From Webometrics to Altmetrics: One and a Half Decades of Digital Research at Wolverhampton

Jonathan M. Levitt
Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK.
Mike Thelwall
Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK.

This article describes and summarises the contributions of the Statistical Cybermetrics Research Group (SCRG) at the University of Wolverhampton in the UK to the information science specialisms of Webometrics and altmetrics. In both cases the group created free computer programs for data gathering and analysis. In Webometrics the SCRG developed counting methods for hyperlink analysis and assessed them for collections of different types of website. In addition, it also developed methods for automatically gathering and analysing text on a large scale, both for web citation analysis and for more general social science purposes. It also developed two Webometric theories. In altmetrics, the SCRG analysed the validity of a range of indicators, including counts of tweets and Mendeley readers for academic articles, finding evidence that they associated with citation counts and hence that they had value as altmetrics. The dual purposes of this paper are to give an overview of a range of methods and free tools for Webometrics and altmetrics, and to give a historical overview of the evolution of one information science research group in the hope that others can learn from its successes and failures.

Introduction

The SCRG was created in December 2000 with the School of Computing and IT at the University of Wolverhampton in response to a perceived need for more computing technologies within Webometrics to address some of its central concerns. Over the next 12 years the group created two computer programs, the web crawler SocSciBot, and the data collection program Webometric Analyst, and used them to investigate Webometric issues. About half way through this period the group attempted to engage a wider social science audience for its methods and software by publishing in journals and conferences outside of information science and my customising some of its software for tasks unrelated to traditional Webometrics. In particular, the group developed methods and software for gathering and analysing tweets and for sentiment analysis. With the advent of altmetrics the group modified Webometric Analyst to gather relevant altmetric data, such as information from Mendeley, and began to investigate altmetric topics. This hagiography summarises some of the research produced by the SCRG, with a focus on altmetrics.

Webometrics

Primarily created by Tomas Almind and Peter Ingwersen in Copenhagen (Almind & Ingwersen, 1997), the research field of Webometrics was concerned with "quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches" (Björneborn & Ingwersen, 2004). It began as an attempt to develop a citation analysis of the web using hyperlinks instead of citations and extending the scope of the hyperlink citation analysis to non-academic topics. This ambitious goal was triggered by the observation that one of the major search engines at the time, AltaVista, had become a citation index (Ingwersen, 1998; Rodríguez i Gairín, 1997) for web hyperlinks through its introduction of methods to search for hyperlinks online. New research was needed, however, to assess the accuracy and comprehensiveness of AltaVista's results and the results of other search engine that followed AltaVista's lead (Bar-Ilan, 1999; Rousseau, 1999). The SCRG attempted to contribute to this debate by developing the web crawler SocSciBot to crawl academic websites and to report the number of hyperlinks between websites in order to help check search engine results, and later also in an attempt to improve on them (Thelwall, 2002).

A second technological development by commercial search engines then changed Webometrics: The provision of Application Programming Interfaces (APIs). These allowed programmers to gain automatic access to search engine results and made it possible to automate the gathering of data for webometric purposes. In response, the SCRG developed a new computer program, LexiURL Searcher (now called Webometric Analyst and also used for altmetrics) to interface with the major search engines to automatically download webometric data. This made much larger scale studies possible using APIs from Google, Microsoft and Yahoo! (e.g., Kousha, & Thelwall, 2008a).

Data gathering for Webometrics became more difficult when the commercial search engines withdrew some or all of their facilities. Currently, no major search engine allows useful hyperlink searches and so it is no longer possible to conduct automated hyperlink data gathering from a major commercial search engine. Moreover, only Bing now offers free API for searches. In response, the SCRG resumed development on its web crawler SocSciBot and developed new types of query for Webometric Analyst that identified citation-like types of inter-document connection that could be searched for automatically in Bing and used as substitutes for hyperlinks. These *URL citations* were mentions of the URL of a target page or website in another website (Kousha & Thelwall, 2007; Stuart & Thelwall, 2006). For example the following query matches pages within the University of Wolverhampton website (www.wlv.ac.uk) that mention the URL of any page in the main BBC News website (news.bbc.co.uk):

```
"news.bbc.co.uk" site:wlv.ac.uk
```

