A History of Webometrics

by Mike Thelwall

EDITOR'S SUMMARY

The application of bibliometric and informetric approaches to study the web, its information resources, structures and technologies, is known as webometrics. Since the name was coined in 1997, the value of webometrics quickly became established through the Web Impact Factor, the key metric for measuring and analyzing website hyperlinks. Link analysis became more focused as link impact analysis and link network analysis, taking the quantity of links as a reflection of research productivity or prestige. Web citation analysis developed to facilitate investigations of links to journal articles, and analysis of keywords and phrases enables linking other types of web content. While webometrics is based in the theory of citation analysis, its methodology and software contributions may offer the greatest value and widest applicability. A study of the ASIS&T site demonstrates link network analysis.

KEYWORDS

webometrics
link analysis
network analysis
citation analysis
research methods
information science history

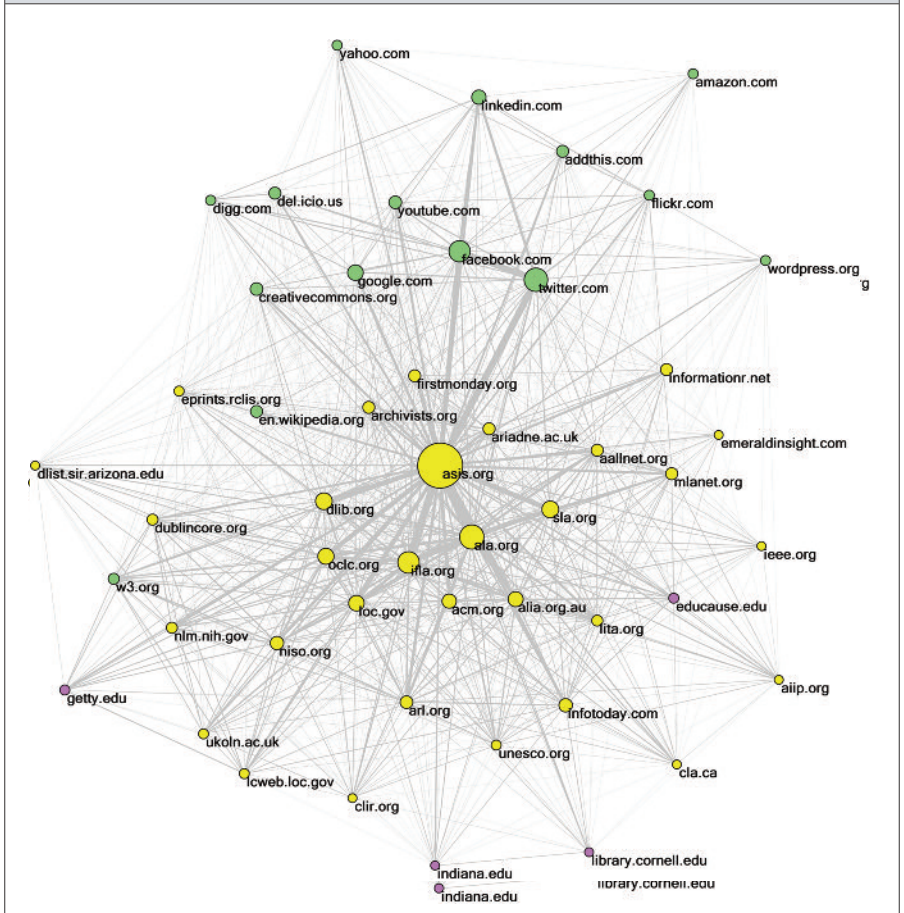
Mike Thelwall is a member of the Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wolverhampton, UK. He can be reached by email at m.thelwall<at>wlv.ac.uk.

The information science field of webometrics is "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the web drawing on bibliometric and informetric approaches" [1] or, more generally, "the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study"[2]. While the former definition emphasizes the informetric heritage of many bibliometric methods, the latter focuses on the value that webometrics could provide to the wider social sciences, reflecting a shift in webometrics over time from more theoretical studies to more applied studies, though retaining an emphasis on methods development. Webometrics currently provides a range of methods and software for various kinds of quantitative analyses of the web, and, despite initial concerns that web data would always be easily manipulated because they are not quality-controlled, the advocates of webometrics claim that it is useful both for studies of aspects of the web itself, such as hyperlinking among academic websites, and studies of offline phenomena that might be reflected online, such as political attitudes reflected in blogs.

