Information Processing & Management Vol. 24, No. 1, pp. 39–44, 1988
Printed in Great Britain.

0306-4573/88 $3.00 + .00

USE OF A VIRTUAL INFORMATION SYSTEM FOR BIBLIOMETRIC ANALYSIS

HILARY D. BURTON
Technology Information Systems, Lawrence Livermore National Laboratory,
P.O. Box 808, 542, Livermore, CA 94550, USA

(*Received* 18 *March* 1987; *accepted* 10 *June* 1987)

Abstract—This article defines and discusses bibliometrics, particularly as carried out in automated systems. The specific requirements to which the data should conform in order to support bibliometric analysis are detailed and explained. Examples of earlier bibliometric work are presented, followed by a discussion of efforts supported by the Intelligent Gateway of the University of California's Lawrence Livermore National Laboratory. Difficulties in doing manual analysis are discussed and the article concludes with a recommendation for greater use of this type of analysis via the increasingly available automated tools.

The recently popular automated front ends and gateways have resulted in access to information resources that we can refer to as *virtual information systems.* The concept of virtual resources that underlies the TIS Intelligent Gateway developed at Lawrence Livermore National Laboratory is originally found in computer science, in virtual memories, virtual storage, virtual relations in data base management systems, etc. It is a particularly useful concept for information science. "The word virtual referring to computer facilities or to data indicates that the item in question appears to exist . . . when in fact it does not exist in that form" [1a].

"An extension of this concept leads to the complete virtual system wherein a . . . programmer . . . may see or visualize whatever system he wants [1b]. In fact, the Intelligent Gateway can provide a series of virtual systems, each capable of supporting the information requirements of a given user community. As new resources are developed, they can be added to the Gateway's repertoire. As users' needs change, the inventory of resources can be shifted accordingly. As new tools become available, they can be integrated" [1].

These virtual information systems allow a user to utilize a distributed set of resources via a single system. Resources can be distributed among multiple vendors or hosts and may involve heterogeneous types of service such as online data base systems, direct computer utilities, and interactive communications systems. Furthermore, in the more sophisticated virtual systems, the user can actually process and repackage the various kinds of data he obtains from the external resources.

One category of processing functions applied to bibliographic data is referred to as *bibliometrics.* According to the 1977 *Annual Review of Information Science and Technology* [2], in which the first review chapter on bibliometrics appeared, the term is attributed to Pritchard who first used it in an article in the *Journal of Documentation* in 1969. Bibliometrics refers to "all studies which seek to quantify the processes of written communication" [2]. Bibliometric studies generally deal with one or more types of raw data. Bibliometrics is generally more extensive than postprocessing, which has come to mean downloading and reformatting of data. Postprocessing can be easily accomplished using one of the many commercially available packages such as PBS, SCIMATE, or even straight word processing software.

Bibliometric analysis is not yet widely practiced, yet it can produce tremendous benefits for the information scientist or research manager. One likely reason for the general lack of occurrence is the difficulty and tedium of doing such analysis manually. Until recently, there has been a lack of generalized software to do such analysis. A notable and major

exception is the software and data bases developed by the Institute for Scientific Information (ISI). The "maps" of subject disciplines which Garfield and others have been producing from subsets of the Science Citation Index and Social Science Citation Index are impressive examples of the kinds of information bibliometrics can yield [3].

The three major classes are:

1. Analysis of source data, that is, how many items were published—either individual articles or journals, etc. How many authors produced them, over what time period, at what cost, etc. These studies use source references as discrete items to produce an analysis of the aggregate.
2. Co-citation studies and bibliographic coupling comprise another aspect of bibliometric analysis. Here the cited or citing works of a publication or item are studied. ISI has done extensive analysis in this area and has provided much of the base data used by others.
3. Combinations of the two kinds of data are most commonly used in indexing studies and sociology of science and science policy studies. Indexing studies attempt to determine if citation information could produce the same or better retrieval terms for an item than the assigned subject indexing set [4]. Sociology of science studies analyze the origin and development of subject disciplines, "styles" of research, and communication networks such as the invisible college [5].

Metainformation, or information about information, is what is actually produced by bibliometric analysis. This metainformation is particularly useful to three categories of information users: (1) to the librarian or information specialist as a means to better understand the environment in which service and support are offered, (2) to the research manager as a means to measure and evaluate productivity of his or her own staff and how it compares to other comparable units, and (3) to the research analyst who may wish to define a new research project or compare his or her own progress to that of a broader spectrum—such as other institutions, related disciplines, or other nationalities.

