Data Science Altmetrics

Mike Thelwall†

Statistical Cybermetrics Research Group, University of Wolverhampton, UK

Mike Thelwall is a Professor of Information Science and leader of the Statistical Cybermetrics Research Group at the University of Wolverhampton, UK. He is also Docent at the Department of Information Studies at Åbo Akademi University, and a research associate at the Oxford Internet Institute. His current research includes identifying and analyzing Web phenomena using quantitative-led research methods, including altmetrics and sentiment analysis, and he has pioneered an information science approach to link analysis. He has developed a wide range of tools for gathering and analyzing Web data, including hyperlink analysis, sentiment analysis and content analysis for Twitter, YouTube, MySpace, blogs and the Web in general. His more than 450 publications include 272 refereed journal articles and two books, including Introduction to Webometrics. He is an Associate Editor of the *Journal of the Association for Information Science and Technology* (*JASIST*) and an Editor of the *Journal of Data and Information Science* (*JDIS*) and sits on three other editorial boards. For more information, visit http://www.scit.wlv.ac.uk/~cm1993/mycv.html.

Citation: Mike Thelwall (2016). Data Science Altmetrics.

Received: Feb. 20, 2016
Accepted: Mar. 10, 2016

Introduction

Within the field of scientometrics, which involves quantitative studies of science, the citation analysis specialism counts citations between academic papers in order to help evaluate the impact of the cited work (Moed, 2006). The reason for this is that if the research reported in a publication makes an important contribution to knowledge it is reasonable to expect it to be cited by other papers that build upon it (Merton, 1973). While there are many exceptions to this rule, citation analysis in many different forms is now widely used and many would accept that, if citations are counted on a large enough scale and with appropriate safeguards for field, time and other differences, citation-based indicators can give a reasonable approximation to the average impact of a body of work. The main advantage of using citation indicators instead of peer review is that peer review is costly and needs a substantial investment in time from experts that might well prefer to be researching rather than evaluating.

† Corresponding author: Mike Thelwall (E-mail: m.thelwall@wlv.ac.uk).

JDIS
Journal of Data and Information Science
Vol. 1 No. 2, 2016
pp 7–12
DOI: 10.20309/jdis.201610
http://www.jdis.org

Citation counts are rarely used on their own but are normally processed in order to generate more informative indicators. One of the most well-known examples of an indicator derived from citation counts is the Thomson Reuters Journal Impact Factor (JIF), which assesses (approximately speaking) the average number of citations per paper for recent articles published in the journal. Following the logic above, a journal with a higher JIF seems likely to have published articles that have made a greater contribution to science than a journal with a lower JIF. There are many limitations to the previous statement so that it is only very broadly accurate. One limitation is that citation counts are not comparable between disciplines. Another limitation is that academic research that has a beneficial impact on society rather than science might not be cited and so applied research may be undervalued by citation counts. This issue is tackled to some extent by altmetrics.

Altmetrics

Altmetrics are relatively new but maturing specialism within scientometrics that is concerned with extracting information from the social Web about the impacts of academic research. The rationale for altmetrics is that posts in the social Web are written by the general public and so indicators derived from the social Web may reflect the value of research to society rather than the value of research to future scholarship. Altmetrics may therefore be useful as an additional source of impact evidence when citation counts are used.

The field of altmetrics was created in 2010 (Priem & Hemminger, 2010; Priem, Taraborelli, Groth, & Neylon, 2010) and has rapidly generated a substantial body of research (for reviews, see: Kousha & Thelwall, 2015; Thelwall & Kousha, 2015a, b). The raw data for altmetrics can be automatically extracted on a large scale from social websites using applications programming interfaces (APIs) and can also be freely downloaded from some altmetric data providers, such as Altmetric.com. The extracted data can then be processed to produce indicators that can be used as evidence of the impact of scholarly work. For example, the social websites Twitter and Weibo have free APIs that can be used to download recent posts. A researcher might use these APIs to count how often a set of academic articles have been tweeted about, and the tweet or Weibo count might then be the altmetric. Other social websites that could be used for altmetrics include Facebook, Google+, Reddit, LinkedIn, and Mendeley (Thelwall et al., 2013; Zahedi, Costas, & Wouters, 2014).