The term *webometrics* was coined in 1997 by Tomas Almind and Peter Ingwersen in recognition that informetric analyses could be applied to the web [3; see also 4]. The field really took off, however, with the introduction of the Web Impact Factor (WIF) metric to assess the impact of a website or other area of the web based upon the number of hyperlinks pointing to it [5]. WIFs seemed to make sense because more useful or important areas of the web would presumably attract more hyperlinks than average. The logic of this metric was derived from the importance of citations in journal impact factors, but WIFs had the advantage that they could be easily calculated

using the new advanced search queries introduced by AltaVista, a leading commercial search engine at the time. Webometrics subsequently rose to become a large coherent field within information science, at least from a bibliometric perspective [6, 7], encompassing link analysis, web citation analysis and a range of other web-based quantitative techniques. In addition, webometrics became useful in various applied contexts, such as to construct the world webometrics ranking of universities [8, 9] and for scientometric evaluations or investigations of bodies of research or research areas [10]. This article reviews a few key areas of webometrics and summarizes its contribution to information science research.

FIGURE 1. A web environment network for ASIS&T created by Webometric Analyst.



Link Analysis: Impact Measurements and Networks

Link analysis drove early webometrics research, primarily through a combination of the development of improved methods and applications to a range of different contexts. Two types of studies emerged, link impact analyses and link network analyses. Link impact studies essentially compare the numbers of hyperlinks pointing to each website within a pre-defined set, such as all universities in a country or all departments within a discipline in a country. Links to university websites and, in some cases departmental websites, were found to correlate significantly with measures of research productivity or prestige, giving evidence of the validity of using link impact metrics as research-related indicator [10, 11, 12]. They have been used in this role to provide an indication of the most important organizations or websites within specific groups. In addition, a breakdown of the sources of links used in the calculations has been used to identify the sources of the impact, such as the country and organization types that host most of the links.

Link network research created network diagrams of the links among specified collections of websites in order to identify connectivity patterns. In addition to networks based upon direct links between pairs of sites, co-inlinks have also been used to indicate connections between pairs of sites. A co-inlink to a pair of websites A and B is a third website C that contains a hyperlink to both A and B [4]. This relation is similar to co-citation in bibliometrics and is particularly useful when investigating websites that are similar but do not necessarily hyperlink to each other. Figure 1 is an example of a co-inlink network diagram for ASIS&T. A direct link network diagram would be likely to exclude links between pairs of sites that were similar in some way but were not directly related to each other.

The nodes in the network are the websites most highly linked to from a set of 741 pages reported by Bing as containing a URL citation to “asis.org.”

Lines between websites indicate co-inlinks between them from the 741 pages. All the organizations represented should be in some way related to ASIS&T. Green nodes are general international sites and pink nodes are university sites in the United States.

Two important components of link analysis are the software and methods to extract link data. Researchers were for many years able to gather hyperlink information from commercial search engines like Bing, AltaVista and Yahoo! via their advanced link search commands, but these tools were all eventually withdrawn. Link data can still be obtained by the use of specialist link analysis web crawlers, including free programs like SocSciBot (http://socscibot.wlv.ac.uk) and IssueCrawler (www.issuecrawler.net) as well as a range of other crawlers developed by individual researchers. The IssueCrawler initiative from sociology [13] seems to have been particularly successful at spreading link analysis methods to the wider areas of social sciences and the humanities. Within information science, hyperlink-based network diagrams have been used to investigate the interconnections between large groups of organizations, such as universities in Europe [14, 15] and organizations within a specific knowledge sector [16, 17].

Some link analysis research has focused on the links themselves, investigating why they are created and why some sites or pages attract more links than others. These studies seem to have focused exclusively on links in academic contexts. Content analyses have shown that links between academic websites tend to be created for scholarly or educational reasons [18], a partial similarity with citation analysis. Statistical tests have also been used to see which attributes of the website owners (other than research productivity or production, which was already a known factor) tend to associate with higher inlink counts, for example finding that research group website owner gender is unimportant [19]. A recent quite comprehensive study used the most advanced statistical modeling approach yet on a large dataset to gain significant insights into the factors behind academic website interlinking in Europe. Among the findings were that country, region, domain specialism and level (whether awarding doctoral degrees or not) were the most important factors predicting hyperlinks, while reputation was the key factor for the top universities [20].

From Web Citation Analysis to Altmetrics

The second type of webometrics to become popular was web citation analysis: counting online citations to published academic documents like refereed journal articles. The rationale behind early research was to assess whether the web could replace traditional citation databases to assess the impact of articles in open access online journals [21] and subsequently also for all journals [22]. This early research found that although counts of web citations correlated with citation counts from traditional databases, many of the web citations derived from non-academic sources, such as online library catalogues. As a result, the web appeared to be an inferior source of citation impact evidence for journals or individual journal articles.