The SCRG developed and applied link analysis for assessing the impact of websites (e.g., Thelwall & Harries, 2004) and also for creating networks of websites built through the links between them (e.g., Thelwall & Zuccala, 2008). In support of the software and methods, the group also introduced a theoretical framework for link analysis theory to guide link analysis research by specifying a minimum set of analyses needed to generate a meaningful link analysis study (Thelwall, 2006). For example, the Framework included content analysis of a random sample of links in order to be able to infer meaning from the network diagrams or link counts generated in a study. The link analysis methods were applied, sometimes in conjunction with other researchers, both inside information science (Barjak & Thelwall, 2008; Eccles, Thelwall, & Meyer, 2012; Mas Bleda, Thelwall, Kousha, & Aguillo, 2014; Tang & Thelwall, 2004) and in the wider social sciences and humanities (Park & Thelwall, 2008). In the latter case the SCRG's goal was to expand Webometrics to analyse "web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study" (Thelwall, 2009).

In addition to variants of link analysis, the SCRG developed text analysis methods for the web, such as a technique to extract trends from news reports delivered from blogs and news websites in RSS format (Thelwall & Prabowo, 2007), later adapting the same methods to identify trends in Twitter (Wilkinson & Thelwall, 2012). At the same time, the SCRG collaborated in the creation of a new theory, that of Information-Centred Research, which posited that information scientists should explore new web-based data sources in order to identify the disciplines in which they may be useful and the methods that may be useful for extracting data from them (Thelwall, & Wouters, 2005; Thelwall, Wouters, & Fry, 2008). This theory essentially argued that information scientist could be pro-active librarians for the web, directing researchers to useful tools and data sources for their problem.

An increasingly important strand of research within webometrics was the generation of metrics for the impact of academic articles using evidence from web searches for mentions of them (following from a previous person-mention approach: Cronin, Snyder, Rosenbaum, Martinson, & Callahan, 1998). These web citations allowed web-based citation analyses to be conducted on a much larger scale and with more data than had been possible with earlier hyperlink-based citation studies. The first research used general searches to look for web citations to academic articles from any web page (Vaughan & Shaw, 2003). Later investigations instead constructed searches for specific types of web page, such as online PowerPoint presentations, blogs or course syllabuses in order to get web indicators for specific types of impact, such as educational impact (Kousha & Thelwall, 2008ab; Kousha, Thelwall, & Rezaie, 2010). At the same time, Google Books was assessed for its ability to report citations from books to books or journal articles, with the findings suggesting that it was possible to automatically extract useful book-based citations from this source (Abdullah & Thelwall, in press; Kousha & Thelwall, 2009; Kousha & Thelwall, in press; Kousha, Thelwall, & Rezaie, 2011).

Altmetrics

The field of altmetrics was created by a group of US and European researchers led by Jason Priem in to study the potential to develop indicators for aspects of the impact or uptake of academic articles through indicators extracted from the social web, using APIs (Priem & Hemminger, 2010; Priem, Taraborelli, Groth, & Neylon, 2010). Altmetrics had become possible because reference sharing sites, such as Mendeley, and social network sites like Twitter were being used by significant numbers of people to share research, creating a large public body of data about the use and sharing of academic articles. Moreover, the companies owing the social web sites often made data collection from them by computer programs possible by offering public API access. Given that academic articles are normally evaluated on a large scale by counting citations to them, two of the key promises of altmetrics were that they could reflect wider uses of articles than just those that led to citations (e.g., educational uses, and uses by practitioners) and that they could be collected much more quickly than could citations, so that altmetrics could be used as indicators for articles soon after publication even though citations might take a year to start to accumulate. This is particularly important for information retrieval since people are often most concerned with research that has been recently published (e.g., for horizon scanning).

