INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

WHAT'S THE USE OF BIBLIOMETRICS ?

J. TAGUE

School of Library and Information Science
University of Western Ontario
London, Ontario, N6G 1H1, Canada
519-661-3542

Abstract

Bibliometrics has been heralded as the theoretical base of information science. But do the mathematical models which it develops have any practical value ? A number of applications have been proposed; few have actually been implemented. Some are based on flimsy evidence, some are hyper-complicated, some are simply impractical. This paper develops a couple of scenarios in which bibliometric theory could have a practical value. The particular application areas are file design and journal collection management. In the examples presented here, procedures are introduced for estimating the parameters of the Zipf/Lotka and Bradford distributions from samples. However, if these applications are to be of maximal use, more information about the errors of estimate should be included. This is an area to which bibliometricians could usefully turn their attention.

Bibliometrics has been heralded as the theoretical base of information science. A number of mathematical models have been proposed for bibliometric variables, that is, variables connected with the production and use of records or text. The models of Lotka[1], Zipf[2], and Bradford[3] are well known, and other useful approaches have been taken by more recent investigators.

Those of us who have studied the forms and relationships of these models have done so, primarily, because of their intrinsic interest and because of the pleasure we take in this kind of analysis. There is, however, another point of view, as the title to this paper attests. It is a complaint frequently heard from students in schools of information science or studies, who have been required to study bibliometrics. Why, they ask, should we have to learn all this math ?

Students do have a way of bringing us down to earth. When I teach bibliometrics, I avoid mathematical demonstrations for their own sake, no matter how elegant. Students are practical. Most of them will become professional information specialists. They need to be convinced, not of the beauty of bibliometrics (and it is that), but of its usefulness.

A teacher can, of course, point to the practical applications that bibliometricians have suggested of the models or "laws" of bibliometrics. However, when these are examined closely, they often turn out to be little more than textbook exercises rather than procedures that could actually be used by a decision-maker.

This paper will develop a couple of scenarios in which bibliometric theory could have a practical value. The particular application areas are file design and journal collection management. In presenting these applications, I hope, also, to point out the gaps in our knowledge, the parts of the theory that need to be more fully developed if the applications are to be of maximal use.

J. Tague

The exposition will focus on the classical bibliometric distributions of Zipf, Lotka, and Bradford, rather than on bibliometric techniques such as co-citation clustering. For the latter, the interesting questions concerning use are part of the application area itself rather than the bibliometric theory; for example, is it legitimate to use quantitative measures of researcher impact ?

1. FILE DESIGN AND THE ZIPF/LOTKA DISTRIBUTION

Lotka's distribution or "law", originally introduced to describe the distribution of publications over authors, and Zipf's distribution or "law', originally introduced to describe the distribution of word types in a text, are special cases of a general model which can be presented in either a size or a rank form:

$f(x) = a/(x+c)^b$, $x=1,2... \; x_{max}$

$g(r) = a'/(r'+c')^{b'}$, $r=1,2... \; r_{max}$

where f(x) is the number of types (authors, words) with x tokens (papers, occurrences) and g(r) is the numer of tokens of the rth ranking type. Usually, the parameters c and c' are assumed to be 0 and sometimes the parameter x_{max} is assumed to be infinitely large.
One application, originally suggested by Samson and Bendell,[4] lies in the area of file design. Suppose, in a boolean online retrieval system or online catalog, an inverted file structure is used to access specific names or keywords or subject headings (entries) in the records. Typically, these entries will be stored in the leaf or terminal nodes of a B-tree, which will point to lists in the postings file of pointers to all the records containing the designated entry. In designing the two files (index B-tree, postings), the system developer needs to know the number of distinct entries, or the number of terminal nodes, in the B-tree (entry types), the number of document pointers in all of the lists in the postings file

(entry tokens), the maximum length of a postings list, and the relationships among these values. This information will enable the developer to determine the storage needed for the two file structures.

The Zipf/Lotka size distribution can be used to answer the developer's questions if it is reasonable to assume that the entry types and tokens follow this pattern. To describe the situation more specifically, let f(x) be the number of entry types with x postings, t the number of entry types, m the number of entry tokens, and x_{max} the number of tokens in the longest postings list. The distribution of entries can then be described by the following function, assuming a Zipf distribution with parameter c=0:

$$f(x) = \frac{t/x^b}{\sum_v 1/v^{b'}}, \quad x=1,2\ldots x_{max}. \qquad (1)$$

and the total number of postings will be

$$m = \sum_x xf(x) = \frac{t \sum_x 1/x^{b-1}}{\sum_x 1/x^b}. \qquad (2)$$

The number of terminal nodes in the B-tree, t, can be determined if the Zipf model is appropriate and if b, x_{max}, and m are known or can be estimated. Usually, these values will not be known in advance of an implementation and so we need to consider how they may be estimated.

