OPINION

D-Lib Magazine
March/April 2009

Volume 15 Number 3/4

ISSN 1082-9873

What's Wrong with Citation Counts?

José H. Canós Cerdá, Ph.D.
Department of Computer Science (DSIC)
Technical University of Valencia, Spain
<jhcanos@dsic.upv.es>

Eduardo Mena Nieto, Ph.D.
Department of Computer Science and Systems Engineering
University of Zaragoza, Spain
<emena@unizar.es>

Manuel Llavador Campos
Department of Computer Science (DSIC)
Technical University of Valencia, Spain
<mllavador@dsic.upv.es>

(This Opinion piece presents the opinions of the author. It does not necessarily reflect the views of D-Lib Magazine, its publisher, the Corporation for National Research Initiatives, or the D-Lib Alliance.)

Abstract

Citation analysis needs an in-depth transformation. Current systems have been long criticized due to shortcomings such as lack of coverage of publications and low accuracy of the citation data. Surprisingly, incomplete or incorrect data are used to make important decisions about researchers' careers. We argue that a new approach based on the collection of citation data at the time the papers are created can overcome current limitations, and we propose a new framework in which the research community is the owner of a Global Citation Registry characterized by high quality citation data handled automatically. We envision a registry that will be accessible by all the interested parties and will be the source from which the different impact models can be applied.

Introduction and motivation

Research is an intrinsically social activity. We work with colleagues, sharing research projects and co-authoring papers; we agree about the utility of peer review for selecting papers for publication, cite others' works, and evaluate ourselves in terms of the impact of our work. In a community that includes supporters for both manual and automatic approaches, both with advantages and shortcomings, there seems to be an overall feeling that current citation-based impact assessment systems are used because they provide the "least bad" solution to the problem of evaluating scientific output, and that people prefer one or another depending on how highly they are ranked in a particular system. And, paradoxically, these measures, obtained from incomplete or incorrect citation counts, are used to make important decisions regarding researchers' careers such as positions and grants.

After years of being dominated by commercial systems, citation analysis needs an in-depth transformation. The proliferation of publications in recent years has made citation capture processes at the commercial level increasingly costly due to the need for human involvement in generating the records; this high cost, in addition to the strong business interests of companies, has led to incomplete coverage of journals and conferences in commercial databases. As an alternative to the commercial entities, several freely accessible systems have emerged that automatically extract citation information from online documents; this has led to a significant increase in publication coverage over that of fee-based systems. CiteSeer (http://citeseer.ist.psu.edu/)

was one of the pioneer open access systems, followed years later by Google Scholar (http://scholar.google.com/) and Microsoft's Live Search Academic. (The Microsoft Live Search Academic system had already been discontinued as of the time of writing.) Unfortunately, automatic citation extraction systems are prone to significant error rates that may seem unacceptable for some purposes. An extreme example was provided by (Jacsó, 2006): according to Google Scholar, an author named "I. INTRODUCTION" has published hundreds of papers. Similarly, according to CiteSeer, the first and third authors of this article have a coauthor named "I. Introducción" for one of their papers written in Spanish (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.98.5433). Nevertheless, these mistakes are inherent to the automatic data harvesting, and do not diminish the relevance, especially in terms of coverage, of such systems.

An increasing number of scientists have begun to acknowledge that the current models of citation analysis cannot be sustained (Adam, 2002), and in the past few years, some alternative impact models, such as the h-index (Hirsch, 2005), have been proposed. However, despite the fact that the new models might seem more realistic in terms of evaluation, they suffer from the same root problem: the rates of citation data accuracy and completeness are not precise enough to make fair assessments. In fact, regardless of the model used to analyze impact, results will be inaccurate if the data set is incomplete or incorrect. Moreover, every current approach has commercial interests behind it that may interfere with what is supposed to be a neutral evaluation process. We feel, therefore, that a new approach is needed to give control of citation data back to the research community, which – in a collaborative fashion – can generate richer data sets from which trusted evaluation methods can be defined and applied.

