Link Analysis: An Informetric Technique

Mike Thelwall and Nigel Payne

{m.thelwall, n.payne} @wlv.ac.uk
School of Computing and IT, University of Wolverhampton, Wufruna Street, Wolverhampton WV1 1SB (UK))

Introduction

In the early years of the web, several information scientists recognised the structural similarity between hyperlinks and citations, noticing that both are inter-document connections and pointers (Larson, 1996; Rodríguez i Gairín, 1997; Rousseau, 1997). This underpinned the creation of a new field, webometrics (Almind & Ingwersen, 1997), defined to be the application of quantitative techniques to the web, using methods drawn from informetrics (Björneborn & Ingwersen, 2004).

The power of the web could first be easily tapped for link analysis when commercial search engines released interfaces that allowed link searches (Ingwersen, 1998; Rodríguez i Gairín, 1997). For example, from 1997 it was possible with AltaVista to submit extremely powerful queries, such as for the number of pages in the world that linked to Swedish pages (Ingwersen, 1998). This meant that with a few hours work submitting search engine queries, the 'impact' of sets of web sites could be compared, assuming links, like citations, to measure the impact of published information. In citation analysis, researchers typically need to pay for access to the citation database of the Institute for Scientific Information, but for link analysis the web 'database' is free, potentially giving it a wider set of users.

Using commercial search engines the impact of many entities were compared, including journals, countries, universities or departments within a country and library web sites (An & Qiu, 2004; Harter & Ford, 2000; Ingwersen, 1998; Smith, 1999; Tang & Thelwall, 2005, to appear; Thomas & Willet, 2000). The early studies showed that care was needed to conduct appropriate link analyses because of many complicating factors such as duplicate web pages and sites, errors in search engine reporting, incomplete search engine coverage of the web, link replication within a site, and spurious or trivial reasons for link creation (Bar-Ilan, 2001; Egghe, 2000; Harter & Ford, 2000; Smith, 1999; Snyder & Rosenbaum, 1999; van Raan, 2001). Nevertheless, link analysis has produced interesting and useful results and has been adopted by several non-information science fields.

Data Sources

Link data can be obtained from commercial search engines, borrowed from web link databases or obtained directly with a link crawler.

At the time of writing, Google's link command could be used to count the number of pages in Google's database that contain a hypertext link to any given page. For example, `link:www.wlv.ac.uk/disclaimer.htm` reports the number of pages linking to this URL. AltaVista's `linkdomain:` command, in contrast, counts the number of pages known by AltaVista to link to any page in the specified domain, a more general search. For instance, the results of `linkdomain:www.wlv.ac.uk` would include all pages linking to any page in the www.wlv.ac.uk domain, not just to the home page. A researcher can conduct a relational analysis of the links between a set of web pages (Google) or sites (AltaVista) by obtaining link counts for all pairs individually from the search engine.

Commercial search engines have problems of coverage (i.e. not crawling some sites and crawling others incompletely), and so are not optimal for link analysis, although their use is often unavoidable (Thelwall, 2004). There is a collection of web link databases online at cybermetrics.wlv.ac.uk/database that includes the university web site link structures of five countries and is free for use by other researchers, including tools to analyse the results in various ways. A free web crawler is available at linkanalysis.wlv.ac.uk for those who need to gather their own data. This can crawl sites of up to 5,000 pages, but is not suitable for very large sites.

Future Directions For Link Analysis

Within information science there are several promising future research directions at the moment.

- *Investigations into why links are created, particularly in academic contexts*. Although there have been a few such studies already (Bar-Ilan, 2004a, 2004b; Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004; Wilkinson, Harries, Thelwall, & Price, 2003), the findings have emphasized the variety of link creation motivations used. This variety makes link classification studies difficult, but it would be interesting to know more about differences in link creation motivations.

- *Time series analyses*. One problem endemic to web link analyses is that the web is

continuously evolving and hence any web study may be out of date by the time it is published in the academic literature. Hence it is very important to know how all types of web link analysis results vary over time. A low rate of variation would lengthen the 'shelf-life' of webometric results.

- *Applying social network analysis measures to information collections*. Following Björneborn (2004), this is a type of research that needs to be applied to web information in order to fully assess its value and give new insights into the structure of information and online groupings such as invisible colleges of academics (Caldas, 2003). One issue that will need to be resolved – perhaps differently in every study – is the fact that link creation is not endemic: the lack of a link between two web sites or pages does not imply that they are unrelated.

- *Supporting wider social sciences research*. Since the web is not exclusively an academic space, it can be used in wider social science research both as an object in its own right (e.g. to study online communities) and as an easily accessible source of information about offline phenomena that happen to be reflected in the web.

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.

An, L., & Qiu, J. P. (2004). Research on the relationships between Chinese journal impact factors and external web link counts and web impact factors. *Journal of Academic Librarianship, 30*(3), 199-204.

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

Bar-Ilan, J. (2004a). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics, 59*(3), 391-403.

Bar-Ilan, J. (2004b). Self-linking and self-linked rates of academic institutions on the Web. *Scientometrics, 59*(1), 29-41.

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space - a Library and Information Science Approach.* Royal School of Library and Information Science, Copenhagen, Denmark.

Björneborn, L., & Ingwersen, P. (2004). Towards a basic framework for Webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216-1227.

Caldas, A. (2003). Are newsgroups extending 'invisible colleges' into the digital infrastructure of science? *Economics of Innovation and New Technology, 12*(1), 43-60.

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

Harries, G., Wilkinson, D., Price, E., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science, 30*(5), 436-447.

Harter, S., & Ford, C. (2000). Web-based analysis of e-journal Impact: Approaches, problems, and issues. *Journal of American Society for Information Science, 51*(13), 1159-1176.

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

Larson, R. R. (1996). *Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace.* Paper presented at the ASIS 59th annual meeting.

Rodríguez i Gairín, J. M. (1997). Valorando el impacto de la información en Internet: AltaVista, el "Citation Index" de la Red. *Revista Española de Documentación Científica, 20*(2), 175-181.

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

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

Snyder, H. W., & Rosenbaum, H. (1999). Can search engines be used for Web-link analysis? A critical review. *Journal of Documentation, 55*(4), 375-384.

Tang, R., & Thelwall, M. (2005, to appear). A hyperlink analysis of US public and academic libraries' Web sites. *Library Quarterly*.

Thelwall, M. (2004). *Link Analysis: An Information Science Approach*. San Diego: Academic Press.

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

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

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science, 29*(1), 49-56.