In practice, altmetrics is very similar to webometrics, where indicators are derived from the general Web rather than from the social Web. For example, it is possible to count how many Web pages cite any given academic article and use this as evidence of its impact (Vaughan & Shaw, 2003). The data for webometrics are often

collected with search engine API queries rather than queries for specific websites (e.g., using Webometric Analyst[①]). This makes it possible to collect large scale data from the general Web for webometrics. Indicators derived from the online book site Google Books are a special case because Google Books has an API that is useful for collecting data (Kousha & Thelwall, 2015a) even though Google Books not a social website. Website log data can also give useful and somewhat similar information (Bollen et al., 2009) but few researchers are able to access the log files of important scholarly websites for this.

On a small scale, an individual might include altmetrics on their curriculum vitae (CV) to support a claim about the type of impact that it had gathered (Piwowar & Priem, 2013). A limitation of this is that social websites are easily spammed and so altmetrics can never form *strong* evidence of research impact (Wouters & Costas, 2012). Altmetrics are also used by some academic publishers to display alongside articles in their digital libraries in order to show visitors which articles have received the most attention in the social Web (Adie & Roe, 2013).

Data Science

Altmetric data are suitable for data science because they are relatively easy to gather on a large scale and there are many interesting practical and theoretical problems that can be investigated with it.

The most promising site for data science altmetrics is Mendeley.com because it focuses on academic research and so its data do not need to be filtered to extract academic-related content. Mendeley is a social reference sharing site (Henning & Reichelt, 2008). It is used by researchers to keep track of articles that they want to read or have read (Mohammadi, Thelwall, & Kousha, 2016). Mendeley reader counts have a strong positive correlation with citation counts (Li & Thelwall, 2012), confirming that they can be analyzed in similar ways to citation counts. Through the Mendeley API, it is possible to extract the number of people that have registered an article within the site. Mendeley also reports the country that the users are from and their academic discipline and status (e.g., researcher or professor). This additional information can be used for data mining to look for international, temporal or disciplinary differences. For example, one study showed that articles tended to be read more often by people from the same country as the authors (Thelwall & Maflahi, 2015).

Large scale altmetric data can be used to assess the validity of specific altmetrics (altmetric) by investigating the extent to which the altmetric correlates with citation

① http://lexiurl.wlv.ac.uk

counts (Sud & Thelwall, 2014). Although some correlations of this type have already been calculated, it is important to calculate more correlations for different sets of documents in order to get a fuller picture of how the correlations vary between disciplines and document types as well as over time. In addition, there is also scope to investigate the challenges facing publishers that are attempting to fully exploit altmetric indicators, such as creating the most informative types of indicators (see also: Lin & Fenner, 2013; Liu & Adie, 2013).

The above research directions are focused on developing and validating altmetrics but there is considerable potential to apply data analytic approaches to investigate each altmetric indicator in more depth. For example, this may be achieved by developing methods to distinguish between different types of authors of the social Web posts used (e.g., Mohammadi et al., 2015). There is even more scope to use altmetric and other Web indicators to investigate scholarly communication more generally (Mohammadi & Thelwall, 2014) or the interface between the public and science (Sugimoto & Thelwall, 2013). For example, which disciplines and types of paper are most interesting to the public or most valued? What types of comment do members of the public make about academic research? Are there hidden patterns of commenting about research that a data mining approach might discover?

Statistical modelling is also relevant to altmetrics data science research (altmetrics and data science research) because few papers have investigated their distributions so far. One exception has shown that the discretized lognormal and hooked power law distributions fit Mendeley reader counts for individual subjects and years, but neither distribution fits all datasets best (Thelwall & Wilson, in press). It would therefore be useful to investigate the distributions of additional altmetrics and to assess whether other distributions might fit the data better.