The SCRG started to investigate altmetrics as a logical extension of its web citation analysis research, mentioned above, and incorporated citation search facilities into its free Webometric Analyst software (http://lexiurl.wlv.ac.uk) for the social reference sharing site Mendeley via its API, as well as features for monitoring Twitter via its API. These facilities were then used to test altmetrics. As for web citation analysis studies, the default initial method to test a new altmetric was to correlate its values against citations from an existing citation database, such as the Web of Science or Google Scholar, with a statistically significant positive correlation being taken as some evidence that the results were not random and were related to scholarly activities in some way, even if not through a cuse-and-effect relationship (Sud, & Thelwall, 2014). The correlation method was used to demonstrate the existence of an association between Mendeley "readers" of an article and its citations (Li, Thelwall, & Giustini, 2012; Mohammadi & Thelwall, in press), for citations from blogs (Shema, Bar-Ilan, & Thelwall, in press), and for scores from the Faculty of 1000 website (Li & Thelwall, 2012). These Faculty of 1000 scores were later shown to be capable of revealing articles that were medically useful despite not attracting many citations, hence performing a useful research evaluation task (Mohammadi & Thelwall, 2013).

The correlation method was found to be inappropriate for some altmetrics because the increasing use of social web sites like Twitter for academic purposes meant that younger articles tended to be mentioned (e.g., tweeted) more due to the increasing use of the site. In response, an alternative method was developed to identify an association between altmetrics and citations that would not be affected by the increasing use of social web sites. This method was used to demonstrate that more tweeted articles tended to be more cited across a range of journals (Thelwall, Haustein, Larivière, & Sugimoto, 2013).

Future work

In addition to Webometrics and altmetrics, the SCRG also conducts sentiment analysis (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010; Thelwall, Buckley, & Paltoglou, 2012; Thelwall, & Buckley, 2013; Thelwall, Buckley, & Paltoglou, 2011) and traditional scientometric research, such as into collaboration (Levitt & Thelwall, 2009; Levitt & Thelwall, 2010; Thelwall & Sud, 2014) and factors associating with high impact articles (Didegah, & Thelwall, 2013ab; Levitt, & Thelwall, 2011). A recent trend within the group that is likely to continue in the future is to use more sophisticated statistical techniques in order to analyse data with multiple simultaneous factors in order to identify which factors are important and which seem to be important because of their association with other factors. For example, one study found evidence that international collaboration tends to be more highly cited not because of the involvement of multiple countries, as had previously been thought, but because of the involvement of additional authors, at least in biochemistry (article currently under review). In addition to the inclusion of more statistical approaches, in the future the group will continue to seek opportunities to exploit new websites or changed in the web for research purposes.

Summary

Overall, the SCRG has attempted to combine (a) an element of web computing in the sense of writing (and sharing) computer programs to gather and analyse data from the web, and (b) simple statistical methods to analyse the data in order to address web-related research questions related to scholarly communication. The purpose of most of the research has been methodological: to develop and assess new methods. In contrast, relatively few articles have focused on discovering something using web data that is irrelevant to the web. Hence the research has had a strong methods focus. Whilst some of the early research described above has become obsolete because of changes in the web and in the services provided by search engines, the overall strand of research has managed to survive through developing existing techniques to address new challenges, such as the rise of the social web and the introduction of altmetrics. As predominantly methods-oriented researchers, however, the success of the group is in the uptake of its methods by others and only time will reveal the extent to which this happens.

References

Abdullah, A. & Thelwall, M. (in press). Can the impact of non-Western academic books be measured? An investigation of Google Books and Google Scholar for Malaysia. Journal of the Association for Information Science and Technology.

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

Bar-Ilan, J. (1999). Search engine results over time - A case study on search engine stability. Cybermetrics, 2/3. http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

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.

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.

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.

Didegah, F., & Thelwall, M. (2013a). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. Journal of Informetrics, 7(4), 861-873.

Didegah, F. & Thelwall, M. (2013b). Determinants of research citation impact in nanoscience and nanotechnology. Journal of the American Society for Information Science and Technology, 64(5), 1055–1064.

Eccles, K.E., Thelwall, M., & Meyer, E.T. (2012). Measuring the web impact of digitised scholarly resources. Journal of Documentation, 68(4), 512-526.

Ingwersen, P. (1998). The calculation of web impact factors. Journal of documentation, 54(2), 236-243.

Kousha, K. & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis, Journal of the American Society for Information Science and Technology, 57(6), 1055-1065.

