Re-citation Analysis: A Promising Method for Improving Citation Analysis for Research Evaluation, Knowledge Network Analysis, Knowledge Representation and Information Retrieval

Dangzhi Zhao[1] and Andreas Strotmann[2]

[1] *dzhao@ualberta.ca*
School of Library and Information Studies, University of Alberta, Edmonton (Canada)

[2] *andreas.strotmann@gmail.com*
ScienceXplore, F.-G.-Keller-Str. 10, D-01814 Bad Schandau (Germany)

Abstract
Citation analysis is used in research evaluation exercises around the globe, directly affecting the lives of millions of researchers and the expenditure of billions of dollars. It is therefore crucial to seriously address the problems and limitations that plague it. Central amongst critiques of the common practice of citation analysis has long been that it treats all citations equally, be they crucial to the citing paper or perfunctory. Weighting citations by their value to the citing paper has long been proposed as a theoretically promising solution to this problem. *Re-citation analysis* proposes to tune out the large percentage of perfunctory citations in a paper and tune in on crucial ones when performing citation analysis, by ignoring uni-citations (mentioned just once in a paper) and counting and analyzing only re-citations (used again and again in a citing paper). By focusing on core connections in knowledge networks, re-citation analysis can help research evaluation become more sensitive to the distinction between essential and perfunctory impact of research. It may benefit citation-link based knowledge representation and retrieval systems with improved precision by better capturing "aboutness" of articles, the essence of subject indexing in knowledge representation and retrieval, rather than merely providing "relatedness" information.

Conference Topic
Theory; Methods and techniques

Introduction
Citation analysis is used in research evaluation exercises around the globe, directly affecting the work and lives of millions of researchers and the expenditure of billions of dollars. It is therefore crucial to seriously address the problems and limitations that plague it. Central amongst critiques of the current practices of citation analysis has long been that it treats all citations equally, be they crucial to the citing paper or perfunctory. This problem is especially serious when tracing or assessing research impact.

Weighting citations by how they are used in the citing paper has therefore long been proposed as a theoretically promising solution to this problem, but in practice it has not been studied closely at a large scale until recently. Increasingly available digital full-text documents and advances in text processing technologies are now making it feasible to conduct large-scale studies on citation counting weighted by in-text citation frequency, location or context. As a result, interest in this type of studies is growing.

Re-citation analysis as defined here may be viewed as a large sub-class of the class of in-text frequency weighted citation analysis schemes, a class which has recently been found to be the most effective one among many features of in-text citations at characterizing essential citations (Zhu, Turney, Lemire, & Vellino, 2014). We discuss in this paper why we consider re-citation analysis a promising method for improving citation analysis for research evaluation, knowledge network analysis, knowledge representation and information retrieval.

Weighted Citation Counting

Citation analysis examines citation patterns and networks in the scholarly literature through statistical analysis and network visualization. It is applied widely in the social sciences to trace knowledge flows, to evaluate research impact, to study the characteristics of scholarly communities and knowledge networks, and to create citation link based knowledge representation and retrieval systems (Borgman & Furner, 2002; Hall, Jaffe, & Trajtenberg, 2005).

The basic assumption underlying citation analysis is that a citation represents the citing author's use of the cited work, and that it therefore indicates that the citing and cited works are related in subject matter or methodological approach (Garfield, 1979; White, 1990). The total number of citations that a document or any aggregate of documents (e.g., author oeuvre, journal) receives (or a score derived from it, e.g., h-index) is therefore used to assess its impact on research in research evaluation. Citation links are used to signify knowledge flow from the cited to the citing group and, along with scores derived from these links, to measure the relatedness between documents or their aggregates in the study of knowledge networks and in the representation and retrieval of related documents.

The assumptions of citation analysis are believed to be in line with Merton's normative view of science (Garfield, 1979; Merton, 1942; White, 1990). Like other activities of science, citation behaviour is assumed to be governed by a set of norms which require authors to cite documents that have influenced them in developing their current works in order to give credit where credit is due (Edge, 1979; Griffith, 1990; Peritz, 1992; Tranöy, 1980). Although citations for reasons other than giving due credit do exist (Cronin, 1984; Edge, 1979), citation analysis has generally been found to produce valid results because it is based on a statistical analysis of the collective perceptions of large numbers of citing authors, most of whom do adhere to the norms most of the time (Small, 1977; White, 1990). This is especially true with citation network analysis and citation link based knowledge representation and retrieval, as even non-normative citations will not refer to unrelated works.

Researchers do cite for various reasons and citations do serve many different functions in citing papers, however (Brooks, 1985, 1986; Case & Higgins, 2000; Chubin & Moitra, 1975; Liu, 1993; Moravcsik & Murugesan, 1975; Shadish, Tolliver, Gray & Sengupta, 1995; Vinkler, 1987). Small (1982), for example, identified five typical distinctions in citation classification schemes: (1) negative or refuted, (2) perfunctory or noted only, (3) compared or reviewed, (4) used or applied, and (5) substantiated or supported by the citing work.

