Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 71, No. 1 (2007) 145–150
DOI: 10.1007/s11192-007-1656-5

Short communication

Using scripts to streamline citation analysis on STN International

CHRISTOPH NEUHAUS,[a] ANDREAS LITSCHER,[b] HANS-DIETER DANIEL[a,c]

[a] *Professorship for Social Psychology and Research on Higher Education, ETH Zurich, Zurich (Switzerland)*
[b] *InfoLit Information Broker, Bern (Switzerland)*
[c] *Evaluation Office, University of Zurich, Zurich (Switzerland)*

The database host STN International allows for extensive citation analysis in the SCISEARCH database (Science Citation Index Expanded) and in the CAplus database (Chemical Abstracts). Along with its powerful browsing, searching and analyzing facilities, STN International also features scripts. In this paper we examine the usefulness of the script language in the automation of citation analysis in SCISEARCH and CAplus.

Introduction

Citation analysis has established itself as a widespread method for the assessment of the research performance of universities, institutes and research groups. Citation counts are seen as quantitative measure of the resonance and impact that their publications have evoked among the scientific community. Most commonly, the main source for citation analysis are the citation indexes produced by Thomson Scientific (formerly Institute for Scientific Information). Cited reference searching is available through the Web-based interface Web of Science and through a variety of database hosts, such as DIALOG, DIMDI and STN International. In 2000, the Chemical Abstracts Service (CAS) introduced citation indexing to its bibliographic database. Cited references are

Received April 4, 2006

Address for correspondence:
CHRISTOPH NEUHAUS
ETH Zurich, Professorship for Social Psychology and Research on Higher Education
Zähringerstrasse 24, CH-8092 Zurich, Switzerland
E-mail: neuhaus@gess.ethz.ch

0138–9130/US $ 20.00
Copyright © 2007 Akadémiai Kiadó, Budapest
All rights reserved

included for journal articles and conference proceedings from 1997 to the present. Patent examiner citations are included for basic patents from the United States Patent & Trademark Office (USPTO), the European Patent Office (EPO), the World Intellectual Property Organization (WIPO), and the German Patent Office from the beginning of 1997, and for British and French basic patents from the beginning of 2003. Data on cited references, however, is available exclusively through the client software SciFinder (Scholar) and on the database host STN International.

Citation analysis can be a laborious and time-consuming task, when assessing research performance of specific target groups such as departments, institutes or research groups. In contrast to bibliometric analysis on a national level, which identifies the relevant publications using the address information available in some bibliographic databases, the research performance of target groups is preferably assessed on the basis of publication lists compiled or verified by the researchers themselves (cf. VAN LEEUWEN et al., 2003). This so-called bottom-up approach ensures that publication data is complete and correct, thereby significantly contributing to the reliability of citation analysis. The short-term impact or long-term impact[*] is assessed on the basis of these publication lists. Basically, such analyses can be performed through the popular interface Web of Science. Determining citation counts during a specified time period, however, requires a step-by-step approach in Web of Science, which takes a lot of time and manual effort. For example, citing publications must be retrieved and analyzed by their publication year and resulting citation counts must be exported separately for further data analysis for each item on the publication list.

The process of data collection must be automated as far as possible to perform citation analysis for specific target groups in an effective manner, both in terms of time and costs. In this regard, we examine the usefulness of the STN International script language for automation of citation analysis in the SCISEARCH database (Science Citation Index Expanded) and in the CAplus database (Chemical Abstracts). We introduce ready-to-use scripts, which automatically determine citation counts for a given set of publications, either for a fixed or a variable citation window. Possibilities as well as limitations of automatic data collection are pointed out.

Performing citation analysis on STN International

The retrieval language of STN International facilitates extensive citation analysis in SCISEARCH and CAplus (MARX et al., 2001; RIDLEY, 2001). Additionally, the client software STN Express with Discover features scripts. The script language is a miniature programming language, including components such as variables, operators, statements

[*] Typically short-term impact relates to a time period ranging between 3 and 5 years, whereas long-term impact assesses the impact over a longer time period, e.g. 10 years (MOED, 1996).

and conditions. Furthermore, the script language provides facilities for reading and writing files. Therefore, scripts are well suited for the execution of repetitive search and display tasks. They enable the preparation of queries offline and the automation of data collection to a large extent.

