Pergamon

Information Processing & Management, Vol. 33, No. 4, pp. 523–527, 1997

0306-4573/97 $17+0.00

PII: S0306-4573(96)00075-1

A NEW EXPLANATION OF THE GEOMETRIC LAW IN THE CASE OF LIBRARY CIRCULATION DATA

THIERRY LAFOUGE and SYLVIE LAINÉ-CRUZEL

Recodoc, Université Claude Bernard Lyon 1, Bât 721, 43 Bd du 11 Novembre, 69 622 Villeurbanne cedex, France

Abstract—In this article we propose a mathematical model using the probability formalism in order to explain why we often observe a geometrical law in the case of library circulation data. To obtain this result we used simple techniques based on convolution theory. © 1997 Elsevier Science Ltd

1. INTRODUCTION

The crucial point of most of bibliometric studies is the observation of the frequencies of events, generally called bibliometric distributions. We can quote:

- the productivity of scientific journals in articles on a given subject
- the distribution of words in a text
- the productivity of scientific authors
- the circulation of journals (periodicals) within a library or a documentation center

In these informetrics studies the notion of a collection of 'sources' producing distinct items is fundamental, and has been pointed out by many authors. Studies have modelled these problems by defining an 'Information Product Process (IPP)'(Egghe, 1990) which generalizes relationships between sources and items. IPPs, as explained above, involve two attributes and so could be called two-dimensional informetric studies. Our study concerns a three-dimensional IPP.

Journals→Articles→Occasions when they are borrowed.

Here we have two source sets (Journals, Articles) and one item set (Uses). We have already studied the distribution of requests in each volume starting from user requests for articles in a host (Lafouge & Delarbre, 1989).

The results obtained allowed us to confirm a classification of the periodicals: i.e. fundamental periodicals, where all the articles seemed 'important', and technical or applied periodicals where requests focused on more 'specialized' or 'practical' problems.

In these infometrics studies, the common point of these distributions is their discrete domain: $r=1,2,3$. The case $r=0$ may also be encountered, the so-called 'no use'. Haitum has studied this type of statistics in particular, which are well known under the name of 'Zipfian statistics' (Haitun, 1982a, 1982, 1983). Many numerous studies consist of adjusting these curves using theoretical models, by multiplying the parameters (Sichel, 1985; Burrel & Fenton, 1993) and the methods of adjustment.

2. PROBLEM

Some authors using urn models to describe statistical after-effects ('contagion') have tried to explain these phenomenas. Polya (Feller, 1968) uses the following statistical model: an urn contains b black (failure) and r red balls (success). A ball is drawn at random, then c balls of the drawn colour are added in the urn. Polya calculates the probability that after n drawings the

result is k red balls.

The limiting form of the Polya distribution is a negative binomial law.

Price (1976) uses this model in an article which became a fundamental basis for bibliometric studies.

In his model, c balls are added in the urn, only if the drawing is a success. Then he calculates the density of the corresponding distribution, knowed under the name of 'law of the cumulative advantage'. This law may be explained as follows: the more items (successes) a source produces, the higher is the probability of new successes. He also demonstrates that the limiting forms are Bradford's law and Lotka's.

Distributions relating to library circulation data are often adjusted with geometric distributions (Burrel, 1986). But the negative binomial law gives better results (Leemans *et al.*, 1992), especially when the 'no use' case is taken into account. (Recall that the negative binomial law is a generalisation of the geometrical law.) In this paper, we will build a model which may explain these geometrical distributions. To obtain this result, we used the following process:

1. We have eliminated the effect of cumulative advantage. A volume is said to have been used i times if i articles of this volume have been used at least once. The user demand for an article is taken into account if the article has been requested at least one time. So, two or more users who have requested the same article are considered as being equivalent to one user for this article.
2. We have eliminated the effect of the case of 'no use': it is very important to take or not to take the case of 'no use' into account, because it often totally changes the adjustment of the distribution. We use standard statistical distribution to adjust these curves.