Kousha, K. & Thelwall, M. (2008a). 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.

Thelwall, M. & Kousha, K. (2008b). 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.

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.

Kousha, K. & Thelwall, M. (in press). An automatic method for extracting citations from Google Books. Journal of the Association for Information Science and Technology.

Kousha, K., Thelwall, M., & Rezaie, S. (2010). Using the web for research evaluation: the integrated online impact indicator. Journal of informetrics, 4(1), 124-135.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. Journal of the American Society for Information Science and Technology, 62(11), 2147-2164.

Levitt, J., & Thelwall, M. (2009). Citation levels and collaboration within Library and Information Science, Journal of the American Society for Information Science and Technology, 60(3), 434-442.

Levitt, J., & Thelwall, M. (2010). Does the higher citation of collaborative research differ from region to region? A case study of economics, Scientometrics, 85(1), 171-183.

Levitt, J., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. Information Processing & Management, 47(2), 300-308.

Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. In Proceedings of the 17th International Conference on Science and Technology Indicators. Montréal, Canada (pp. 451-551).

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. Scientometrics, 91(2), 461-471.

Mas Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I. (2014). Successful researchers publicizing research online: An outlink analysis of European highly cited scientists' personal websites. Journal of Documentation, 70(1), 148-172.

Mohammadi, E. & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. Scientometrics, 97(2), 383-395.

Mohammadi, E. & Thelwall, M. (in press). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology.

Park, H. W., & Thelwall, M. (2008). Developing network indicators for ideological landscapes from the political blogosphere in South Korea. Journal of Computer-Mediated Communication, 13(4), 856-879.

Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. First Monday, 15(7). http://firstmonday.org/ojs/index.php/fm/article/view/2874

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. altmetrics.org.

Rodríguez i Gairín, J. M. (1997). Valoración del impacto de la información en Internet: AltaVista, el Citation Index de la red. Revista española de documentación científica, 20(2), 175-181.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight, Cybermetrics, 2/3. http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Shema, H., Bar-Ilan, J., & Thelwall, M. (in press). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. Journal of the Association for Information Science and Technology.

Stuart, D. & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: A case study of the UK West Midlands automobile industry. Research Evaluation, 15(2), 97-106.

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. Scientometrics, 98(2), 1131-1143.

Tang, R. & Thelwall, M. (2004). Patterns of national and international web inlinks to US academic departments: An analysis of disciplinary variations. Scientometrics, 60(3), 475-485.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.

Thelwall, M. & Buckley, K. (2013). Topic-based sentiment analysis for the Social Web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8), 1608–1617.

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*.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. PloS ONE, 8(5), e64841.

Thelwall, M. & Prabowo, R. (2007). Identifying and characterising public science-related fears from RSS feeds. Journal of the American Society for Information Science and Technology, 58(3), 379-390.

Thelwall, M. & Sud, P. (2014). No citation advantage for monograph-based collaborations? Journal of Informetrics, 8(1), 276-283.

Thelwall, M., Wouters, P., & Fry, J. (2008). Information-Centred Research for large-scale analysis of new information sources, Journal of the American Society for Information Science and Technology, 59(9), 1523-1527.

Thelwall, M. & Wouters, P. (2005). What’s the deal with the web/Blogs/the next big technology: A key role for information science in e-social science research? CoLIS 2005, Lecture Notes in Computer Science 3507, 187-199.

Thelwall, M. & Zuccala, A. (2008). A university-centred European Union link analysis, Scientometrics, 75(3), 407-420.

Thelwall, M. (2002). A comparison of sources of links for academic Web Impact Factor calculations. Journal of Documentation, 58(1), 66-78.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, 57(1), 60-68.

Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. Synthesis lectures on information concepts, retrieval, and services. New York: Morgan & Claypool.

Thelwall, M., & Harries, G. (2004). Do the Web sites of higher rated scholars have significantly more online impact? Journal of the American Society for Information Science and Technology, 55(2), 149-159.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: what is the difference? Journal of the American Society for Information Science and Technology, 54(14), 1313-1322.

Wilkinson, D. & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison. Journal of the American Society for Information Science and Technology, 63(8), 1631-1646.