghe, 2000). It is therefore of interest to experiment with the Web graph to find possible patterns of citation type features. Such features may facilitate insight in respect of an important emerging need which is to understand issues of presence and visibility within the Web. Associated with the broad issue of visibility in the Web is the determination of impact within the Web (Ingwersen, 1998) which is based on in-degree and which can be thought of as analogous to "journal impact" (for example, Garfield, 1999).

All such Web research is based on analysing a Web graph that is constructed from Web crawl data. Obtaining crawl data for Web research is itself

problematic and requires the use of special software tools (Björneborn & Ingwersen, 2001). The data generally comprises a large set of Web-pages together with their associated outlinks. It can therefore be thought of as a directed graph (digraph) where each node or vertex in the digraph represents a Web-page and each arc represents a hyperlink from one Web-page to another. In this context the term Web-page includes both html and non-html resources although outlinks are usually collected from html resources only.

Figure 1 shows the Web-page digraph illustrating a link crawl (Cothey, 2004) of the “cindoc.csic.es” (cindoc) sub-domain within the Spanish country level domain. A similar link crawl of the “scit.wlv.ac.uk” (scit) sub-domain within the United Kingdom country level domain was also carried out. The link crawl data for each sub-domain has been simplified by removing all loops and multiple arcs. The cindoc Web-page digraph initially comprised 3577 vertices and 11737 arcs but simplification and removing the vertices that correspond to Web-pages outside the cindoc sub-domain (and their associated arcs) reduces the digraph to 1371 vertices and 8988 arcs as in Figure 1. The scit Web-page digraph had 85073 vertices and 1617565 arcs. This is similarly reduced to 55038 vertices and 418370 arcs.

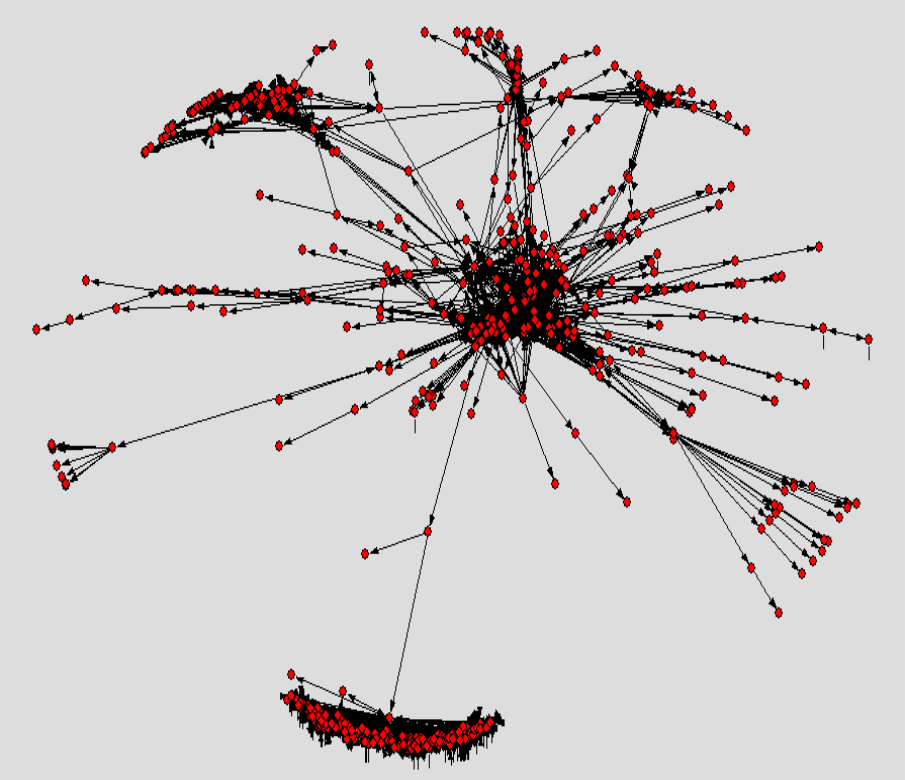

Figure 1: Simplified Web-page digraph of the cindoc.csic.es sub-domain

In general the Web-page digraph is intractable since it soon becomes massive. In consequence attention turns to the notion of a "Website" that comprises a collection of Web-pages. Thus the Web is thought of as a smaller network of interlinked Websites rather than the more massive Web-page graph. Bharat et al. (2001) argue that in any event this is a more appropriate level of abstraction.

The Web-page digraphs illustrated are drawn by the Pajek graph visualisation tool (Batagelj & Mrvar, 2003). The visualisations distribute the vertices according to how well connected they are in the graph. It is immediately apparent that the digraph appears to contain clusters of Web-pages. By way of comparison, Figure 2 illustrates a similar Pajek visualisation for a simple random digraph having the same number of vertices and arcs as the cindoc Web-page digraph in Figure 1. The random digraph lacks any visual clustering.