But if we may accept easily an adjustment of a curve in the discrete interval [1. ∞], where we may admit that the measured phenomenas are homogeneous, it is much more difficult to consider the result at the point 0 ('no use') as being of the same nature: actually, when an article is not requested, it is not a failure but more a non-event. Unfortunately, the position of this particular point will be of a great importance for the result of the adjustment.

Egghe (1994) carries out a continuous approach of an equivalent problem based on convolution theory. He gives a new explanation of the historical Lotka's law. He also shows (see Section 3.3) that we have the following exact stability result for the geometrical distribution. In his paper he uses the continuous model. Similar techniques, using a discrete model, will be used in this article.

3. MODEL

3.1. Main principles

The circulation of volumes of a collection of journals, each containing a certain number of articles, is observed. We suppose that the use frequency for each article over a given period is known. A volume is said to have been used i times if i articles of this volume have been used at least once. It can be noted:

C_{ji}: Number of combinations of i elements in a set of j elements($i \leq j$).
$B(j\ p)$: The binomial distribution of parameters j and p ($0 \leq p \leq 1$ $j \in \mathbb{N}$.)
$P(i)$: Probability that a volume be used i times, $i=0,1,2,...$
$p(i)$: Probability that a volume contains i articles, $i=1,2,...$
p_0: Probability that an article will not be used ($0 \leq p_0 < 1$).

We will define a general model for P:

We suppose that we have: $B(j\ 1-p_0)(i)$ the fraction of the volumes having j articles, used i times (each used article is only counted once). It is equivalent to the fraction of volumes in which i articles have been used at least once, $i \leq j$.

We can therefore write

$$P(0)=\sum_{j=1,\infty} B(j1-p_0)(0)p(j) \tag{0}$$

$$P(i)=\sum_{j=i,\infty} B(j1-p_0)(i)p(j)\ i=1,2\ldots$$

It is easy to show that P is a probability.

P designates the probability that i articles belonging to a volume have been used at least once.

If the expectation of p is m, then we can easily demonstrate (Lafouge, 1995) that the expectation of P is $m(1-p_0)$.

The case without 'no use'

We suppose now that P never takes the value zero. This means that all the volumes have been used at least once. It is possible, however, that some articles have never been used, so we define the distribution of volumes, such that we note PP as:

$$PP(i)=\frac{P(i)}{1-P(0)}\ i=1,2\ldots \tag{1}$$

where we have: $P(0)=\sum_{j=1,\infty} p_0^j p(j)$.If we choose the probability defined in Equation (1), it can be noted that if $p_0=0$ then we have $P=p$, so we give the following theorem.

Theorem 1. Let p be a geometrical distribution of expectation m taking values for $i=1,2\ldots$ then the probability defined in Equation (1) is a geometrical distribution of expectation $m(1-p_0)+p_0$.

Proof. p being geometrical, we can write: $p(j)=q(1-q)^{j-1}, j=1,2\ldots$ where the expectation m can be written: $m=1/q$ with $0<q\leq 1$.

Calculation of the denominator:

$$\sum_{j=1,\infty} p_0^j p(j)=\sum_{j=1,\infty} p_0^j q(1-q)^{j-1}=qp_0\sum_{j=1,\infty} p_0^{j-1}(1-q)^{j-1}=qp_0/(1-(1-q)p_0)$$

therefore: $1-P(0)=(1-p_0)/(1-(1-q)p_0)$.