The importance of weighing citations by their role in the text has therefore long been recognized (Herlach, 1978; Narin, 1976). In recent years, with increasingly available digital full-text documents and advances in technologies for text processing, interest in studying weighted citations has finally picked up. Studies have experimented with weighing citations by the frequency with which they are referred to in the text (e.g., Ding, Liu, Guo, & Cronin, 2013; Hou, Li, & Niu, 2011; Zhu, Turney, Lemire, & Vellino, 2014), by the citation impact of citing papers (Ding & Cronin, 2011), or by the location and context in which they are cited (Boyack, Small, & Klavans, 2013; Jeong, Song, & Ding, 2014). It has been found that frequency-weighted citation ranking can outperform traditional citation ranking of top authors, and that in-text citation frequency was the best of many other full-text features to help spot citations that were considered crucial to the citing papers by their authors, at least in a hard science field studied (Zhu, Turney, Lemire, & Vellino, 2014).

Depending on what functions they serve in a given citing paper, citations likely appear more or less frequently there: perfunctory ones once only, negative or contrastive ones a couple of times, and used or substantiated ones many times. By weighing citations by their frequency of appearance in a scholarly paper, it is hoped that essential citations could be assigned greater weight than perfunctory ones so that citation analysis can focus on the more profound

influences and on organic relationships. If so, this could improve traditional citation analysis significantly as a high incidence of perfunctory citations has been observed (Small, 1982). For example, Teufel, Siddharthan, & Tidhar (2006) found that only a fifth of the references are essential for the citing papers, and Moravcsik & Murugesan (1975) noted that 40% references were perfunctory, frequently simply copied from other papers without ever having been read (Dubin, 2004).

Re-citation analysis: motivation and innovation

Perfunctory citations can thus be considered a serious source of noise if the signal that one wants to detect is the direct and substantial flow of knowledge in the literature. There are two obvious types of approaches to dealing with this problem: (1) to amplify the signal or (2) to filter out the noise. The ultimately best approach is likely some combination of the two. All frequency-based weighing schemes studied so far used the former approach by assigning a weight based on the in-text citation frequency such as assigning a weight of N or N^2 to a citation that appears N times in a citing paper.

By contrast, re-citation analysis, a concept we introduced recently (Zhao & Strotmann, 2015), uses the latter approach: it attempts to filter out perfunctory citations from the analysis by removing uni-citations (i.e., documents referenced only once in the text of a work) in order to analyze only re-citations (i.e., references that appear more than once in the text of a citing paper). The degree to which a cited work is used or has impacted research can be further differentiated by assigning weights to different re-citation frequencies. Re-citation analysis can thus combine the noise filtering and signal amplification approaches, offering the potential to find an optimal weighing scheme for in-text citation frequency.

Thus, the fundamental difference between re-citation analysis and all other frequency-based weighing schemes and hence the innovation of re-citation analysis is that the former attempts to make the fundamental qualitative distinction between those citations that represent real use by, or core impact on, the citing paper (which it tends to retain for analysis) and those that are merely mentioned in passing as related work that the author is aware of but did not directly rely on (which it tends to remove). The basic assumption of re-citation analysis is that papers are very likely to be cited again and again in a publication that relies heavily on them, while perfunctory citations should appear once only in a citing paper almost by definition.

Re-citation analysis can also avoid potential technical problems associated with simply amplifying multi-citations. Since the noise created by perfunctory citations is very strong (40% or more), the signal amplification required to counter it tends to be so strong that it can cause serious distortions. For example, Zhao & Strotmann (2015) found that a simple weight of N does not suffice to make non-perfunctory citations stand out. N^2 is the minimal power of N that fulfills this requirement, but tends to be seriously affected by ultra-meticulous in-text citing styles of a few authors as it overweighs high in-text frequencies. Weighing re-citations avoids this problem.

Promises of Re-citation Analysis

Re-citation analysis can be expected to contribute significantly to the theory and methods of citation analysis. It addresses head-on an old and fundamental concern with citation analysis, especially with evaluative citation analysis. By proposing to filter out the strong noise caused by a high incidence of perfunctory citations rather than simply amplifying multi-citations, it also opens up a new way of thinking about weighing citations at a time when the study of weighted citation counting based on full-text analysis is still in its infancy.

Re-citation analysis is promising in improving citation analysis for research evaluation, knowledge network analysis, knowledge representation and information retrieval.