Following the usual procedure, the two parameters of the Zipf distribution, b and x_{max}, may be estimated from a random sample of the database to be input. Nicholls[5] and Tague and Nicholls[6] have described the use of maximum likelihood estimators for b in this situation and Tague and Nicholls have derived an approximate upper confidence limit for the x_{max} parameter.

A particular situation in which these estimators might be used is in the design of an online catalog structure for an existing card catalog. In this case, estimates of m, the total number of entry tokens, and b and x_{max}, the Zipf distribution parameters may be obtained using two random samples. The first sample would be of catalog card drawers, the second of cards within the drawers. From the sample of drawers, the average number of catalog cards per drawer could be estimated and hence the total number of postings (m) as the product of the number of drawers and the average number of cards per drawer. The sample of cards constitutes a sample of types; the number of cards with that heading (author, title, subject) could be determined, and from these figures, which constitute a sample of x values, the frequency distribution parameters b and x_{max} could be estimated using one of the techniques described in Tague and Nicholls. In particular, the exponent b can best be estimated by its maximum likelihood estimator, the solution to the following equation,

$$\sum_i \log_e x_i/n = \sum_x (\log_e x)/x^b \;/\; \sum_x 1/x^b \qquad (3)$$

where the summation on the left side of Equation (3) is over all n sample values and that on the right side over all possible values of x up to its maximum value of x_{max}. A "quick and dirty" estimate of b is based on the first two empirical frequencies, f'(1) and f'(2).

$$b' = \log_e(f'(1)/f'(2))/\log_e 2. \qquad (4)$$

The maximum likelihood value for x_{max} is its sample value x_{max}'.
Another useful estimate is the approximate 100 (1-p)% upper confidence limit U given by:

$$U = [p^{1/n}/(p^{1/n} - 1 + 1/x_{max}')]^{1/(b-1)} \quad (5)$$

Once m, b, and x_{max} have been estimated, t may be estimated from Equation (2).
The sample thus provides the information needed to determine the storage requirements for the B-tree, and the postings file and the average and worst case number of disk accesses to retrieve the locations of all items relating to a particular entry in the catalog. If k represents the number of index file entries per disk block and h the number of postings file entries per disk block, then the average number of accesses a and the maximum number of accesses w will be:

$$a = \log_k t + m/(th)$$

$$w = \log_k t + x_{max}/h$$

where, in the latter expression, x_{max} could be estimated by the expression in (5).
A rank rather than a size frequency approach could have been taken in this estimation problem, although the dependent nature of ranks makes it difficult to apply to them the standard statistical procedures based on independent observations.
Nelson and Tague[7] suggest that a mixed or split size-rank distribution best fits the distribution of terms used to index journal papers: in other words, a size-based distribution function is best for low frequency terms and a rank-based function for high frequency terms.

2. JOURNAL COLLECTION MANAGEMENT AND THE BRADFORD DISTRIBUTION

Two models have been proposed relating to journal collection management: Bradford's distribution describing the scattering of papers in a subject area over journals and the negative exponential distribution for the obsolescence of journal papers. In this paper we will concentrate on applications of the former distribution, as the appropriateness of the latter model has been a matter of some controversy.
Bradford's "law" has a number of forms, but all essentially describe the relationship between the rank of a journal when journals are arranged in order of decreasing productivity in a subject area and the cumulative number of papers in journals of that rank or better. It has been proposed that the Bradford distribution can provide library managers with information about what journals to collect if they wish to cover a specified percentage of the literature of a subject area.
Does Bradford's distribution in fact provide this information ? I suggest the answer to this question is "no". Bradford's law, in its commonly-stated forms, has very little predictive power. In order to decide whether or not to include a journal in a library collection, the library manager must know the rank or zone of the journal. To determine the rank or zone, the manager must find out the number of relevant papers in every journal which publishes papers in the subject area and then order journals by this number.
Once the library manager has carried out the procedure described in the preceding paragraph, he or she has all the information needed to make a decision concerning which journals to include in the collection in order to cover a desired percentage of the literature. Fitting to a Bradford distribution will provide no additional information. For example, Table 1 shows the distribution of papers over journals in terms of r, the cumulative number of journals, and G(r), the cumulative proportion of papers in the most productive r journals. It is immediately apparent that approximately 1/2 of the journals provide approximately 3/4 of the papers. Whether or not the distribution fits a Bradford distribution is of no particular concern.