Each of these three kinds of metainformation users may look at different aspects and draw from different sets of base data, but the basic analytical capability required is the same. Citations must be capable of being parsed into consistent, discrete elements; and it must be possible to calculate correlations and distributions both across citations and within citation elements (i.e., inter- and intracitation analysis).

In manual analysis, the human generally creates the citation consistency through his or her own intellectual effort. For example, a human can easily identify an author's name whether printed in upper and lower case or in all upper case. Increasingly more intelligent decisions are often an implicit part of any manual analysis. However, the computer's requirement for literal and explicit data presentation results in a need for stringent citation translation before any automated analysis can begin. Once an effective translation is achieved, it is relatively straightforward to apply some simple analyses such as frequency distributions and permutations. More complex analyses will depend on the level of detail into which the citations have been parsed and the availability of appropriate statistical techniques.

Why the degree of detail is important is not always intuitively obvious. Two brief examples may be helpful. If a citation has four authors and seven subject headings in the respective author and keyword fields, there are multiple ways to treat them. The simplest translation would produce a string of characters labeled author (Example 1, Translated 1) and a second string labeled keywords or subjects (Example 2, versions 1 and 2). A more useful translation of the author field would produce a list of items where each item included the complete set of information as given for an author (Example 1, Translated 2). There would be four items in the list. Even more useful would be the production of an author list with four entries using a standardized format such as last name, first initial, second initial. In this version, case, punctuation and spacing would be standardized as well as the substitution of initials for full names, etc. (Example 1, Translated 3).

Example 1 (authors)

Original version of author information.

Brown, John A., Smith, D. B., Scott, Frances, L. X., and G. Anderson

Translated version

1. Brown John A Smith DB Scott Frances LX G Anderson
2. Brown John A, Smith DB, Scott Frances LX, G Anderson
3. Brown JA, Smith DB, Scott FL, Anderson G*

For the subject field, a similar procedure is necessary. A single string of indexing terms yields considerably less information than one that indicates term relationships.

In the following example, the punctuation of the subject terms becomes the determinant of what the content is. For instance, does "Green" modify Apple or Aphids? Does "California" modify Regulations or Chemical Control? Clear and consistent translation of such data are fundamental to useful analysis.

Example 2 (subject terms)

Unformatted indexing terms. Apples Green Aphids Control Chemical California Regulations

Formatted indexing terms—version 1. Apples, green; Aphids; Control, Chemical; California, Regulations

Formatted indexing terms—version 2. Apples; Green Aphids; Control; Chemical; California; Regulations

Superficial bibliometric analysis can be carried out on even roughly translated collections of data. However, as one obtains greater data specificity and consistency the value and variety of bibliometric analyses increase. In the past, customized software has been developed to operate on specially collected data to produce analyses required for individually funded projects. For example, in the 1970s, Computer Horizons carried out extensive bibliometric analyses for the National Institute of Health. This program was to investigate "the utility of publication and citation analysis in studying the structure and dynamics of research communications in the biosciences" [6].

Subsequently, Computer Horizons developed techniques "to measure the interactions between basic and clinical activities, to study the relations among biomedical disciplines, and to assess the effect of federal funding policies on publications [6]. Their data base was jointly comprised of project research data, of the type that the Smithsonian Science Information Exchange (SSIE) or U.S. Department of Agriculture's CRIS system collected, and data on biomedical publishing from the Corporate Index tapes of Science Citation Index. From these two sources, Computer Horizons created a subset of data relevant to their area of interest and then developed various analytical routines.

There are instances of other projects developing both their own data bases and analytical tools [7,8]. But, over the past five years, there has been a growing awareness of the value that analysis of data already available in the major online information systems can provide. Most recent projects have involved downloading from one or more services and then analyzing either the search term postings, the citation content, or some combination of both [9, 10]. In a few cases, downloaded data have been reformatted for analysis by a spreadsheet program such as LOTUS 1-2-3. But the use of powerful analytical tools as were developed in the 1970s has not occurred. Instead, most of the current work consists of search term analyses done while connected online to the search service or analysis done on a PC after downloading a set of citations.

An ongoing project at the Lawrence Livermore National Laboratory involves development of a generalized bibliometric capability to be used with the Technology Information Systems (TIS) Intelligent Gateway. The Gateway provides the user with a "virtual

*Some systems will add a symbol to indicate that no second initial is available.

information system" and the bibliometric tools allow analysis of data drawn from disparate sources [11].