One data science technique that should be useful for studying altmetrics but has not been used yet is natural language processing. This is because the context in which articles are mentioned in the social Web is particularly important and natural language processing methods can give the most detailed information about this context. These methods are already used in scientometrics (e.g., Li et al., 2012), although not on a large scale. Natural language processing could only be used for altmetrics where a citation is associated with text, as in the case of Twitter and Sina Weibo, but not when there is no associated text, as in the case of Mendeley.

In summary, altmetrics is a promising research topic for data science because it has a rich supply of data and challenging problems. Researchers that are able to gather relevant data and process them on a large scale can expect to make substantial contributions to scholarship.

References

Adie, E., & Roe, W. (2013). Altmetric: Enriching scholarly content with article-level discussion and metrics. Learned Publishing, 26(1), 11–17.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. PLOS One, 4(3), e4803.

Henning, V., & Reichelt, J. (2008). Mendeley-a last.fm for research? In IEEE Fourth International Conference on eScience (pp. 327–328). Los Alamitos, CA: IEEE Press.

Kousha, K., & Thelwall, M. (2015a). An automatic method for extracting citations from Google Books. Journal of the Association for Information Science and Technology, 66(2), 309–320.

Kousha, K., & Thelwall, M. (2015b). Web indicators for research evaluation. Part 3: Books and non standard outputs. El Profesional de la Información, 24(6), 724–736.

Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. In Proceedings of the 17th International Conference on Science and Technology Indicators (Vol. 2, pp. 451–551). Montreal Quebec, Canada.

Li, Z., Tate, D., Lane, C., & Adams, C. (2012). A framework for automatic TRIZ level of invention estimation of patents using natural language processing, knowledge-transfer and patent citation metrics. Computer-aided Design, 44(10), 987–1010.

Lin, J., & Fenner, M. (2013). Altmetrics in evolution: Defining and redefining the ontology of article-level metrics. Information Standards Quarterly, 25(2), 20.

Liu, J., & Adie, E. (2013). Five challenges in altmetrics: A toolmaker's perspective. Bulletin of the American Society for Information Science and Technology, 39(4), 31–34.

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. Chicago, IL: University of Chicago Press.

Moed, H. F. (2006). Citation analysis in research evaluation. Berlin: Springer Science & Business Media.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. Journal of the Association for Information Science and Technology, 66(9), 1832–1846.

Mohammadi, E., Thelwall, M. & Kousha, K. (2016). Can Mendeley bookmarks reflect readership? A survey of user motivations. Journal of the Association for Information Science and Technology, 67(5), 1198–1209. doi:10.1002/asi.23477.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, 65(8), 1627–1638.

Piwowar, H., & Priem, J. (2013). The power of altmetrics on a CV. Bulletin of the American Society for Information Science and Technology, 39(4), 10–13.

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

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

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

Sugimoto, C.R., & Thelwall, M. (2013). Scholars on soap boxes: Science communication and dissemination via TED videos. Journal of the American Society for Information Science and Technology, 64(4), 663–674.

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., & Kousha, K. (2015a). Web indicators for research evaluation. Part 1: Citations and links to academic articles from the Web. El Profesional de la Información, 24(5), 587–606.

Thelwall, M., & Kousha, K. (2015b). Web indicators for research evaluation. Part 2: Social media metrics. El Profesional de la Información, 24(5), 607–620.

Thelwall, M., & Maflahi, N. (2015). Are scholarly articles disproportionately read in their own country? An analysis of Mendeley readers. Journal of the Association for Information Science and Technology, 66(6), 1124–1135. DOI:10.1002/asi.23252.

Thelwall, M., & Wilson, P. (in press). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. Journal of the Association for Information Science and Technology. DOI:10.1002/asi.23501.

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.

Wouters, P., & Costas, R. (2012). Users, narcissism and control: Tracking the impact of scholarly publications in the 21st century (pp. 847–857). Utrecht: SURF foundation.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics, 101(2), 1491–1513.