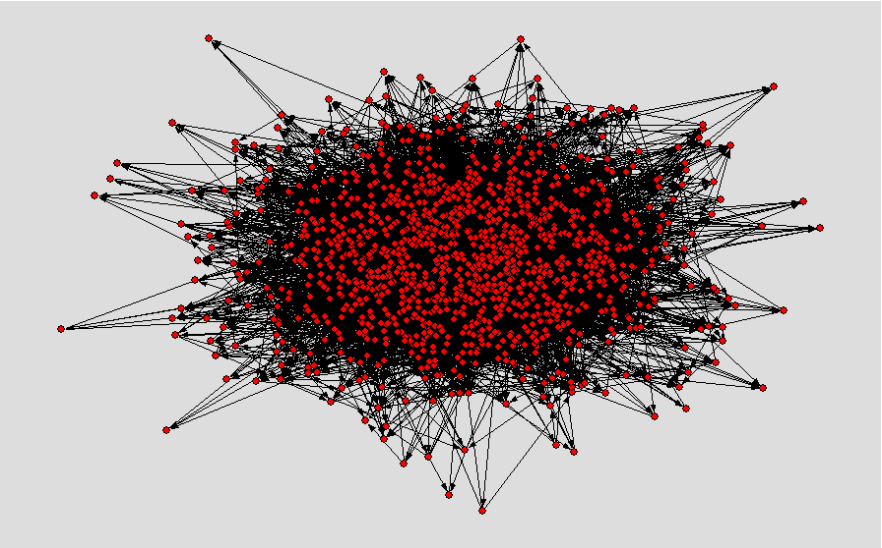

Figure 2: Simple random digraph of equivalent size to cindoc

Although the notion of a Web-page can be operationalised in a straightforward way as the resource corresponding to a particular url, the same is not true of "Website". There is no consensus as to what a "Website" is (Deo & Gupta, 2001) although in practice the term is frequently used to suggest a coherent assemblage of Web-pages that is intended to be considered as a unit by the reader. Ayan et al (2002) use the term "logical domain" to describe this coherent unit.

The lack of effective operationalisations for "Website" hinders progress in the research areas mentioned. In particular there is a need to understand the effect that different pragmatic analytic choices may have on the results of such Web-graph analyses in order to aid interpretation of their results (Thelwall, 2004).

In this paper we discuss some operationalisations of "Website", that is techniques for clustering together collections of Web-pages so that the consequential link structure between the "Websites" can be analysed. It is to be expected that as the number of clusters so formed reduces, so the frequency of the inter-cluster links will also reduce. Therefore the metric of interest here is how the impact or relative frequency of inlinking to a Website may be affected by the clustering technique.

In Sections 2 and 3 we discuss two operationalisation techniques and provide examples of their implementation. In Section 4 we present an indegree analysis of the different Website digraphs and compare the impacts of the Websites. We also discuss our techniques in the context of research in the semantic clustering of Web-pages. The paper concludes with Section 5 which also suggests further work.

2. Url-lexical clustering of Web-pages to form “Websites”

This Web-page clustering technique exploits the lexical structure of the url that labels each Web-page or vertex, for example, , This url syntax comprises the host server name "www.cindoc.csic.es", the path component "/info/example/" and the filename "infrev2.html".

The vertex label in the digraph can be modified by including only part of the host server name and path components. This generates a number of different levels of url-clustering. At level 1, the modified label consists of just the host server name. At level 2, the modified label includes the host server name and the first segment of the path component demarcated by the "/" character, at level 3, the host server name and the first two segments of the path component, and so on.

Hence for the previous example, the url label , becomes at level 1, "www.cindoc.csic.es/", at level 2, "www.cindoc.csic.es/info/" and at level 3 and above, "www.cindoc.csic.es/info/example/"

Once this url-lexical modification has been carried out then the digraph can be analysed and shrunk so that vertices having the same (modified) label coalesce. An interpretation of the effect of this procedure for levels 2 and greater is that Web-pages that share their host directory are aggregated to form "Websites".

The effect of this url-lexical clustering on the cindoc and scit Web-page digraphs at levels 2 and 3 is shown in Figures 3 to 6. Figures 7 and 8 show the level 1 url-lexical (or host server name) clustering of cindoc and scit Web-page graphs. The clustered Web-page digraphs are simplified (no loops nor multiple arcs) and, for Figures 3 to 6, clusters with indegree less than two are excluded. The impact distribution revealed by this clustering is discussed in Section 4.