- Evaluative citation analysis ranks authors, journals, institutions or other components of the scholarly communication system by their citation counts or by derivative scores such as the h-index. Scores based on re-citation counting can be expected to boost those researchers or groupings whose publications receive close scrutiny and to introduce a bias against those whose work mainly provides convenient background information. Such re-citation metrics should thus be better at measuring research impact than traditional citation metrics.
- In citation-based knowledge network analysis and visualization, results based on re-citations can be expected to be significantly more detailed and "crisp" than those based on citations since re-citation based relations (e.g., direct re-citation, co-recitation, or re-citation coupling) should represent core relationships where citation-based relations include many peripheral ones. The price might be an underestimation of interrelatedness between distant parts of a science map.
- For information retrieval (IR), re-citation based similarity metrics can likely provide a considerably enhanced precision of the "Similar documents" or "More like this" feature that many IR systems provide nowadays, compared to citation-based ones. The latter can be expected to show better recall, however, so that a (weighted) combination of the two may work better than either one alone.
- For knowledge representation, it is well understood that citations in scholarly publications serve as concept symbols (Small, 1978). One would expect the presence of a certain set of citations in a paper to translate fairly straightforwardly to the assignment of that paper to a specific subject category. However, subject categories are meant to capture the paper's "aboutness", but a large percentage of citations merely provide "relatedness" information. We suspect that re-citations, on the other hand, do correspond to a considerable degree to concept symbols with an "aboutness" semantics. A re-citation based form of computer-aided subject indexing might therefore be feasible.

Re-citation analysis may thus have a profound impact on the future of the scholarly communication system and of Scientometrics as re-citation analysis values and thus encourages research that is worth following in depth, whereas traditional citation analysis has encouraged review publications that tend to be cited widely.

Finally, as they rely on access to the full text of scholarly publications rather than on citation databases such as Web of Science and Scopus, re-citation analysis methods and metrics are as easily available to the study and evaluation of the social sciences and humanities as to that of the natural and life sciences. Unlike the latter, the former have never been treated fairly by traditional citation analysis due to the insufficient coverage of their literature by these databases.

References

Borgman, C.L. & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 3-72.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co‐citation clustering using full text. *Journal of the American Society for Information Science and Technology, 64*(9), 1759-1767.

Brooks, T. A. (1985). Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223-229.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34-36.

Case, D. O. & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Cronin, B. (1984). *The Citation Process. The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.

Chubin, D. E. & Moitra, S. D. (1975). Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423-441.

Ding, Y. & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management, 47*, 80–96.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Dubin, D. (2004). The Most Influential Paper Gerard Salton Never Wrote. *Library trends, 52*(4), 748-764.

Edge, D. (1979). Quantitative measures of communication in science: A critical review. *History of Science Cambridge, 17*(36), 102-134.

Garfield, E. (1979). *Citation indexing – Its Theory and Application in Science, Technology, and Humanities.* New York: John Wiley & Sons.

Griffith, B. C. (1990). Understanding science: Studies of communication and information. In C. L. Borgman (ed.). *Scholarly Communication and Bibliometrics*, 33-45. Newbury Park, CA: Sage Publications, Inc.

Hall, B.H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics, 36* (1), 16–38.

Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science, 29*(6), 308-310.

Hou, W., Li, M., & Niu, D. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays, 33*, 724-727.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics, 8*(1), 197-211.

Liu, M. (1993). The complexities of citation practice: A review of citation studies. *Journal of Documentation, 49*, 370–408.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology, 1*, 115-126.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86-92.

Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Washington, D. C.: Computer Horizons.

Peritz, B. C. (1992). On the objectives of citation analysis: Problems of theory and method. *Journal of the American Society for Information Science, 43*(6), 448-451.

Shadish, W. R., Tolliver, D., Gray, M., & Gupta, S. K. S. (1995). Author judgements about works they cite: three studies from psychology journals. *Social Studies of Science, 25*(3), 477-498.

Small, H. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Social Studies of Science, 7*(2), 139-166.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327-340.

Small, H. (1982). Citation context analysis. *In* B. J. Dervin & M. J. Voigt (eds.)*, Progress in Communication Sciences, 3* (pp. 287-310). Norwood, NJ: Ablex.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function*. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 103-110)*. Stroudsburg, PA, USA.

Tranöy, K. E. (1980). Norms of inquiry: Rationality, consistency requirements and normative conflict. In *Rationality in Science* (pp. 191-202). Springer Netherlands.

Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics, 12*(1), 47-72.

White, H. D. (1990). Author co-citation analysis: Overview and defense. *In* C. L. Borgman (ed.), *Scholarly Communication and Bibliometrics* (pp. 84-106). Newbury Park, CA: Sage.

Zhao, D. & Strotmann, A. (2015). Dimensions and uncertainties of author citation rankings: Lessons learned from frequency-weighted in-text citation counting. *Journal of the Association for Information Science and Technology*, doi: 10.1002/asi.23418.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2014). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*. Early view (DOI: 10.1002/asi.23179).