Calculation of the numerator:

$$\sum_{j=i,\infty} B(j1-p_0)(i)p(j)=\sum_{j=i,\infty} c_{ji}(1-p_0)^i p_0^{j-i} q(1-q)^{j-1}=(1-q)^{i-1}(1-p_0)^i q\sum_{j=i,\infty} C_{ji}p_0^{j-i}(1-q)^{j-i}$$

Given $h=(1-q).p_0$ we have:

$$\begin{aligned}\sum_{j=i,\infty} C_{ji}p_0^{j-i}(1-q)^{j-i}&=\sum_{k=0,\infty} C_{k+i\,i}h^k=(1/i!)\sum_{k=0,\infty}(k+1)(k+2)\ldots(k+i)h^k,\\&=(1/i!)\sum_{k=0,\infty}[h^{i+k}]^i([\ ]^i\text{:derivative of order } i)\ h \text{ being less than 1 we have:}\\&=(1/i!).[(h^i/(1-h)]^i\end{aligned}$$

We can show that: $[(h^i/(1-h)]^i=(i!/(1-h)^{i+1})$,therefore:

$$\Sigma B(j1-p_0)(i)p(j)=(1-q)^{i-1}(1-p_0)^i q/(1-p_0(1-q))^{i+1}$$

So we can write:

$$PP(i)=q(1-q)^{i-1}(1-p_0)^{i-1}/(1-p_0(1-q))^i\ i=1,2\ldots$$

If q is replaced by $1/m$ we have:

$$PP(i)=(m(1-p_0)+p_0-1)^{i-1}/(m(1-p_0)+p_o)^i\ i=1,2\ldots$$

Therefore, PP is a geometrical distribution of expectation $m(1-p_0)+p_0$.

4. COMMENT

When the number of articles is fixed, all the articles have the same probability $(1-p_0)$ of being requested at least once (that is a consequence of the hypothesis of a binomial law). In this model, we have introduced a distribution of probability which will quantify the number of

articles in each volume.

Consequently, given g, a geometrical distribution taking values 1, 2, 3,... of expectation M, the number of possible values for m (expectation of the geometrical distribution p) and of values for p_0 (proportion of 'no use') which verify the equation is infinite.

$$g(i)=\frac{\sum_{j=i,\infty} B(j1-p_0)(i)p(j)}{1-\sum_{j=1,\infty}(p_0)j\cdot p(j)}$$

where $i=1,2...,\ i<=j\ j=1,\ 2,\ ...$

We therefore have the relationship: $m=(M-p_0)/(1-p_0)$, which allows us to calculate p_0 if we know the value of m.*The case with 'No Use'*

If we suppose p is a geometrical distribution in Equation (0), we can easily demonstrate that P is not a geometrical distribution. We extend to the limit to get the same result.

Theorem 2. Let p be a geometrical distribution of expectation m and p_0 a number between 0 and 1. Then the distribution defined by:

$$H(0)=\sum_{j=1,\infty} B(j1-p_0)(0)p(j) \tag{2}$$

$$H(i)=\Sigma B(j1-p_0)(i)p(j)\ i=1,2...\ j=i,\infty$$

converges in law towards a geometrical distribution of expectation M; when p_0 tends towards 1, m towards infinity, in such a way that $m(1-p_0)$ converges towards a finite limit denoted M.

Let us note that $m(1-p_0)$ is the expectation of the distribution H.

Proof. According to the calculations of the previous theorem we can write: $H(0)=qp_0/(1-(1-q)p_0)=p_0/(m(1-p_0)+p_0)$

For $i>=1$ we have: $H(i)=(1-q)^{i-1}(1-p_0)^i q.(1/(1-p_0(1-q))^{i+1} i=1,2.$

If q is replaced by its value: $H(i)=(1-p_0)^i(1/m)((m-1)/m)^{i-1}(m/(m(1-p_0)+p_0)^{i+1}$

$H(i)=(m(1-p_0)+p_0-1)^i(m/(m-1))/(m(1-p_0)+p_0)^{i+1}$

We note: $M=$limit $m(1-p_0)$, $p_0\rightarrow 1$, $m\rightarrow\infty$.

If we extend to the limit we obtainlimit $(H(i)=M^i(M+1)^{i+1}\ i=0,1,2...,\ p_0\rightarrow 1,\ m\rightarrow\infty$.