Table 1

Cum. Num. Journals (r)	Cum. Prop. Journals	Cum. Prop. papers G(r)
1	1.3	11.5
2	2.5	18.4
3	3.8	21.8
5	6.3	27.6
9	11.3	36.8
18	22.5	52.3
39	48.8	76.4
80	100.0	100.0

What one would like to be able to do is to predict from a sample of papers what the over-all shape of the Bradford distribution is, so that one could then draw some conclusions about numbers of papers and journals in the population of all papers in the subject area. For example, suppose a library manager wishes to develop a collection of journals in some subject area. The manager has access, in printed or online form, to indexing or abstracting services which cover the subject area. He or she does not have time or resources to look at all entries in this database to determine all of the papers and journals in the area. However, he or she does have time to take a sample of records from the database, i.e., a sample of papers from the field.

As in the earlier application, the major problem is how to estimate the distribution parameters from a sample. This problem will be examined using the form of Bradford's law which specifies that journals in a discipline be ordered according to number of papers in the discipline which they contain and then divided into some number of zones, z, in such a way that each zone contains an equal number, p, of papers. Then, according to Bradford theory, the number of journals t_j in the jth zone is given by:

$$t_j = r_0 k^{j-1}, \quad j=1,2 \ldots z,$$

where r_0 is the number of papers in the first or nuclear zone.
So the total number of journals is

$$t = \sum_{j=1}^{z} t_j = r_0(k^z - 1)/(k-1)$$

and the total number of papers is m=zp.

Suppose a random sample of papers is selected from the indexing/abstracting database, i.e., a sample is selected in such a way that each paper in the database has the same chance of being in the sample. Since, according to Bradford theory, the zones have equal numbers of papers, this procedure means that each zone will have an equal probability of being represented by the papers in the sample. The chance that a particular journal will be in the sample will thus depend on the number of papers which it contains in the subject area.

The manager wants, from this sample, to estimate the number of journals in each zone in the population and from this how many journals will be needed to achieve a specified coverage. Thus, the manager wants to estimate r_0 and k, the number of papers in the nuclear zone and the Bradford multiplier, given that there are z zones. The number of zones, z, would be assigned arbitrarily.
The expected number of papers in each zone in the sample will be the same for all zones. Thus, if the total sample size is n, the expected number of papers in each zone will be n/z=q. Then the zone ranges can be estimated by determining the journal of each paper in the sample and the number of sample papers from each sample journal, then ranking the sample journals by productivity, and finally dividing the sample journals into z zones, each containing an equal number, q, of papers. The expected number of journals in the jth zone will be qr_0k^{j-1}/p, but the actual number will be a random variable.
The expected number of papers per journal in the jth zone of the population is $p/(r_0k^{j-1})=m_j$. This value can be estimated by the average number of papers per journal in the sample from the jth zone, i.e., by the sample mean q/t'_j, where t'_j is the number of journals in the jth zone in the sample. These estimates thus provide a set of z equations:

$$q/t'_j = p/(r_0k^{j-1}), j=1...z,$$

which can be transformed to linear form:

$$\log_e q - \log_e t'_j = \log_e(p/r_0) - (j-1)\log_e k, \quad j=1... \ z \ .$$

We can find the least squares estimates for $\log_e(p/r_0)$ and $\log_e k$ and from these estimates m_1' and k' of p/r_0 and k. In order to uniquely determine r_0, additional information is needed. The most likely piece of information to be available is the number of papers in the most productive journal, since this journal will likely be known to the library manager.
As Egghe[8] and others have shown, the Bradford zonal distribution described above implies that the distribution of journal ranks will have the form originally introduced by Leimkuhler:

$$R(r) = \{p \log_e[1 + ((k-1)/r_0)r]\}/\log_e k \ . \qquad (6)$$

Where R(r) is the cumulative number of papers in the first r journals. If the most productive journal is known to be one with p_0 papers, then, substituting $R(r)=p_0$
and r=1 in Equation (6) and using the estimate m_1' and k' of p/r_0 and k derived from the earlier analysis we get two equations in the two unknowns p and r_0. Solving each for p we get:

$$p = m_1'r_0 \qquad (7)$$

$$p = p_0\log_e k/\log_e[1+(k-1)/r_0] \qquad (8)$$

To solve these for r_0 and p we may graph the two equations, as in Figure 1, and find the intersection of the two graphs, or solve by iterative methods the equation:

$$k^{p_0/m_1'} = [1 + (k-1)/r_0]^{r_0} .$$

Once r_0 and k have been estimated, the number of journals and papers in each zone may be estimated.

