Developing Bibliometric Indicators of Research Performance in Computer Science

Martijn S. Visser and Henk F. Moed

{visser,moed}@cwts.leidenuniv.nl
Leiden University, Centre for Science and Technology Studies (CWTS),
P.O. Box 9555, 2300 RB Leiden (The Netherlands)

Abstract
This paper presents the first preliminary findings of a study that aims to develop bibliometric indicators for the measurement of research performance in the field of Computer Science. A database of the complete publication output during 1996 – 2001 of 52 research groups at universities in the Netherlands was created. ISI coverage indicators suggest that ISI journals only capture a small share of the impact of Computer Science research. In a first attempt to provide a more complete picture of the research impact of Dutch computer scientists, both the impact of their ISI-publications as well as the impact of their non-ISI publications such as proceedings papers and book chapters in ISI journals was determined. In the next phase of this study, not only these *targets* for citation will be included in our study but also the universe of citing *sources* will be expanded. The study aims to develop bibliometric indicators on the basis of insight into the system of written communication in the various fields of Computer Science. The indicators will be further validated by comparing them with peer judgements obtained from a national research assessment exercise.

Background and purpose
In the past decades bibliometric indicators have proven to be useful tools in the assessment of research performance in many scientific disciplines. In most studies those indicators were based predominantly on data derived from the Science Citation Index (SCI) and related Citation Indexes published by the Institute for Scientific Information (ISI). The ISI Citation Indexes processes articles published in approximately 9,000 leading international journals from all scientific disciplines. A usual method applied in such studies collects articles that were published by the unit to be assessed in journals processed for the ISI Citation Indexes (in so called ISI source journals), and analyses their citation impact, by counting the number of times they are cited in other papers published in ISI source journals (Van Raan, 1996). Such a methodology solely based on ISI source publications is denoted as a 'standard ISI impact assessment'.

The basic conditions for such a methodology to adequately measure research productivity and impact are the following:

- The research article is the principal output carrier of scientific/scholarly activity in the field under inquiry.
- Scientific journals constitute the principle media of (written) communication in the field.
- The ISI Citation Indexes cover the journal communication system sufficiently well.
- Practitioners in the field perceive some kind of hierarchy of journals on the basis of their importance, roughly categorizing these in terms of 'top or very good' journals, 'good', and 'less good' journals.
- The field under inquiry has a strong international orientation, and communication is seriously not hampered by national or language barriers.

During the past years, evidence was obtained suggesting that several of the assumptions listed above may not be valid - or only partly valid - in the field of Computer Science. Such evidence came from statements by experts in the field, and from empirical results of bibliometric research. Some experts qualify Computer Science as a relatively new, rapidly developing field with a strong multi-disciplinary orientation, embracing both basic and more applied research activities. It was claimed that, particularly in the youngest subfields, proceedings volumes rather than scientific journals constitute the main channel of written communication; that the ISI citation indexes do not adequately cover the journal literature in all subfields; that there may be no clear hierarchy of journals based on their perceived

Proceedings of ISSI 2005, pages 275–279.
Edited by P. Ingwersen and B. Larsen.

importance; and that in some subfields the scientific article may not even be the most important carrier of scientific output.

Several studies have demonstrated the frequent use of conference papers in the field of Computer Science (e.g., Goodrum et al., 2001). However, the importance of conference papers was not only claimed based on their frequent use but also based on the particular role conference papers play in the communication process. As many conference papers do not end up in journals, Drott (1995) put forward that a conference paper may very well be a final product in itself. Based on discussions with computer scientists Goodrum et al. (2001) concluded that in Computer Science a research article in a proceedings volume may serve the same purpose for the author as a journal article. Therefore, researchers may not seek to get their conference papers published in journals.

Bibliometric research has been conducted to operationalize the concept of adequacy of coverage of the ISI Citation Indexes (Visser and Moed, 2004). It was proposed that the share of cited references contained in papers from ISI journals that are themselves published in ISI journals, provides an 'internal' indicator of this adequacy of coverage. If the reference lists of papers that belong to a particular scientific field contain relatively few citations to ISI publications, it may indicate that one or more of the conditions described above are not fulfilled. Empirical findings show that this is actually the case in the broad field of Engineering that is believed to also cover most research activities of Computer Science. As Garfield (1974) suggested, this may be typical for the applied nature of the research activities in the field.