Performing citation analysis for a specific target group is a repetitive task. Based on publication lists, citation counts are determined by looking up the reference field in SCISEARCH or CAplus, respectively, how often the given publications have been cited. Basically, this process requires only the name of the first author, the publication year, the volume number, along with the beginning page. A publication is well defined with this data (MARX et al., 2001). Unlike Web of Science, which enables searching only for the cited author, cited publication year and the cited work, STN International provides separate data fields for each element in cited references, both for SCISEARCH and CAplus. Searching for cited author and cited publication year as well as for cited volume and cited beginning page is available. Citing publications can thus be retrieved in systematic manner.

Combining these powerful search capabilities and the script language of STN International is an excellent way to efficiently perform citation analysis. In the following we introduce ready-to-use scripts (NEUHAUS & LITSCHER, 2006), which determine citation counts for a given set of publications, applying either a fixed or a variable citation window and using either SCISEARCH or CAplus as a data source. When a fixed citation window is applied, each paper on the publication list is analyzed during a fixed period (MOED, 1996). In the case of a 3-year citation window, for instance, citations to papers published in 1997 are counted over the time period 1997–1999. Applying a variable citation window, this time period depends on the publication year of the respective paper. For example, citations to papers published in 1997 are counted over the time period 1997–2001, resulting in a 5-year citation window, whereas citations to papers published in 1998 are analyzed only over a 4-year time period. The scripts do not exclude self-citations from analysis, as identification by author name would be negatively affected by homonyms, spelling variants and misspellings of author names.

The scripts operate as follows: The bibliographic information, specifically first author, publication year, volume number and beginning page, are stored in a text file along with a unique record number for convenient data processing (see Figure 1). The script reads the bibliographic information line-by-line, transfers it into the reference format of SCISEARCH or CAplus, respectively, and searches for citations in the database reference field. For variable citation windows, citations are given year-on-year, providing a powerful way in which to explore citation data over time. Finally, citation counts are written along with the respective record number to a comma-separated file, which can be easily imported into a spreadsheet (e.g. Excel) or a statistical package for further data analysis (see Figure 2).

```
1200
Vaneeden F J M
1998
23
65
1201
Nussleinvolhard C
1994
266
572
```

Figure 1. Part of an input file for SCISEARCH, including record number, name of first author, publication year, volume number and beginning page of two papers published by Christiane Nüsslein-Volhard, who shared the 1995 Nobel Prize in Physiology or Medicine with Edward B. Lewis and Eric F. Wieschaus.

```
id;cctotal;cc1;cc2;cc3;cc4;cc5
1200;24;1;5;8;5;5
1201;38;0;7;12;14;5
```

Figure 2. Part of an output file listing record number, citation count from publication year to present, and citation counts for five consecutive years of two papers.

In a nutshell, the user must prepare the input file, including the bibliographic information, and load the script using the client software STN Express with Discover – all the rest is done by the script automatically. For both fixed and variable citation windows the script provides citation data that would require dozens of mouse clicks in Web of Science and manual data processing. Some analyses are even impractical using Web of Science. For example, Web of Science enables ranking of search results by publication year up to 2,000 records, thereby making it impossible to apply a citation window to a publication set cited in entirety more often than 2,000 times.

Limitations

Principally, citation analysis using the Science Citation Index can also be performed with Web of Science and other database hosts. With its script language, however, STN International simplifies the process of determining citation counts for a given set of publications since the entire process can be extensively automated.