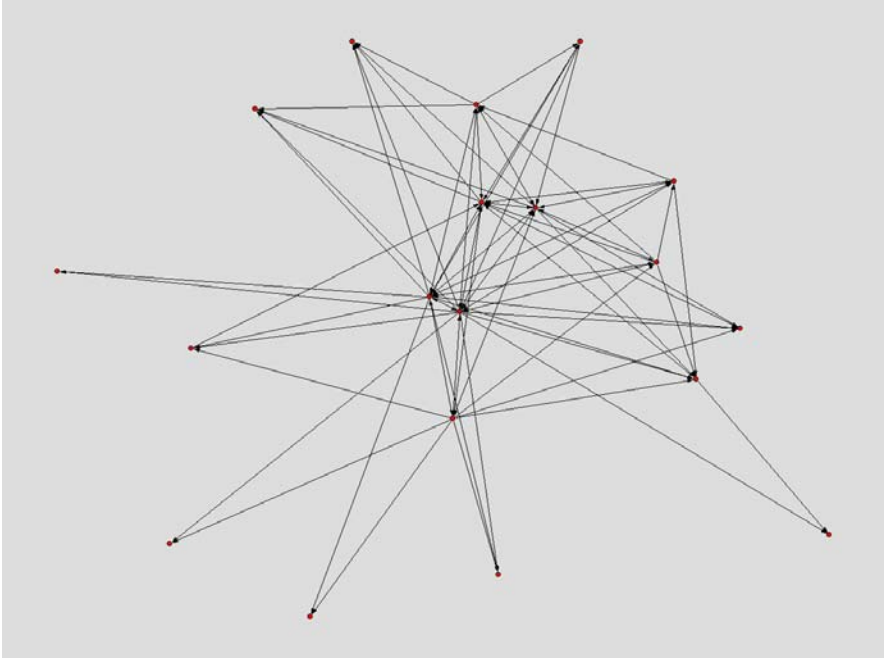

Figure 3: Web-page clustering for cindoc at url-lexical level 2 (19 clusters)

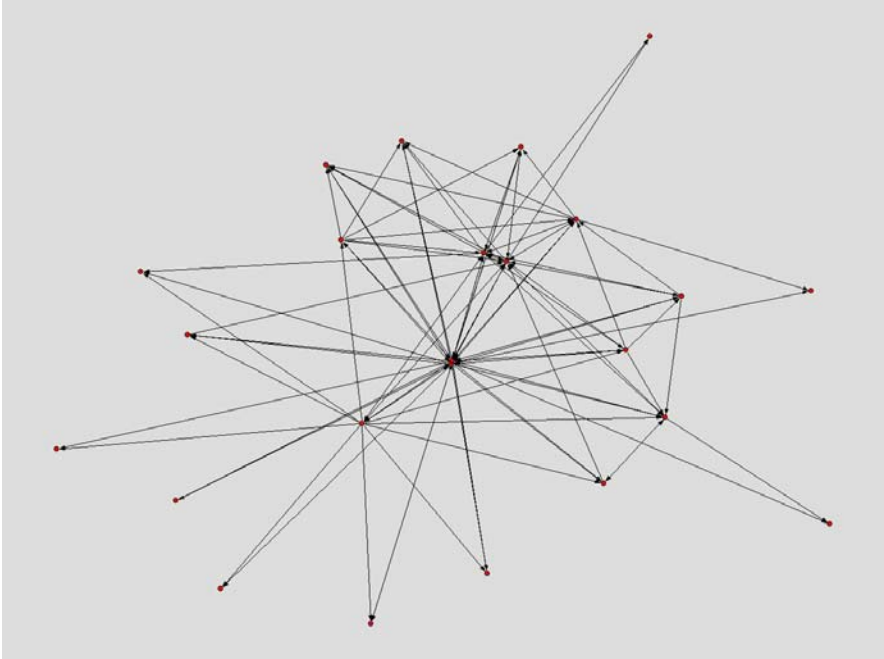

Figure 4: Web-page clustering for cindoc at url-lexical level 3 (24 clusters)

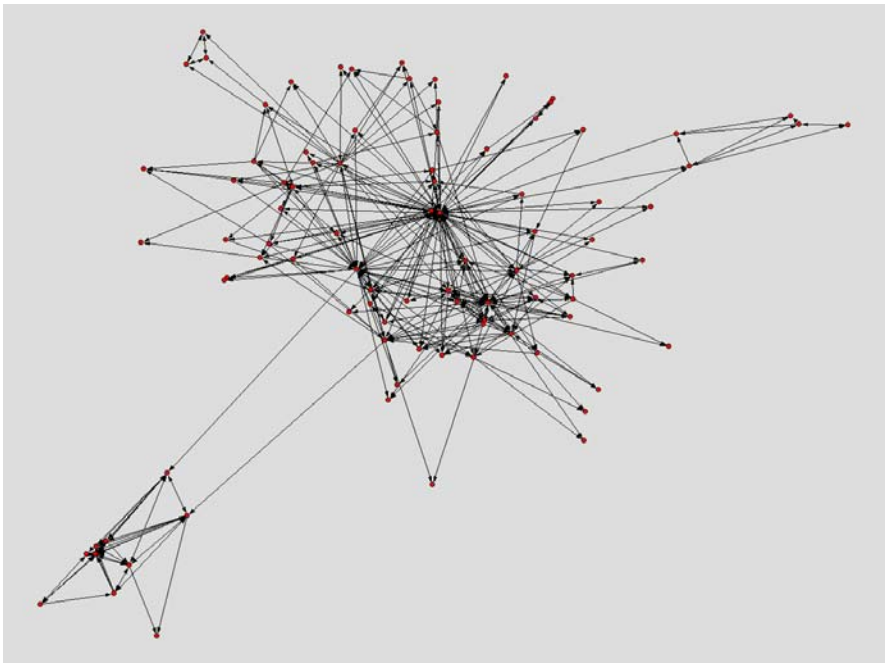

Figure 5: Web-page clustering for scit at url-lexical level 2 (95 clusters)