In view of the expert views described above and the available empirical evidence a measurement based solely on papers published in ISI journals may provide an unreliable and unrepresentative picture of the production and citation impact of research groups active in Computer Science. This paper reports on a study to obtain valid indicators for research performance in the field of Computer Science. The object of study is formed by the research activities conducted by researchers in the field of Computer Science at universities in the Netherlands. The study is methodological: it aims at the development - i.e. construction and testing - of bibliometric indicators of research performance in the various subfields of Computer Science, based upon insight into the system of written communication in those subfields. The study focuses on the development of impact indicators, measured by the number of times articles published by a research group are cited in other documents published in sufficiently important communication media.

Methods

Table 1 presents a tentative classification of four types of bibliometric studies in function of the adequacy of ISI coverage of the field of inquiry (see Moed, 2005 for an elaborate discussion of this scheme). Essentially, this classification represents the central hypothesis that determines the approach adopted in this study. Our goal is to decide which type of study is most appropriate for the field of Computer Science and to objectify the criteria on which this decision is based. In order to do so we need to further operationalize the concept of adequacy of ISI coverage as in Table 1, the qualifications of ISI coverage are still purely qualitative and need to be further quantified.

Table 1. Four types of citation impact studies

ISI coverage	Type of Study	Target / Cited	Source / Citing
Excellent	Standard ISI	ISI	ISI
Good	Target expanded	ISI+non ISI	ISI
Moderate	Source expanded	ISI+non ISI	ISI+non ISI
Poor	No citation analysis at all		

To further explain this classification, some technical terms should be specified. In a citation impact analysis one should distinguish a citing or source side and a cited or target side. Target articles are those that are subjected to a citation analysis. Source articles are documents from which cited references are extracted.

In fields with an excellent ISI coverage, it is generally sufficient in a citation impact analysis to take into account as target articles only those that are published in ISI source journals, and to use the total collection of cited references in ISI source journals as citation universe. This type of analysis is labelled above as the standard ISI analysis.

If ISI coverage in a field is not excellent, but can nevertheless be qualified as good, the scheme suggests expanding the collection of target articles analysed in the standard ISI analysis by including target articles that are not published in ISI source journals (*Target Expanded* in Table 1). In this approach it is assumed that, although the collection of source (citing) documents is incomplete, their cited references still may provide reliable citation impact estimates, to the extent that the ISI source articles constitute a representative sample of a wider population of citing sources.

For instance, if in a field a limited number of important journals or even proceedings volumes is not covered by ISI, whereas ISI coverage can still be qualified as good, it can be expected that papers published in those not-covered media are on average sufficiently frequently cited from ISI source journals in order to assess their impact.

If ISI coverage of a field is moderate, however, it becomes questionable whether such an assumption of representativeness of cited references in ISI source journals is still valid. For instance, if proceedings of annual international conferences play a dominant role in the communication system in a field, it can be expected that there is heavy citation traffic among proceedings from subsequent years. If such proceedings are not covered by the ISI Citation Indexes, these citation relationships remain invisible if citations are merely extracted from ISI source journals. In that case, it seems appropriate to expand the universe of citing sources with articles in proceedings volumes from a range of subsequent years. Such an approach is labelled as *Source Expanded* in Table 1.

Finally, if ISI coverage in a field is poor, it is questionable whether it is useful conducting a citation analysis based on ISI data, even if target or source universes are expanded. This is particularly true in fields that are fragmented into schools of thought and by national or linguistic barriers. It is to be expected that in such fields alternative approaches, not based on citation data, are more fruitful than citation impact analyses.

Data

A bibliometric database of the complete publication output of 52 research groups at Dutch universities during 1996 – 2001 has been created containing bibliographic details on almost seven thousand publications including information about all co-authors for each publication. The database was constructed in such a way that it enabled us to conduct statistical analyses, and link it - even on an article-by-article basis - with other, external literature databases. In addition to the bibliographic details it also provides information on the way each publication was classified (e.g. book, chapter in edited book, thesis, journal article, report etc.) and whether the publication was peer reviewed or not.