For all of its benefits, citation analysis has some methodological and technical limitations that must be considered (e.g. VAN RAAN, 2005). First of all, SCISEARCH on

STN International goes back only to 1974, in contrast to Web of Science with back-years to 1945 and even to 1900 for a small subset of journals.* In research evaluation, however, this limited coverage back in time hardly matters, as the time period under assessment will rarely exceed the past ten years. The second point concerns inconsistencies and errors in cited references. When citing publications, authors may misspell author names, omit the middle initial, or confuse volume, issue and page numbers. Furthermore, author names with umlaut (e.g. Müller), with internal punctuation or spaces (e.g. O'Brian, van Eeden) and hyphenated names (e.g. Nüsslein-Volhard) are transcribed variously in the author field and the reference field of SCISEARCH and CAplus, respectively. Therefore, it is recommended to browse cited references in advance, to check which variants are relevant and should be included into the input file (see Figures 3 and 4). Both SCISEARCH and CAplus provide browsable indexes for cited author and cited work for looking up variants. Thirdly, the script is suited to determine citation counts for journal articles only. Counting citations to other publication types such as books and conference proceedings does still require manual effort, as such publications are not defined by first author, publication year, volume number and beginning page. Because cited work is abbreviated in an inconsistent and unpredictable manner, browsing cited references is inevitable in this case. Patent examiner citation counts, in contrast, can be determined with ease by searching the patent number in the cited work field in Web of Science or in the cited reference patent number field (RPN) on STN International.

```
=> FILE SCISEARCH
=> EXPAND COLANERI N F, 1990/RE
...
E5      1   COLANERI N F, 1990, V41, P11670, PHYS REV B/RE
E6      1   COLANERI N F, 1990, V42, P11160, PHYS REV B/RE
E7      2   COLANERI N F, 1990, V42, P1167, PHYS REV B/RE
E8    182   COLANERI N F, 1990, V42, P11670, PHYS REV B/RE
E9     35   COLANERI N F, 1990, V42, P11671, PHYS REV B/RE
E10     1   COLANERI N F, 1990, V42, P261, PHYS REV B/RE
E11     3   COLANERI N F, 1990, V42, P670, PHYS REV B/RE
E12     3   COLANERI N F, 1990, V42, PHYS REV B/RE
...
```

Figure 3. Cited references may be listed through the EXPAND command to look up inconsistent or erroneous citations. For example, a paper of N. F. COLANERI et al. published in 1990 in *Physical Review B* was cited correctly 182 times and incorrectly 46 times.

* The Century of Science initiative makes available around 850,000 publications from 262 journals published between 1900 and 1944.

```
L1   ANSWER 1 OF 1  SCISEARCH  COPYRIGHT (c) 2006
     The Thomson Corporation on STN
AU   Van Eeden F J M (Reprint); Holley S A; Haffter P;
     Nusslein-Volhard C
TI   Zebrafish segmentation and pair-rule patterning
SO   DEVELOPMENTAL GENETICS, (1998) Vol. 23, No. 1,
     pp. 65-76.
     ISSN: 0192-253X.


=> E VANEEDEN F J M, 1998/RE
...
E4       43    VANEEDEN F J M, 1998, V23, P65, DEV GENET/RE
...
```

Figure 4. The representation of author names in the author field and the reference field of SCISEARCH may be different, e.g. spaces within author names are removed in the reference field.

Conclusions

The database host STN International provides powerful browsing, searching and analyzing facilities, which allow for extensive citation analysis in SCISEARCH and CAplus. In conjunction with its script language, which is well suited for repetitive tasks, citation analysis can be significantly automated.

*

Note added in proof: Meanwhile Web of Science enables to analyze up to 100,000 records and to breakdown the citation history for a set of up to 10,000 records with the new Citation Report feature.

References

MARX, W., SCHIER, H., WANITSCHEK, M. (2001), Citation analysis using online databases: Feasibilities and shortcomings, *Scientometrics*, 52 : 59–82.

MOED, H. F. (1996), Differences in the construction of SCI based bibliometric indicators among various producers: A first overview, *Scientometrics*, 35 : 177–191.

NEUHAUS, C., LITSCHER, A. (2006). STN Express scripts for citation analysis in SCISEARCH and CAplus. Retrieved from http://www.psh.ethz.ch/people/neuhaus/stn

RIDLEY, D. D. (2001), Citation searches in on-line databases: Possibilities and pitfalls, *Trends in Analytical Chemistry*, 20 : 1–10.

VAN LEEUWEN, T. N., VISSER, M. S., MOED, H. F., NEDERHOF, T. J., VAN RAAN, A. F. J. (2003), The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence, *Scientometrics*, 57 : 257–280.

VAN RAAN, A. F. J. (2005), Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods, *Scientometrics*, 62 : 133–143.