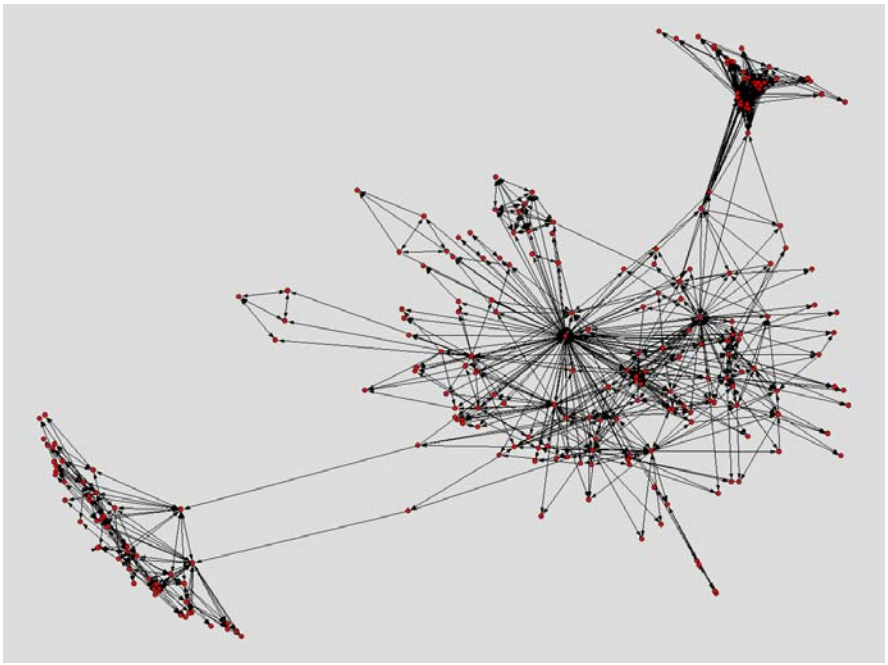

Figure 6: Web-page clustering for scit at url-lexical level 3 (279 clusters)

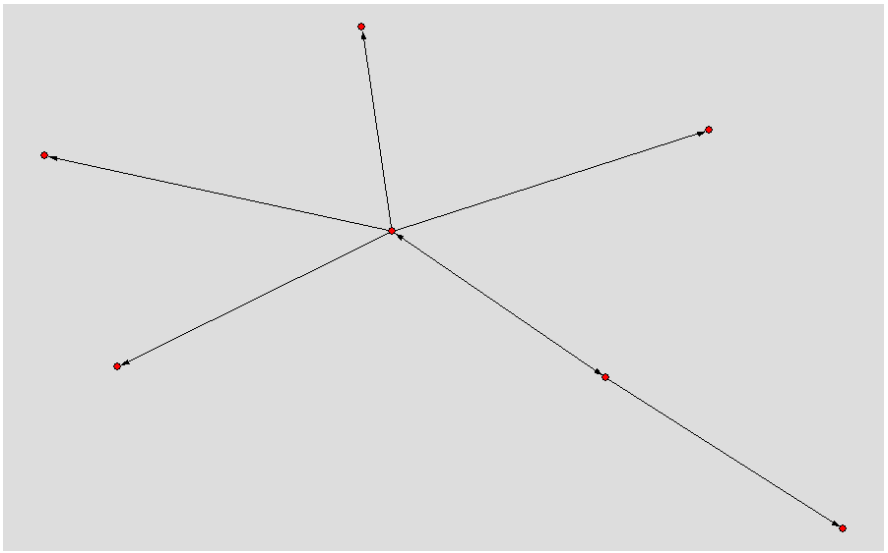

Figure 7: Host server name Web-page clustering for cindoc (seven clusters)