Late vs. early citation management

A look at the internal processes of both commercial and free citation management systems shows that citation harvesting, which uses costly techniques such as optical character recognition, machine learning, and others, start *after papers have been published*. Notice, however, that in a high number of papers, citation information is generated, and hence can be collected, much earlier. In fact, most scientists prepare their papers using word processing systems that have accompanying bibliography management utilities. BibTeX, for instance, is able to generate bibliography lists in LaTeX documents from metadata stored in the so-called ".bib" files. Microsoft Office Word 2007 has a built-in bibliography manager, and users of earlier versions can manage their bibliographies using third-party applications such as EndNote or RefWorks. All these bibliography managers are aware of the citations included in papers, but such information is systematically discarded when the camera ready copies of the papers are sent for publication. At that time, citation records must be built again from scratch, which results in additional costs, errors and delays. Moreover, different companies maintain different citation databases with highly overlapping content, possibly in different formats that complicate interoperability.

Better management of the citation data throughout the lifecycle of a paper will improve data quality and significantly reduce the cost of citation generation. Instead of viewing scientific publishing as a number of disconnected activities, we claim that a framework should be defined for a global workflow, from document creation to publication, involving different actors who would participate collaboratively. Citation data would be generated only once – at the time of document creation – after which such data could flow from one activity to the next. Consequently, there would no longer be a need to harvest citation data again after a paper's publication. The citation records thus generated should be stored in a Global Citation Registry (GCR), maintained by independent organizations similarly to the way in which Internet domain names or ISBN codes are managed. As envisioned, the GCR would be freely accessible for queries; and updates to it would be made by the entities responsible for the publications of papers, that is, companies or organizations acting as publishers.

Requirements for early citation management

To make this workflow possible, bibliography management systems must become citation-aware, that is, new functionality must be added to implement a citation meta-model. Inserting a citation in a paper comprises several steps:

1. The metadata of the cited paper is retrieved by the bibliography manager from a reference retrieval service such as DBLP (dblp.uni-trier.de);
2. The author defines the citation context that gives an assessment that is as precise as possible regarding the relevance of the citation for the work described in the paper. By context we mean metadata attached to a citation that can be used to define richer impact models that include some qualitative measures. An example of context was pointed out in Parnas (2007): it may be relevant to a researcher to know if a paper cites another one to criticize it, to list it as related work, or to acknowledge it as a main source for the work being described.
3. The bibliography manager should check to be sure that every paper referenced in the bibliography list is cited explicitly in the paper's body.
4. Finally, the bibliography manager should generate a citation record in the GCR format from the references included in the paper.

Discussion

We feel the effort required to implement this approach is worth it, given the advantages that doing so returns. Authors and researchers would benefit from being able to use the enhanced bibliography managers in different ways, including: handling citations automatically; specifying citation contexts as explained above; and being able to make citation-aware bibliographic queries like "find papers about global warming cited more than 50 times in the last three years".

After the paper has been submitted, reviewers can validate the context of the citations during the reviewing task, so that only pertinent citations count in favor of the cited papers. Then once editors accept a paper for publication, the publisher's content management systems would request the citation file for the accepted paper and submit it to the GCR. As a counterpart, publishers could exploit citation data to better index the papers they publish.

Finally, there would be no need for post-publication citation harvesting, as citations would be included in the validated files. Every time a new paper is published, the publisher would only need to associate an identifier (e.g., a Digital Object Identifier (DOI), <http://www.doi.org>) with the paper and submit the citation file to the GCR.

Table 1 A look at the different paper lifecycle tasks before and after the proposed framework