In three recent projects, several thousand bibliographic records were downloaded from both federal and commercial data base services. Records were downloaded from several files offered by BRS, Lockheed DIALOG, and Pergamon's ORBIT service as well as from NASA and the Department of Energy's RECON systems, and from the Department of Defense online system, DROLS. Access and downloading were accomplished using the Intelligent Gateway. The Gateway allows a user to issue a "connect" command (i.e., connect SDC or connect DROLS) and be connected to the target system via the most efficient telecommunications channel available. This can range from 300 baud in the worst case to 4800 or 9600 baud where higher-speed access is supported.

After connecting to the desired resource, the user can save the results of the session by typing a simple three-key sequence and naming a file where the downloaded information will be stored. At any point, the downloading can be interrupted or stopped; but if no such interruption occurs, the file will be closed at the conclusion of the session when the user logs out.

Obtaining information using the Gateway is very easy and large files can be built very quickly. However, several points should be stressed: (1) Each data base producer has his or her own policy on downloading charges and it is the user's responsibility to comply with that policy. Depending on the producer, these charges can be substantial. (2) Unevaluated masses of data become more of a burden to the end user than an aid. (3) Merging data from several sources creates both physical and intellectual problems.

Policies on downloading charges range from enlightened to autocratic. Indeed, until recently, one of the world's largest search services did not offer retrieval output in a downloading format in part because of the economic questions raised. A review of current downloading policies of the major online services can be found in a recent proceedings [12].

In the projects discussed below, negotiation was conducted directly with each data base producer. Online search and display costs were paid at the standard commercial rates of the search service and separate payment was made to the appropriate vendors for the downloading activity. These varied from royalties of $1.00 per record retrieved to a request to acknowledge the data base producer in any report listing the citation. Note that one of the projects was an in-house research project from which no explicit citation lists resulted. Another project resulted in a large bibliography that had very limited in-house distribution. However, it certainly appears that data base producers have a far more enlightened viewpoint than they did 5 or 10 years ago [13].

Following is an outline of the steps that were necessary to build a usable file. In projects that obtain all of their information from a single data base on one host, all of these steps may not be necessary. However, formats may change over time or a vendor's treatment of the data may vary even within a single data base. One example of a simple change that has major ramifications is the inclusion—after 1975—of an explicit date of publication field in most data bases. Time series analysis on pre-1975 is not feasible until a special program can be written to scan the data and create a date field.

To build an analyzable file:

- Refine the search strategy until satisfactory retrieval results are obtained.
- Download the citation in the fullest and most explicitly tagged available format (e.g., format 4 in DIALOG; 'print full' in ORBIT).
- Repeat steps 1 and 2 for each relevant data base, creating discrete citation files for each data base.
- Add missing fields as necessary for analysis (e.g., year of publication, country of origin, language).
- Translate records to common format.
- Identify duplicate citations. Duplicates can occur within the same data bases as well as across data bases. It is easier (but not easy) to identify duplicates after the records have been translated. Even after translation, there are amazing differences in citation content and format that make duplicate identification difficult. Within the

Gateway processing routines, we have developed a program that identifies duplicates using a default key of author, title, date, and source or allows a user to define his or her own key using whichever combination of fields he or she prefers.

- Eliminate less-preferred form(s) of duplicates. There are occasions when a search will produce the same citation from three or more data bases. Which is the preferred version? In some cases, it may be data base dependent; in others, system dependent; and in others, variable. One sophisticated solution to this problem is that used by the University of California's MELVYL system [14], which creates a merged record carrying the unique portions of each "duplicate."

Initial bibliometric analyses will likely be obvious (e.g., how may publications were produced, by whom, on what subjects, etc.). But as the effort proceeds and more metainformation is created, new questions will develop. Often, new correlations or analyses of subsets of the data will be requested. For example, do the authors from academic institutions publish in different journals than those from other institutions such as government? Do authors from different geographic locations vary in their productivity? Examples of the kinds of more complex questions that are satisfied by bibliometric analysis can be found in many disciplines, but the biomedical field probably provides the best model. McAllister and Narin completed a study in 1982 to consider "the relationships among the number of papers published, the citation influence and subject emphasis of the journals in which the papers are published, and various external characteristics of the (medical) schools such as public versus private control, geographic location, NIH funding, and peer ratings [15]. Specific kinds of questions dealt with the strength of the relationship between funding and numbers of papers produced and differences in medical school output with respect to subject, research emphasis (clinical vs. basic) and citation influence.