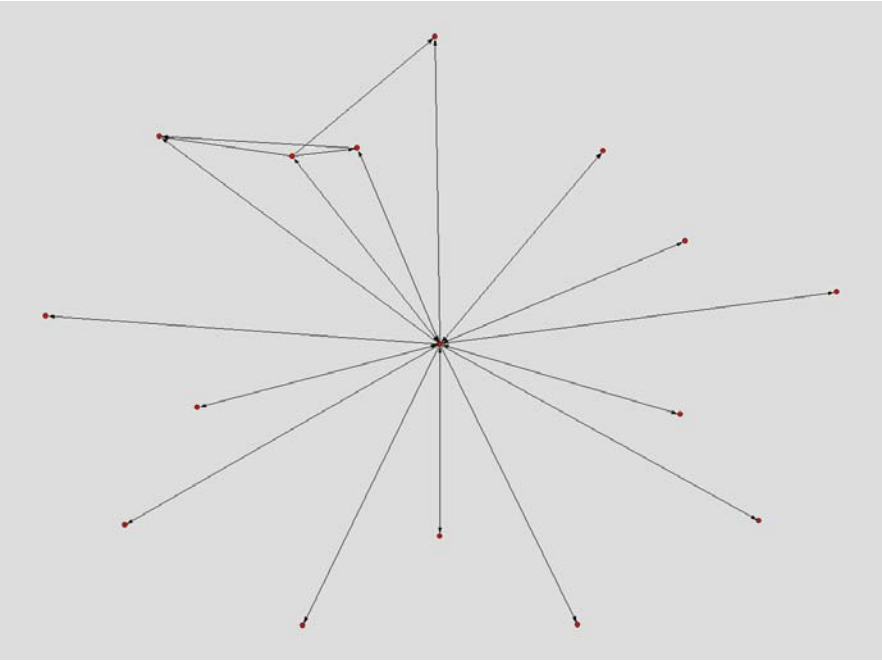

Figure 8: Host server name Web-page clustering for scit (sixteen clusters)

3. Topological clustering of Web-pages to form “Websites”

Leydersdorf (2004) provides an example of a purely topological analysis of a citation type graph in order to cluster the nodes of the graph and to provide insight to the underlying structure. In the context of the Web, adopting a topological perspective offers a basis for clustering Web-pages into Websites that is an alternative to the argument that Web-pages that share a directory form a "Website" because of their administrative cohesion. It therefore provides an insight to the underlying structure of the Web that is unbiased with respect to the administrative operation of the Web.

It can be argued topologically that the Web-pages within a Website will be more strongly interlinked than they will be linked to Web-pages not in the Website. The intuition here is that a Website is constructed by the author as single hyperdocument (Bharat et al. 2001). In consequence the Web-pages of Websites are no longer assumed to share the same directory structure in their url labels but are instead assumed to be more strongly linked together. Menczer (2004) also leans towards aggregating Web-pages that are topologically close into sites (although his procedure is mediated by analyses of Web-page content).

The topological operationalisation for Website considered here is a clique feature appearing within the Web-page digraph. Strictly, in a clique every vertex links to every other vertex. In practice this condition is too strong and here we use the weaker p-clique feature with p=10%. In a p-clique, each vertex links to at least a proportion p of the other vertices in the p-clique. Clearly the value of p can be varied. Generally, given a collection of Web-pages, each has links to relatively few other pages. Increasing the value of p reduces the chance of there being a p-clique of a given size so that over some threshold value no p-cliques will be found.

Figures 9 and 10 show the effect of clustering the cindoc and scit Web-page digraphs according to their p-cliques. This clustering procedure is entirely topological since it is based on the link structure within the Web-page digraph without any reference to the url labels of the vertices.

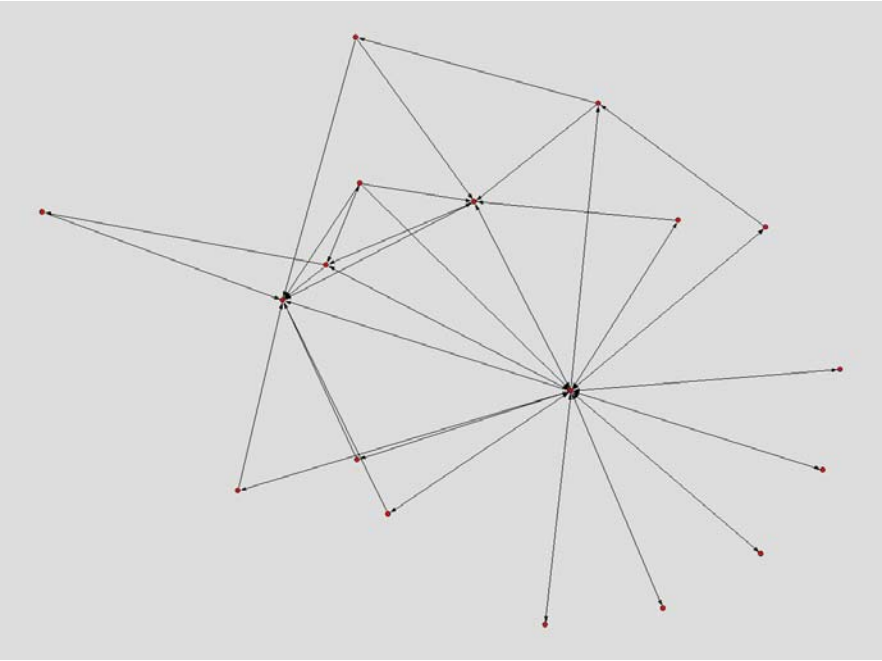

Figure 9: Topological Web-page clustering for cindoc at p=10% (18 clusters)