Task	Edit	Submit papers	Order review	Review	Order publication	Publish
Actor	Author	Author	Editor	Reviewer	Editor	Publisher
Before	Use of bibliography manager (optional) to build the bibliography	Only the paper is submitted	Paper is sent to reviewers	Reviews paper using his/her own efforts	Only the camera-ready paper is sent to the publisher	Put the paper in the digital library (optionally, generate citation information by hand)
	Problems: Citations as text require post-publication citation harvesting	Problems: Citations as text require post-publication citation harvesting	Problems: Reviewer cannot manually check many situations	Problems: Manual verification of citations and related work	Problems: Publisher cannot index papers as desired	Problems: Citation information must be extracted from the paper's full text
After	Citation-aware bibliography manager (mandatory) produces citations as objects	The paper + the citation file are submitted	The paper + the citation file are sent to reviewers	Citation data is used to validate references in the paper	The camera-ready paper + the citation data are sent to reviewers	The new paper is registered globally. High quality citation data are used to index papers
	Benefits: It makes post-publication citation harvesting unnecessary	Benefits: It improves the quality of biblio data and eases the tasks of reviewers, publishers, and authors!	Benefits: • Quality of reviews is improved. • Accepted papers have correct bibliography	Benefits: • Citation context can be validated easily • Automatic detection of inconsistencies	Benefits: Publishers can exploit the citation data to index papers	Benefits: • No need for citation harvesting • High quality cross references

Our proposal does not leave current stakeholders out of the business. Separating citation data management from processing will allow companies to apply their impact evaluation models from the data stored at the GCR instead of applying them to local and possibly incomplete citation databases. This will also help companies to eliminate the costs of harvesting citation data – as all citations will already reside in the GCR – and the companies can then focus on improving their analysis models, whose trustworthiness will be increased because evaluation will be applied to higher quality citation data. Moreover, data will be the same for all the systems, which can launch an interesting competition between impact models in terms of what researchers expect from them: realistic impact evaluations.

Conclusion

Citation data is a very valuable asset that reflects the social side of research. It is generated by the scientific community as a whole and should be considered part of the community's heritage. The existence of a global, community-maintained citation registry, generated via the early collection of citation data, is the first step towards reaching the goal of trustworthy evaluation mechanisms that will replace or improve the currently incomplete ones. Twelve years ago Robert Cameron proposed the creation of a Universal Citation Database as the solution to some of the problems of current systems (Cameron,

1997). The technology available at that time made a practical implementation of his proposal difficult. However, currently available technologies allow the implementation of a global citation management workflow over a service-oriented architecture. The Framework for Early Citation Management we presented at the European Conference on Digital Libraries 2008 (Canós et al., 2008) includes the GCR, as well as other related data sources such as an Author Registry – to enable name disambiguation – and a Global Publication Registry. All these sources are accessible to feeding and querying client applications via corresponding public application interfaces, facilitating the creation of numerous applications that exploit the publicly available citation data for use by various communities involved in the collective task of communicating research results.

Acknowledgements

The work of J. H. Canós and M. Llavador is partially funded by the Spanish *Ministerio de Educación y Ciencia* (MEC) under project META (TIN2006-15175-C05-01), the *Junta de Comunidades de Castilla-La Mancha* under project INGENIO (PAC08-0154-9262) and the *Generalitat Valenciana* under grant ACOMP07/216. M. Llavador is the holder of the MEC-FPU grant no. AP2005-3356. The work of E. Mena is supported by the MEC under project TIN2007-68091-C02-02.

References

Jacsó, P., 2006. "Deflated, inflated and phantom citation counts". *Online Information Review*, Vol.30 No.3, pp.297-309. <10.1108/14684520610675816>.

Adam, D., 2002. "The counting house". *Nature*, Vol. 415 (February 2002), pp. 726-729. <doi:10.1038/415726a>.

Hirsch, Jorge E., 2005. "An index to quantify an individual's scientific research output," *PNAS* 102(46):16569-16572 (November 15 2005). <doi:10.1073/pnas.0507655102>.

Parnas, D. L., 2007. "Stop the numbers game". *Commun. ACM* 50, 11 (Nov. 2007), 19-21. <doi:10.1145/1297797.1297815>.

Cameron, R., 1997. "A Universal Citation Database as a Catalyst for Reform in Scholarly Communication". *First Monday* 2(4), at <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/522/443>. Accessed on October 20, 2008.

Canós, J.H., Llavador, M., Mena, E., and Borges, M., 2008. "A Service-Oriented Infrastructure for Early Citation Management". *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*. LNCS 5173, Springer, 2008. <10.1007/978-3-540-87599-4>.

Top | Contents
Search | Author Index | Title Index | Back Issues
Editorial | Next article
Home | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

doi:10.1045/march2009-canos