First Results

Pure ISI analysis

The publication output of Dutch academic research groups in Computer Science was linked to all source publications processed for the ISI Citation Indexes. In this way a subset of 1,219 papers was defined that encompass all publications in ISI journals. As a first step, to gain insight in the adequacy of the ISI coverage for Dutch Computer Science research publications, we analysed the references of the ISI-papers included in this subset. The cited references of these papers published in 1996 - 2001 were matched against all ISI publications processed since 1980. Self-citations were included, as we could not exclude all self-citations for non-ISI documents. The combined bibliographies comprised on average almost 25 items. Of these 25 cited references, 91% were dated between 1980 and 2001 and 30% from those could be matched to publications in the ISI Citation Indexes. However, this figure probably overstates the importance of ISI-publications for researchers in Computer Science somewhat, as the research groups that cite ISI publications relatively often also tend to publish more frequently in ISI-journals.

Table 2. Distribution of internal and external ISI coverage rates among research groups

ISI Coverage	Min	Q1	Median	Q3	Max
% Cited References	11%	20%	25%	32%	57%
% Publications	0%	14%	20%	29%	46%

Legend: Min, Max: the minimum and maximum score in the distribution;
Q1, Q3: the first and third quartile of the distribution

Table 2 (Column Median) shows that for half the research groups 25% or less of their references cite other ISI publications. These percentages of 'internal' ISI coverage are very low when compared to scientific disciplines such as Physics and Chemistry where one usually can observe percentages around 90%. Coverage rates are even lower when the ISI coverage rate is expressed as the share of ISI publications in the total collection of all publications made by the groups involved. However, this observation may well be attributed to the fact that all publications are assumed to be equally important whereas the ISI Citation Indexes are selective in their coverage and include only the most influential journals.

Nevertheless, it is obvious from these results that a pure ISI analysis captures only a small share of the impact that other research has on Dutch Computer Science. For that reason it seemed appropriate to expand our citation analysis to publication categories other than ISI journal articles.

Target expanded analysis

In the next step we matched all publications included in our Dutch Computer Science database against the cited references of source publications processed for the ISI Citation Indexes during the years 1996 - 2003. If a linkage could be established, the author lists of the citing and cited publication were compared with one another. Linked publications with one or more authors in common were defined as self citations and were excluded from the impact analysis.

Table 3. Distribution of output and ISI-impact over publication categories
(publication years: 1996 – 2001, citation years: 1996 – 2003)

Publication Category	**Total Pubs**	**%**	**%Cited**	**Total Cites**	**CPP**	**Max Cites**
ISI journal articles	1,214	20	60%	3,712	3.06	125
Non-ISI refereed journal articles	436	7	37%	828	1.90	213
Refereed Conference papers	3,375	57	26%	2,836	0.84	323
Theses	400	7	42%	495	1.24	32
Books	72	1	60%	1,435	19.93	346
Chapters in edited book	456	8	25%	405	0.89	24
Other (non refereed)*	627	-	5%	70	0.11	14

Legend: CPP: Citations per publication
Max Cites: The number of citations to the most frequently cited publication of a particular type
*The figures on "Other (non refereed)" are given for reference purposes only as not every university provided data for this category.

Table 3 presents a breakdown of the total output into publication types along with impact indicators for the different categories. Conference proceedings are by far the most frequently used publication outlet. The expansion of the analysis with the additional targets more than doubled the total citations received by Dutch researchers in Computer Science. However, the majority of publications that were added to our target universe remains invisible in ISI journals. Only around a quarter of the conference papers and book chapters were cited in journals processed by ISI. The level of visibility for non-ISI journal articles, theses and especially books is higher but these are less frequently used types of media.

Figure 1 depicts for each of the research groups involved the average impact of their ISI publications in comparison to the average impact of their non-ISI publications. No strong correlation can be observed and the picture even suggests a division between groups with rather different impact characteristics in relation to the importance of non-ISI publications versus ISI publications.