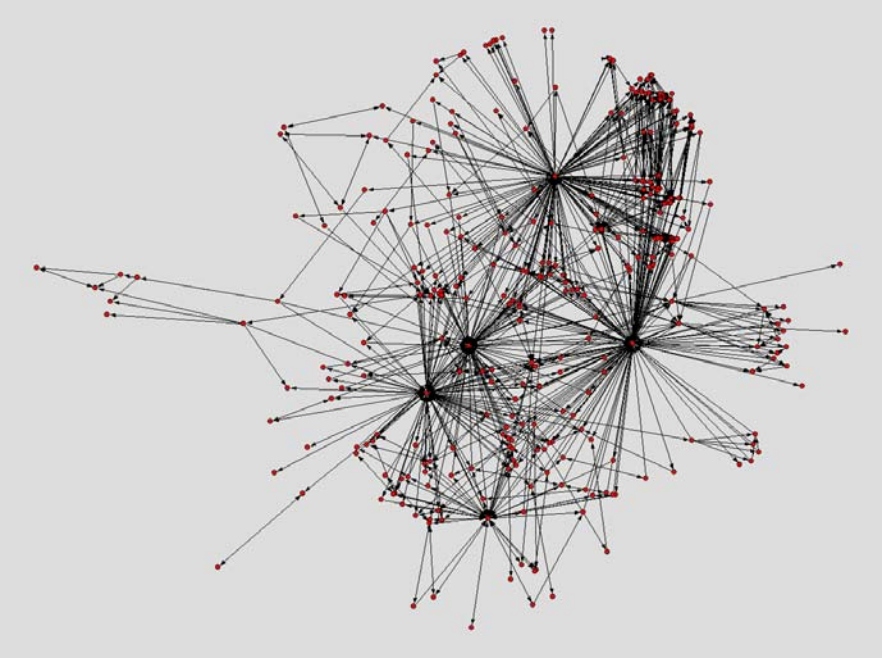

Figure 10: Topological Web-page clustering for scit at p=10% (538 clusters)

The existence of topological Websites as shown by Figures 9 and 10 is caused by a real topological clustering in the underlying Web-page digraph. An equivalent analysis of the random digraph mentioned earlier produces just two clusters. This distinction between the Web-page graph and a random graph is a characteristic of the "small-world" phenomenon (Watts & Strogatz, 1998).

4. Results and discussion

The effect of the different clustering operationalisations just considered is compared by using the inter-Website link structure that is revealed to compute a Website impact metric. This metric is cluster indegree divided by the average cluster indegree. For the cindoc Web-page digraph the Website impacts are given (in descending order) in Table 1 together with the number of Web-pages contained within each cluster/Website. Table 2 reports the equivalent analysis for the scit Web-page digraph.

url-lexical level 3 (24 sites)	**url-lexical level 2 (19 sites)**	**url-lexical level 1 (7 sites)**	**p-clique (18 sites)**

2.6 (8 pages)	2.1 (8 pages)	1.0 (1268 pages)	4.8 (341 pages)
2.0 (100 pages)	1.7 (101 pages)	1.0 (71 pages)	3.6 (483 pages)
2.0 (81 pages)	1.5 (205 pages)	1.0 (28 pages)	2.4 (114 pages)
1.7 (66 pages)	1.5 (74 pages)	1.0 (1 page)	1.2 (2 pages)
1.5 (34 pages)	1.3 (107 pages)	1.0 (1 page)	0.8 (1 page)
1.5 (4 pages)	1.3 (30 pages)	1.0 (1 page)	0.4 (338 pages)
1.3 (1 page)	1.1 (29 pages)	1.0 (1 page)	0.4 (35 pages)
1.3 (1 page)	1.1 (15 pages)		0.4 (17 pages)
1.1 (350 pages)	1.1 (3 pages)		0.4 (14 pages)
0.9 (64 pages)	1.1 (1 page)		0.4 (10 pages)
0.9 (21 pages)	0.9 (350 pages)		0.4 (2 pages)
0.9 (5 pages)	0.9 (129 pages)		0.4 (2 pages)
0.7 (106 pages)	0.6 (143 page)		0.4 (2 pages)
0.7 (88 pages)	0.6 (23 pages)		0.4 (2 pages)
0.7 (23 pages)	0.6 (2 pages)		0.4 (2 pages)
0.7 (3 pages)	0.4 (5 pages)		0.4 (2 pages)
0.7 (2 pages)	0.4 (2 pages)		0.4 (2 pages)
0.4 (18 pages)	0.4 (2 page)		0.4 (2 pages)
0.4 (5 pages)	0.4 (1 page)		
0.4 (2 pages)			
0.4 (2 pages)			
0.4 (1 page)			
0.4 (1 page)			
0.4 (1 page)			

