How to build your own citation index: First hand experiences from Web of Science, Scopus and CSA reference data.

William Dinkel, Frank Sawitzki, Andreas Strotmann

GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

Links between citing and cited documents are essential for bibliometric analysis: Evaluative bibliometrics uses the citation frequencies of documents as a proxy for their impact, descriptive bibliometrics as representations of cognitive linkage between documents, co-citation and bibliographic coupling data help in describing knowledge flows based on citation links.

However, limitations and biases of the technical procedures applied in generating and standardizing reference data as well as in the subsequent matching of references to target documents have received limited attention in the recent development of bibliometric methods. While there are some new approaches on a conceptual level, little "hands-on" experience is available yet. By outlining and comparing our experience (Dinkel 2011; Strotmann et al. 2010) in cleaning, standardizing, and matching references from Web of Science (WOS), Scopus and Cambridge Sociological Abstracts (CSA) we hope to provide insight into the limitations and potentials of these data sources.

Views of References in Citation Analysis

In traditional citation analysis, the citation graph has usually been modeled as a bipartite network consisting of citing papers as one partition, and references as the second partition, linked whenever the citing paper represented by a node contains the corresponding reference. This model corresponds to the realities faced for a long time by an external user of the (Social) Science(s) Citation Index of Thomson Reuters, where the information provided on citing papers and on references contained in them were vastly different and searching (linking) the citation index for the full record that corresponded to a reference was all but impossible.

From a naïve non-bibliometrician's perspective, this citation network model is quite unintuitive. In a citation index, users nowadays expect there to be an actionable link from the text of a reference contained in a citing paper to some kind of web page providing access to some version of the corresponding full paper, or at least its full metadata. In other words, the intuitive citation network model is that of a uni-mode network of document nodes connected via directed citation links from the citing document's node to the node that corresponds to a document that is referenced in it.

Citation Matching for Citation Network Analysis

The former traditional model lends itself to direct analysis of data provided by a citation index, be it WoS, Scopus, or CSA. In the case of author citation network analysis, for example, this means that cited authors' names are available only in last-name-plus-initial form, greatly hampering cited author name disambiguation where necessary.

By matching a reference to the (most likely) document that it refers to, and by completing the cited nodes in the citation graph with full (meta-)data relevant to an analysis, considerably more interesting analyses on the cited nodes of the network become possible. (Strotman et al., 2010) describes how this enables the development of improved author cocitation analysis techniques that take into account all authors of papers, which is particularly important in highly collaborative research areas where the last author of a paper may be just as important as its first author.

Citation Matching in Scopus

A typical reference contained in Scopus contains the names of up to eight authors, the publication year, the full title of the paper, and the journal name, volume, and number where it appeared, with the first page number of the paper in that number. Author

names are standardized to last-name-plus-initials, but little else is standardized. Nevertheless, (Strotmann et al., 2010) found that it was possible to parse Scopus references to extract these pieces of information, and to construct reasonably successful and precise search strategies within Scopus to locate the corresponding full record for a reference (in the case of a biomedical field, more than 90% success and accuracy rates).

Citation Matching in WoS

While the parsing of references from WoS records is much more straightforward than for those from Scopus, WoS search facilities for a long time did not provide filtering capabilities for some of the most distinctive fields they contain (e.g., volume, number, and starting page), while lacking one of the most telling fields, namely, the title of the referenced document. Searches within WoS for the document that corresponds to a reference have thus long been likely to yield large numbers of hits, from which the "correct" one needed to be filtered out, making the matching process quite inefficient.

Citation Matching in CSA

As the provider of CSA citation databases for the German National Academic Licensing programme for academic databases, we at GESIS have recently begun looking to include links between records of citing and cited documents within our Sowiport integrated social science literature database at gesis.org/sowiport, both for user browsing support and for bibliometric analysis support. The information contained in CSA references is similar to that in Scopus, with subfield parsing even simpler than in WoS, at least for journal references. Fairly straightforward matching routines allowed a precise (>95%) match for about a third of all references found in the CSA databases we carry (almost 10 million).

In the social sciences, however, non-journal publications play a major role, and the subfield parsing for the corresponding references did not work well. We are therefore working on improving the subfield extraction for these types of references, as well as on improving fuzzy matching techniques for identifying referenced works from badly OCRed or otherwise broken references. Especially for cited non-journal literature, these techniques are required not just for CSA, but also for Scopus and WoS, we have found.

Conclusion

In our contribution we will report on our experience with limitations and potentials of deterministic and probabilistic algorithmic approaches at different stages of reference matching for different citation indexes, as well as with different data storage and handling techniques. By making available our experience with building a custom citation index we aim to foster an understanding of technical and conceptual limitations of "off-the-shelf" data generally available from WoS or Scopus. We also provide some starting points for extending citation analysis to data sources beyond WoS and Scopus.

References

Dinkel, William (2011). How do matchkeys affect citation counts? First steps towards an error calculus for bibliometric indicators. In: Noyons, Ed / Ngulube, Patrick / Leta, Jacqueline (eds.), 2011: Proceedings of the ISSI 2011 Conference, 13th International Conference of the International Society of Scientometrics and Informetrics, Durban, South Africa, 04.-07.07.2011, Volume I. Leiden: ISSI, 175-180.

Strotmann, Andreas; Zhao, Dangzhi (2010). Combining commercial and open access citation databases to delimit highly interdisciplinary research fields for citation analysis. Journal of Informetrics, 4 / 2, pp. 194-200.