In addition to subset analyses, questions concerning the population outside the data set may occur: What professional affiliations are not represented in the data? What subject areas have dropped out or not yet shown up? And, more broadly and problematically, what kinds of data may be needed but are not available? In some cases, more sophisticated analytical routines can compensate for shortcomings in the data. However, in spite of the mass of data available online, there are still gaps where data will have to be collected manually and entered into the data base to ensure comprehensive analysis. The situation is tremendously improved over what it was just 10 years ago, however.

As Narin and Moll predicted in 1977: "The bibliometrician will benefit from the expanding scope and standardization of information in computerized data bases as well as from their increased availability. . . . The pace of development of future techniques and applications will be closely related to the economies of time and money made possible by improvements in both computerized data bases and citation indexes" [2, p. 50].

Progress in the 1980s has included extensive increases in data base coverage, rapid development of new data bases, and release of a wide variety of "user friendly" tools to improve and facilitate access to existing services. With the increasing availability of these data and the means to obtain them, it is now time to utilize bibliometric tools to produce information about information.

This metainformation can assist the information specialist in better managing his or her own resources. But it can also assist the specialist in better understanding and even predicting changes in the structure of the information resource environment in which he or she exists.

REFERENCES

1. Burton, H. D. The Intelligent Gateway: A dynamic resource environment. Information Service and Use (in press) 1986. The referenced works are:
   a. Martin, J. Computer Database Organization. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1977: 16.
   b. Meadow, C. Man Machine Communication. New York: Wiley; 1970: 96.
2. Narin, F.; Moll, J. K. Bibliometrics. Annual Review of Information Science and Technology. Volume 12. Chapter 2. White Plains, NY: Knowledge Industry Publications, Inc.; 1977: 35.
3. Garfield, E.; Sher, I. H.; Torpie, R. J. The Use of Citation Data in Writing the History of Science. Philadelphia, PA: Institute for Scientific Information; 1964. 76p.

4. White, H. D.; Griffith, B. C. Quality of indexing in on line databases. Information Processing and Management. (In press) 1986.
5. Zuckerman, H. Nobel laureates in science: Patterns of productivity, collaboration, and authorship. American Sociological Review 32(3): 391–403; 1967.
6. Frame, J. D.; Narin, F. NIH funding and biomedical publication output. Federation Proceedings 35(14): 2531; December 1976.
7. Salisbury, G. W. Research productivity of the state agricultural experiment station system: Measured by scientific publication output. University of Illinois. Agricultural Experimental Station Bulletin 762; July 1980. 65p.
8. David, H. G.; Piip, L.; Haley, A. R. The examination of research trends by analysis of publication number. Journal of Information Science 3: 283–288; 1981.
9. Hibbs, J. E.; Bobner, R. F.; Newman, I.; Dyer, C. M.; Benz, C. R. How to use online databases to perform trend analysis in research. Online 8(2): 59–64; 1984.
10. Bobner, R. F.; Newman, I.; Hibbs, J.; Benz, C. R.; Dyer, C. M. Historical policy capturing through use of computerized databases and trend analysis techniques. Paper presented at the Annual Meeting of the American Educational Research Association. New York: ED 216957 30p; 1982.
11. Hampel, V. E.; Bailey, C.; Kawin, R. A.; Lann, N. A.; McGrogan, S. K.; Scott, W. S.; Stammers, S. M.; Thomas, J. L. "TIS"—An Intelligent Gateway computer for information and modeling networks—Overview, UCRL-53439, Lawrence Livermore National Laboratory, Livermore, CA, August 1983. 9p.
12. Weinberg, B. H.; Benson, J. A., eds. Downloading/uploading online databases and catalogs. Library Hi Tech Special Studies Series, Number 1. Ann Arbor, MI: Pierian Press; 1985. Appendix A. 79–85.
13. Hawkins, D. T. Machine readable output from on-line searches. Journal of American Society for Information Science 32(4): 253–256; 1981.
14. In-Depth: University of California MELVYL. Parts 1, 2, Information Technology and Libraries 1: 4 (December 1982): 350–381; 2: 1 (March 1983): 58–115.
15. McAllister, P. R.; Narin, F. Characteristics of the research paper of U.S. medical schools. Journal of American Society for Information Science 34(2): 123–131; 1983.