Table 1: cindoc Website impacts and Web-page counts for different Website operationalisations

url-lexical level 3 (279 sites)	**url-lexical level 2 (95 sites)**	**url-lexical level 1 (16 sites)**	**p-clique (538 sites)**
10.0 (286 pages)	7.3 (286 pages)	5.1 (53192 pages)	74.5 (13277 pages)
5.4 (157 pages)	4.1 (5683 pages)	2.2 (13 pages)	47.3 (17589 pages)
5.0 (185 pages)	4.1 (1600 pages)	1.5 (32 pages)	31.9 (623 pages)
4.8 (165 pages)	2.9 (6830 pages)	1.5 (14 pages)	24.3.2 (454 pages)
4.8 (48 pages)	2.9 (191 pages)	0.7 (1610 page)	12.5 (1062 pages)
4.6 (186 pages)	2.3 (40 pages)	0.7 (111 page)	6.8 (8377 pages)
4.6 (179 page)	2.0 (29 pages)	0.7 (18 pages)	3.9 (156 pages)
4.2 (177 page)	1.8 (493 pages)	0.7 (17 pages)	3.9 (7 pages)
4.2 (174 pages)	1.8 (370 pages)	0.7 (10 pages)	3.6 (1134 pages)
3.8 (269 pages)	1.8 (112 pages)	0.7 (8 pages)	3.2 (1131 pages)
3.8 (183 pages)	1.8 (55 pages)	0.7 (3 pages)	2.9 (31 pages)
3.6 (164 pages)	1.8 (20 pages)	0.7 (2 pages)	2.9 (2 pages)
3.6 (12 pages)	1.8 (5 pages)	0.7 (2 pages)	2.5 (31 pages)
3.6 (4 pages)	1.6 (16816 pages)	0.7 (2 pages)	2.1 (2 pages)
3.2 (181 pages)	1.6 (268 pages)	0.7 (2 pages)	1.8 (2964 pages)
3.2 (135 pages)	1.6 (242 pages)	0.7 (1 page)	1.8 (31 pages)
3.2 (10 pages)	1.6 (87 pages)		1.8 (9 pages)
3.0 (183 pages)	1.6 (6 pages)		1.8 (9 pages)

2.9 (172 pages)	1.4 (2144 pages)		1.8 (7 pages)
2.7 (191 pages)	1.4 (146 pages)		1.8 (2 pages)
2.5 (226 pages)	1.4 (28 pages)		1.8 (1 pages)
2.5 (177 page)	1.4 (5 pages)		1.8 (1 pages)
2.5 (119 page)	1.4 (5 pages)		1.4 (18 pages)
2.5 (116 page)	1.4 (3 pages)		1.4 (16 pages)
etc.	etc.	etc.	etc.

Table 2: scit Website impacts and Web-page counts for different Website operationalisations

The results tabulated show the range of impact reducing as the url lexical level reduces. At level one the analysis is flattened and for cindoc reaches the extreme of all the seven (host server name) Websites having the same impact. This level one url-lexical clustering generates a Website graph that is equivalent to the host-graph of Bharat et al. (2001). Url-lexical clustering is similar in principle to the alternative document model (ADM) approach for analysing Web crawl data developed by Thelwall (2002) but differs in implementation since it introduces the idea of level. In the ADM approach nodes in the Web-page digraph are clustered according to whether they share a url-lexical path component so that nodes having different levels are combined. Also in the ADM an arc indicates just the existence of one or many links between any Web-page in either cluster.

As expected the number of different Websites identified url-lexically also reduces as the level reduces which indicates that, administratively, a hierarchical directory structure is being used. However it is evident that the Web-pages in the scit digraph make greater use of a deeper directory type hierarchy than do Web-pages in cindoc. This reflects the need to differentiate administratively the much larger number of Web-pages in scit because of the size of the principal host server.

For cindoc the number of Websites revealed by the topological p-clique analysis is similar to the url-lexical analysis at level 2 which includes the first component of the url path. In the scit digraph, the topologically defined Websites outnumber those at url-lexical level 3. This corresponds to the deeper administrative hierarchy being used in scit. It also suggests that use of a url-lexical operationalisation to determine Websites needs to take account of the size of a host name server (in order to apply an appropriate level) in order to be effective whereas the topological approach can be applied irrespective of the size of the host name server.

The topological approach also differentiates more strongly those Websites that have high impact. In addition, comparison of Website page counts shows that the topological operationalisation is able to both disaggregate Websites that share an administrative directory hierarchy and aggregate Web-pages from different directory hierarchies into a Website.

The impact computation used is not normalized. Hence a Website comprising many pages may obtain a large impact value by virtue of its size which provides a greater potential for linking to by other Websites. However this effect does not appear inevitable and both cindoc and scit provide examples of large Websites with low impact and vice-versa.

