Bibliometric Models for Management of an Information Store. I. Differential Utility among Items

Ralph H. Parker
School of Library and Informational Science, University of Missouri-Columbia, Columbia, MO 65211

Differential demand for use among the items in an information store is a necessary condition for management of the store. Using bibliometric techniques for determination of the distribution of demand, the hypothesis of hyperbolic distribution, and an index of differential demand are developed.

Introduction

Limitation of the size of a working information store through management techniques is of interest to administrators of many types of information storage and retrieval systems. The economic pressures of storage are particularly great in situations where storage costs are relatively high.

Solution of the problems of collection management is based ultimately on the assumption that some items in the store are more useful than others, and that utility can be measured by demand (i.e., probability of use). Thus, in matters of selection and acquisition, of storage and arrangement, and of retention or disposal, the utility of each item is the most important factor. The same concerns with differential utility apply whether the system is a library, a book store, or a computer-based information service; whether the item is a book or some other information-bearing object such as an index entry.

A corollary assumption, that utility is a function of age, complicates the management problem. We shall consider that subset of problems in another article.

Until recently, little had been done to identify or analyze the specifics of differential utility of items in a collection. Chen [1] has suggested that a collection of information sources may be divided into those which are living and those which are dead, and that there are significant differences between the two. This appears to be a distinction of convenience rather than an accurate description of reality. One may more logically assume that the variation in probability of use of items is essentially continuous from the highest probability to the lowest, approaching but never reaching zero as a limit. This assumption is made throughout this study.

Morse [2], Chen [1], and others have intimated, but not explicitly stated, that frequency distribution of living titles by rate of use is exponential; Booth [3] and others have suggested that the distribution is according to Bradford's [4] law of scattering. There have been, within the last few years, a number of studies using mathematical and statistical techniques for optimization of retention and disposal policies or of storage methods. Trueswell [5] has proposed methods for selecting low-utility items for disposal and similar proposals have been made by Ash [6] and by Fussler and Simon [7].

The Setting and Methodology

For reasons which will be developed in this article, observation alone is inadequate in determining the extent of differential utility of items in an information store. This article undertakes the determination through setting up hypothetical distributions and testing them against empirical data collected for a study of storage of and ready access to library materials by the four campuses of the University of Missouri [8].

This study utilized a sample of recorded uses (circulations) of material from the Elmer Ellis (Main) Library of the University of Missouri-Columbia occurring between August 1, 1972 and August 1, 1973, representing most of the recorded use of Ellis Library for the year. Because of technical problems some records were eliminated, leaving 385,989 usable records. The number of titles used and the total number of uses are plotted for the entire year in Figure 1. After 10,000 uses, for example, approximately 8,500 different titles had been used; after 300,000 uses approximately 118,000 titles had been used.

Received May 1980; revised April 1981; accepted November 13, 1981

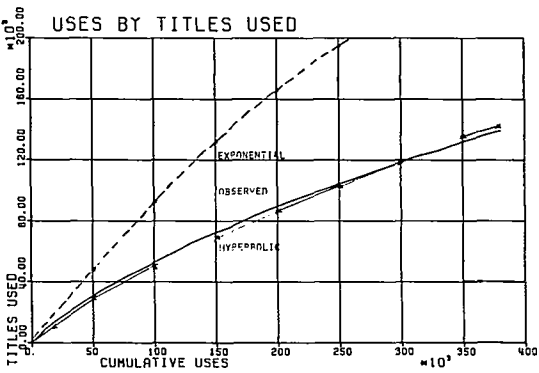

FIGURE 1. Cumulative titles used by cumulative circulation.

Differential Probabilities of Use

The first condition necessary for effective management of the size of an information store is that there be significant differences between the rates of demand for the items in the store. This study, and many other studies of this type, assume that the use of an item is a random process.

One characteristic of random occurrences is that they tend to follow a Poisson distribution. If there are on the average two users arriving per unit of time, there will be some units in which two users arrive, some in which more than two arrive, some in which only one arrives, and some in which none arrives. The equation of this distribution is

$$P_n = \lambda^n e^{-\lambda}/n!, \tag{1}$$

where P_n is the probability of n arrivals in one unit of time, λ is the average number of arrivals (demand rate) per unit of time, n is the number of arrivals, and e is 2.71728 (the exponential).