This strand of webometric research gave way to more specialized investigations into particular types of web citations to academic publications, such as citations from PowerPoint presentations [23], online syllabi [24] and Google Books [25] on the basis that within these restricted domains, web-based citation counts could reveal different types of impact from the scholarly impact reflected by traditional citation counts. For example, online syllabus citations could reflect the education impact or value of articles. This line of research was subsequently overtaken by the altmetrics initiative, discussed elsewhere in this issue.

A promising but relatively little studied type of webometrics is the analysis of mentions of keywords or phrases – not necessarily citations. This type of analysis was started by an investigation into the context of online mentions of academics [26], but the keyword approach has also been used to map concepts online [27] and interactions between concepts online by tracking co-words in web pages [28].

Theoretical Perspectives and Information-Centered Research

Webometrics has been a methods-centered field, developing methods to gather and analyze data from the web. Perhaps as a result of this focus, the theoretical component of most webometric studies has typically been drawn from citation analysis rather than being created specifically for web data. For example, many early studies assessed whether web citation counts or web link counts correlated with traditional citation counts, drawing upon

Robert Merton's theoretical discussion of citation norms in science. Hence, such studies assessed to some extent how well web data fitted Merton's theory. The lack of development of specialist theory for the most developed area of webometrics, link analysis, reflects the web being a far more varied and complex space than academic journal databases, with theory development in the latter being recognized as problematic and controversial [29].

One partial exception to the lack of native theory for webometrics is information-centered research [30], a style of research theorized to be particularly appropriate to webometrics. An information-centered research study focuses on a new information source, such as a type of web data, and attempts to identify the social science research problems that the data is most suited to address rather than using *a priori* intuitions to match the data with a research problem and then to assess the value of the data for the problem. This theory was used to justify the development of a range of different methods to analyze web data and to match the methods to a variety of social science problem areas.

Web Data Analysis as a Service for the Social Sciences

Webometrics research has expanded from general or academic web analyses to investigations of social websites, often by automatically downloading data from those websites either through a web crawler or through data requests sent through permitted routes (application programming interfaces). For example, exploiting the information-centered research approach, blogs and RSS feeds have been analyzed to detect public fears about science while social network sites have been investigated to detect friendship patterns and language use. Twitter has been analyzed for the sentiment of public reactions to major media events and YouTube for the factors associated with discussions attached to online videos. In all cases, the methods of the research have been webometrics – large scale data gathering and analysis for social science purposes – but the findings of the research have been targeted at disciplines outside information science, such as media studies [31], politics [32, 33, 34] and science communication [35]. Many of the programs used are now publicly available in the free software Webometric Analyst.

Summary

The field of webometrics has now developed a range of different strands of research, from link analysis to social web analysis, and is used in mainstream information science applications such as research evaluation in addition to applications in the wider social sciences. Its outputs are methods-based, including free software and a range of analysis methods.

Perhaps one advantage of webometrics that has been insufficiently exploited within information science education is that webometric software supports the quick gathering of data targeted to a topic of interest –whether defined by a specific set of websites or a set of keywords. This speed is useful for student projects since it allows students to devote more time to data analysis and research design rather than data collection. It also allows projects to be more targeted to student interests, including almost anything that is discussed significantly online. The data quality issues discussed above are even a positive advantage for student projects since they give them scope for an extensive and intelligent discussion of validity issues. ■

Resources Mentioned in the Article

[1] Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216-1227.
[2] Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences.* New York: Morgan & Claypool.
[3] Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to "Webometrics." J*ournal of Documentation, 53*(4), 404-426.
[4] Rousseau, R. (1997). Sitations: An exploratory study. *Cybermetrics, 1*(1), Retrieved June 6, 2012, from www.cybermetrics.info/articles/v1i1p1.pdf.
[5] Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation, 54*(2), 236-243.

Resources continued on next page

Resources Mentioned in the Article, cont.