Neither of the approaches demonstrated to operationalise "Website" (url-lexical or topological) have exploited the (non-link) content of Web-page in order to aggregate Web-pages. In contrast both Menczer (2004) and Ayan et al. (2002) investigate semantic operationalisation procedures whereby Web-pages that are about the same topic may be clustered together. Compared to the url-lexical or p-clique procedures, the semantic procedures demand more analysis (for example, of Web-page content) and possibly access to some form of Web-page classification. Menczer reports an empirical validation of the notion that topologically close Web-pages are similar. This finding supports the intuition that authors construct Websites as coherent hyperdocuments and which is exploited by the p-clique analysis.

Operationalising a Website semantically presents at least two problems. First,

it is unlikely that very many Web-pages would be about just a single narrowly defined topic. (Of course the inference of “topic”, or what the Web-page means, is itself problematic as is indicated, for example, by Fry (elsewhere this issue)). Thus either one accepts that a Web-page can be in two or more Websites which makes for difficulties of analysis or each topic has to be more broadly defined. This exacerbates the second problem which is that for popular topics the Website would tend to “explode” since the collection of Web-pages that are about the a popular topic may become arbitrarily large. Hence while semantic clustering provides benefits in respect of information retrieval and accessing content, it presents challenges when used to define a Website.

A difficulty with the topologically aggregated Website is that there is no natural way of naming the Website since it comprises a collection of Web-pages that have potentially arbitrary url labels. Following Ayan et al. (2002), a possible basis for a naming description could use the url-labels by which the p-clique is linked to by other Websites. As with the url-lexical approach, it would have to be understood that the Website name label stood for a collection of Web-pages and not for just one Web-page.

5. Conclusion

The experiment using the cindoc and scit sub-domains demonstrates the feasibility of operationalising topographically the notion of a Website in the Web-page digraph. A benefit of this approach is that it avoids the administrative bias inherent in the url-lexical approach. In consequence the topological approach is unaffected by the size of the Web host servers that are included in the Web-page digraph. The topological approach is thus a more flexible operationalisation for Website when compared with a url-lexical approach. It also offers what may be regarded as an appropriate level of abstraction on which to base further investigations into the structure of the Web graph.

In the sub-domains investigated the topological operationalisation provided greater discrimination between the impact of the Websites revealed and appeared effective at both aggregating and disaggregating Websites that are not determined by the lexical components of their constituent Web-page urls.

The topologically operationalised Websites reveal the clusters of Web-pages that their authors have related in order to provide the reader with a particular hypertext. There is thus an intrinsic connectedness of informational content within these Websites that is not available elsewhere. And, equivalently, there is also an informational distinctiveness between these Websites that is not otherwise identified.

It should be noted that the analysis required when using the topological operationalisation described here is computationally far more demanding than that needed for the url-lexical operationalisations. This will impose constraints on larger scale investigations that may therefore demand a url-lexical operationalisation.

However the initial experience of a topological operationalisation is considered sufficiently promising to extend the analysis over a larger sample of the Web-page digraph. In particular such analyses should investigate the occurrence of multi host Websites, that is Websites that comprise Web-pages from more than one host server.

Acknowledgements

This work was supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technological Development" of the European Commission. It is part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015).

References

Ayan N. F., Li W. and Kolak O. (2002). Automating extraction of logical

domains in a web site. **Data & knowledge engineering**, 43(2), 179-205.

Batagelj V. and Mrvar A:, (2003). Pajek - Analysis and Visualization of Large Networks. In Juenger, M., and Mutzel, P. (eds.), **Graph drawing software**. pp. 77-103, Berlin: Springer.

Bharat K., Chang B., Henzinger M. and Ruhl M., (2001). Who links to whom: mining linkage between Web sites. In **Proceedings of the 2001 IEEE International Conference on Data Mining**, pp. 51-58.

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

Cothey V., (2004). Web-crawling reliability. **Journal of the American Society for Information Science and Technology**, 55(14), 1228-1238.

Deo N. and Gupta P., (2001). World wide web: a graph-theoretic perspective. Technical report CS-TR-01-001, School of Computer Science, University of Central Florida.

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

Garfield E., (1999). Journal impact factor: a brief review. **Canadian Medical Association Journal**, 161(8), 979-980.

Ingwersen P. (1998). The calculation of Web impact factors. **Journal of Documentation**, 54(2), 236-243.

Leydesdorff L. (2004). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. **Journal of documentation**, 60(4), 371-427.

Menczer P., (2004). Lexical and semantic clustering by Web links. **Journal of the American Society for Information Science and Technology,** 55(14), 1261-1269.

Thelwall M., (2002). 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., (2004). Methods for reporting on the targets of links from national systems of university Web sites. **Information processing and management**, 40(1), 125-144.

Watts D. J. and Strogatz S. H. (1998). Collective dynamics of small world networks. **Nature**, 393(6684), 440-442.

Received 15/Dec/2005
Accepted 31/Jan/2006

DISCUSSION

Discussion of "Operationalising "Websites": lexically, semantically or topologically?"
Gaston Heimeriks