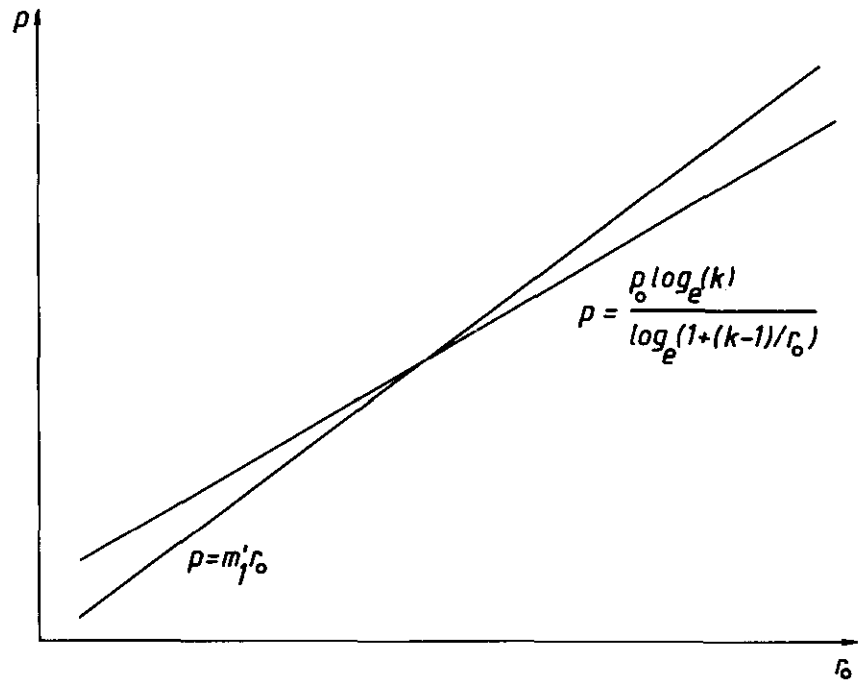

Figure 1: Graphical solution to determine r_0, number of journals in nuclear zone

Implications

In a recent letter to the editor of the Journal of Documentation[9], Line suggested that any question concerning bibliometrics should have attached to it two other questions: "Who precisely wants to know ?" and "For what precise purpose is the information wanted?" He suggests that only in this way can we avoid consideration of useless questions.

The major point of this paper is that, to be truly useful, bibliometrics needs to pay more attention to statistical inference, to ways of deriving conclusions about bibliometric variables from samples. In the examples presented here, procedures have been introduced for estimating the parameters of the Zipf/Lotka and Bradford distributions from samples. However, to be truly useful, the estimates should include some information about the distribution of the errors of estimate, at least the standard deviation of these errors. This is an area to which bibliometricians interested in further developing the theory could usefully turn their attention.

REFERENCES

[1] Lotka, A.J., The frequency distribution of scientific productivity. Journal of the Washington Academy of Science 16(12) (1926) pp. 317-323.

[2] Zipf, G.K., Human Behavior and the Principle of Least Effort, (Addison-Wesley, 1949).

[3] Bradford, S.C., Sources of information on specific subjects. Engineering 137 (1934) pp. 85-86.

[4] Samsom W.B., and Bendell, A., Rank order distributions and secondary key indexing. Computer Journal 28 (1985) pp. 309-312.
[5] Nicholls, P.T., Empirical validation of Lotka's law. Information Processing and Management 22 (1986) pp. 417-419.
[6] Tague, J.M., and Nicholls, P.T., The maximal value of a Zipf size variable: sampling properties and relationship to other parameters. Information Processing and Management 23 (1987) pp. 155-170.
[7] Nelson, M.J. and Tague, J.M., Split size-rank models for the distribution of index terms. Journal of the American Society for Information Science 36 (1985) pp. 283-296.
[8] Egghe, L., Consequences of Lotka's law for the law of Bradford. Journal of Documentation 41 (1985) pp. 173-189.
[9] Line, M.B., Obsolescence studies: a plea for realism. Journal of Documentation 42 (1986) pp. 46-47.