[6] Åström, F. (2007). Changes in the LIS research front: Time-sliced co-citation analyses of LIS journal articles, 1990-2004. *Journal of the American Society for Information Science and Technology, 58*(7), 947-957.
[7] Zhao, D., & Strotmann, A. (2008). Information science during the first decade of the Web: An enriched author co-citation analysis. *Journal of the American Society for Information Science and Technology, 59*(6), 916-937.
[8] Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology, 57*(10), 1296-1302.
[9] Thelwall, M. (2010). Webometrics: Emergent or doomed? *Information Research, 15*(4). Retrieved July 18, 2012, from http://informationr.net/ir/15-4/colis713.html.
[10] Li, X., Thelwall, M., Musgrove, P. B., & Wilkinson, D. (2003). The relationship between the WIFs or inlinks of computer science departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics, 57*(2), 239-255.
[11] Tang, R., & Thelwall, M. (2003). U.S. academic departmental website interlinking: Disciplinary differences. *Library & Information Science Research, 25*(4), 437-458.
[12] Thelwall, M., & Harries, G. (2004). Do the websites of higher rated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology, 55*(2), 149-159.
[13] Rogers, R. (2006). *Information politics on the web.* Cambridge, MA: MIT Press.
[14] Ortega, J. L., & Aguillo, I. F. (2008). Visualization of the Nordic academic web: Link analysis using social network tools. *Information Processing & Management, 44*(4), 1624-1633.
[15] Ortega, J. L., Aguillo, I. F., Cothey, V., & Scharnhorst, A. (2008). Maps of the academic web in the European Higher Education Area: An exploration of visual web indicators. *Scientometrics, 74*(2), 295-308.
[16] Heimeriks, G., & van den Besselaar, P. (2006). Analyzing hyperlink networks: The meaning of hyperlink-based indicators of knowledge. *Cybermetrics, 10*(1). Retrieved August 1, 2011 from http://cybermetrics.cindoc.csic.es/articles/v10i1p1.html.
[17] Heimeriks, G., Hörlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics, 58*(2), 391-413.
[18] Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management, 41*(3), 973-986.
[19] Barjak, F., & Thelwall, M. (2008). A statistical analysis of the web presences of European life sciences research teams. *Journal of the American Society for Information Science and Technology, 59*(4), 628-643.
[20] Seeber, M., Lepori, B., Lomi, A., Aguillo, I., & Barberio, V. (2012). Factors affecting web links between European higher education institutions. *Journal of Informetrics, 6*(3), 435-447.
[21] Smith, A. G. (1999). A tale of two web spaces; Comparing sites using Web Impact Factors. *Journal of Documentation, 55*(5), 577-592.
[22] Vaughan, L., & Hysen, K. (2002). Relationship between links to journal websites and impact factors. *ASLIB Proceedings, 54*(6), 356-361.
[23] Thelwall, M., & Kousha, K. (2008). Online presentations as a source of scientific impact? An analysis of PowerPoint files citing academic journals. *Journal of the American Society for Information Science and Technology, 59*(5), 805-815.
[24] Kousha, K., & Thelwall, M. (2008). Assessing the impact of disciplinary research on teaching: An automatic analysis of online syllabuses. *Journal of the American Society for Information Science and Technology, 59*(13), 2060-2069.
[25] Kousha, K., & Thelwall, M. (2009). Google Book Search: Citation analysis for social science and the humanities. *Journal of the American Society for Information Science and Technology, 60*(8), 1537-1549.
[26] Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the web. *Journal of the American Society for Information Science, 49*(14), 1319-1328.
[27] Thelwall, M., Vann, K., & Fairclough, R. (2006). Web issue analysis: An integrated water resource management case study. *Journal of the American Society for Information Science and Technology, 57*(10), 1303-1314.

Resources continued on next page

Resources Mentioned in the Article, cont.

[28] Khan, G. F., & Park, H. W. (2011). Measuring the triple helix on the web: Longitudinal trends in the university-industry-government relationship in Korea. *Journal of the American Society for Information Science and Technology, 62*(12), 2443-2455.
[29] Moed, H. F. (2005). *Citation analysis in research evaluation.* New York: Springer.
[30] Thelwall, M., Wouters, P., & Fry, J. (2008). Information-centered research for large-scale analysis of new information sources. *Journal of the American Society for Information Science and Technology, 59*(9), 1523-1527.
[31] Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology, 62*(2), 406-418.
[32] Kim, J. H., Barnett, G. A., & Park, H. W. (2010). A hyperlink and issue network analysis of the United States Senate: A rediscovery of the web as a relational and topical medium. *Journal of the American Society for Information Science and Technology, 61*(8), 1598-1611.
[33] Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections, 25*(1), 49-61.
[34] Romero-Frias, E., & Vaughan, L. (2010). European political trends viewed through patterns of web linking. *Journal of the American Society for Information Science and Technology, 61*(10), 2109-2121.
[35] Koteyko, N., Thelwall, M., & Nerlich, B. (2010). *From carbon markets to carbon morality: Creative compounds as framing devices in online discourses on climate change mitigation. Science Communication, 32*(1), 25-54.