Therefore H is a geometrical distribution of expectation M.

5. COMMENT

Such conditions of extension to the limit are very frequent in probability such as the well-known approximation of a Poisson's law by a binomial law.

It must be remembered that an article used i times is counted just once. This result can be compared to that of Egghe (1994) where he defines the distribution F by:

$$F(i)=\sum_{j=1,\infty} P_j(i)p(j)\ i\geq 0 \tag{3}$$

$P_j(i)$: probability that a volume containing j articles be used i times, where P_j is the convolution product of j geometrical continuous distributions, and p is a geometrical continuous distribution.

$(p(i)=pq^i\ i\geq 0)$. In this case he has supposed that an article used i times is counted i times. So he has supposed that the distribution of the use of articles is of a geometrical nature. Egghe therefore shows that F is a geometrical distribution.

6. CONCLUSIONS

Bradford's law says that the repartition of articles of periodics dealing with a specific area is of a geometrical kind. In our study, it is when we suppose that the quantification of the number of articles in a periodical is geometrical that we get (Theorem 1) a geometrical distribution of the circulation.

The passage to the limit in Theorem 2 corresponds to a more complex situation $p_0 \rightarrow 1$. This means that the majority of articles are not used, which is associated for the passage to the limit to the fact that $m \rightarrow \infty$. Therefore the number of accessible articles is increasing, and consequently the number of unrequested articles is increasing (with the mean value of the distribution which has a finite limit). This increasing phenomena becomes more and more frequent in actual information systems.

We get a result which could seem quite surprising: it is when the weight of 'no use' becomes important that we again find a geometric distribution (when we had partially eliminated its influence in Theorem 1, where we only took into account the articles not used in requested volumes).

Note: all the distributions of circulation cannot be built from the model we defined. (cf. Equation (0)). By example, if we try to apply it to the circulation of books in a library, we will have to define a grouping criteria, which could be the area, or the acquisition date, etc., but these grouping criteria don't mean the same thing as grouping articles in a periodical.

REFERENCES

Burrel, Q. L. (1986). Library circulation distributions: some observations on the PLR sample. *J. Documentation*, *42*(1), 22–45.

Burrel, Q. L., & Fenton, M. R. (1993). Yes the Gigp really does work and is workable. *J. Amer. Soc. Information Sci.*, *44*(2), 61–69.

Egghe, L. (1990). The duality of Informetric systems with application to the empirical laws. *J. Information Sci.*, *16*, 17–27.

Egghe, L. (1994). Special features of the author publication relationship and a new explanation of Lotka's law based on convolution theory. *J. Amer. Soc. Information Sci.*, *45*(6), 422–427.

Feller, W. (1968). *An introduction to probability theory and its applications*. New York: John Willey & Sons (3rd edn).

Haitun, S. D. (1982). Stationary scientometrics distributions, Part 1: Different approximations. *Scientometrics*, *4*(1), 5–25.

Haitun, S. D. (1982). Stationary scientometrics distributions, Part 2: Non Gaussian nature of scientific activities. *Scientometrics*, *4*(2), 89–104.

Haitun, S. D. (1983). Stationary scientometrics distributions, Part 3: Role of the Zipf distribution. *Scientometrics*, *4*(3), 181–194.

Lafouge, T. (1995). Stochastic information field. *Int. J. Scientometrics Informetrics*, *1*(2), 57–64.

Lafouge, T., & Delarbre, A. (1989). Des statistiques à la bibliométrie. *Revue Française de Bibliométrie*, *4*, 179–190.

Leemans, M. J., Maes, M., Rousseau, R., & Ruts, C. (1992). The negative binomial distribution for circulation data in flemish public libraries. *Scientometrics*, *2*, 47–57.

Price, S. D. (1976). A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Information Sci.*, *39*(4), 292–306.

Sichel, H. (1985). A bibliometric distribution which really works. *J. Amer. Soc. Information Sci.*, *36*(5), 314–321.