In view of the observations above in combination with the low coverage rates of references in ISI journal papers it remains uncertain whether the outcomes of a target expanded citation analysis are reliable and representative for the total citation impact of the groups involved.

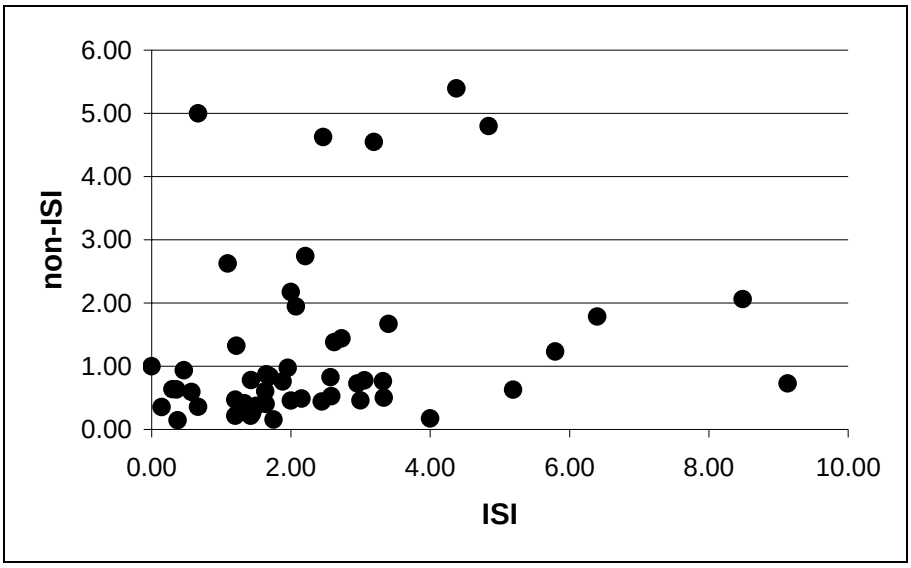

Figure 1: Impact of ISI publications versus that of non-ISI publications for 52 research groups in Dutch Computer Science

Remaining Work

Source expanded analysis

To gain more insight into the total citation impact of Dutch research groups in Computer Science we will expand our analysis with reference lists taken from publications not included in the ISI citation indexes. We will not aim at achieving completeness of coverage. The crucial point is that important sources are selected, and a main challenge is to operationalize importance of sources in an objective way. From the results obtained in the previous stages of the study a list will be compiled of important sources of publications not included in the ISI citation indexes. The importance will be assessed on the basis of the following types of information:

a) The frequency at which Dutch researchers have published papers in those sources.
b) The frequency at which sources are cited in ISI-covered journals.
c) Rankings of sources based on citation analyses carried out by others.
d) Views expressed by experts in the various subfields.

Testing and validating of bibliometric indicators.

The results from this last phase of our study will be used to test and validate the outcomes of the pure ISI analysis and target expanded analysis. The bibliometric findings will also be compared with the outcomes of the peer evaluations of Dutch research activities in the field of Computer Science conducted in 2003/early 2004. In addition, bibliometric outcomes will be discussed with prominent researchers from the subfields involved.

References

Drott, M.C. (1995) Reexamining the Role of Conference Papers in Scholarly Communication. *Journal of the American Society of Information Science & Technology* 46, 299-305.

Garfield, E. (1974) Some comments on pure and applied research – stimulated by a list of works cited by applied chemists. *Current Contents*, December 4, 1974. In: *Essays of an information Scientist*, 2, 184–197. Philadelphia: ISI Press.

Goodrum, A.A., McCain, K.W., Lawrence, S. and Giles, L. (2002) Scholarly publishing in the Internet age: a citation analysis of computer science literature *Information Processing and Management,* 37, 661-675.

Moed, H.F. (2005), *Citation Analysis in Research Evaluation*. Springer, forthcoming.

Van Raan, A.F.J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36, 397–420.

Visser, M.S. and Moed, H.F. (2004), Measuring the impact of non-ISI source items, Presentation at the 8th Science and Technology Indicators Conference 2004. Slides available at: http://conference.cwts.nl/Downloads/ppt/128_Visser.pdf.