For a book with a demand rate of one use per year ($\lambda = 1$), the probability of exactly one use (the value of P_n) becomes 0.367, or 37%; the same probability exists that there were *no* uses observed during the year. While a book with a demand rate of 1 ($\lambda = 1$) will probably be used *once* 37% of the time, one with twice the rate ($\lambda = 2$) will be used once 27% of the time, and one with a rate of 3 ($\lambda = 3$) will be used once 15% of the time. Thus, it is impossible to determine (by observation alone) the demand rate of an item.

Before attempting to validate any hypothesis of differential demand, it is necessary to examine the distribution of use when the demand rate of all items is equal. Let us examine the operation of a system consisting of N distinct information sources, which will be called titles (T). In the system it is assumed that (1) uses occur sequentially, and (2) each request results in a use: there is no circulation interference, i.e., no user interferes with any subsequent request.

The probability that the next request will be for a specific title is $1/N$. The probability (P_v) that the next title selected will be the first use of that title will be $1/N$ times the number of titles not yet used (R_u). The mathematical equation for this relationship becomes

$$P_v = R_u/N, \tag{2}$$

where N is the number of items in the store, R_u is the number of items remaining unused after U uses, and P_v is the probability that the next use after U uses will be of a previously unused item. Equation (2) may, by transposition, be written

$$R_u = NP. \tag{3}$$

The number of items used after N uses becomes

$$T_u = N - R_u = N - NP. \tag{4}$$

As the value of U increases, the number of items used (T_u) increases and the remaining unused items (R_u) declines, the probability of selecting an unused item declines with continued use. The probability for any use may be expressed as

$$P_v = \left(\frac{N-1}{N}\right)^u. \tag{5}$$

In a store of 1,000 times ($N = 1{,}000$), for example, after 1,000 uses, 368 items will not have been used, 632 titles will have been used once or more, and some items will have been used up to five times despite the equality of demand. The conclusion from this discussion is that even though every item in an information store has an equal probability of use, there will be an unequal distribution of observed use unless the sample is exceedingly large.

Under conditions of equal probability of use, the number of items used is dependent on the values of N (size of the store) and of U (number of observed uses); under conditions of unequal probability of use, the expected number of items used in a sample of a given size is additionally a function of (1) differential rates of demand, and (2) the frequency distribution of items within the range of rates of demand. The demand for an item in an information store can be conceived in terms of the mean time between expected uses of the item: a few hours, a few days, a year, or many years; this may be converted into a rate (λ) by relating it to the reciprocal of the interuse interval. Thus, if the greatest interuse interval is taken as the unit, the rates will vary from one to the inverse of the shortest interval.

Hypothetical Distributions

As indicated earlier, it is difficult to determine empirically the rate of demand for an item, the range of probability, or the frequency distribution of items within the range. Because of these difficulties, distribution of the collection by rate of probable use can best be determined by creating and testing hypothetical models against empirical data. A number of possible models may be considered.

That titles are distributed:

(1) uniformly throughout the range; i.e., number with

λ (demand) equal to 1 is the same with λ equal to 1,000, or any other value;

(2) essentially normally; i.e., the greatest number of titles is concentrated around the mean value of λ, with smaller numbers as the extremes are approached;

(3) with the greatest number of the lowest value and decreasing linearly as λ increases;

(4) exponentially according to the value of λ such that the proportion for any rate can be expressed by

$$P_\lambda = ae^{-a\lambda}, \tag{6}$$

where a is a constant less than 1, characteristic of the information store;

(5) hyperbolically, such that the frequency of occurrence (f) of items of any demand rate (λ) is inversely related to the rate (λ); stated in another way the yield (i.e., the product of frequency and rate) for any value of λ is a constant $K(f\lambda = K)$.

It is demonstrated easily that distributions (1)-(3) do not conform to experimental results and are not considered further in this article.

To arrive at a convenient method for computation in testing these hypotheses, it is assumed that the contents of an information store can be segmented into an appropriate number (g) of groups or subsets, according to demand rate, such that each group will yield equal use, and that items within each group may be presumed to have identical demand, the mean value of λ for the group. We may now proceed to compute for each segment the expected number of titles that would be used in any specified number of uses by applying the formulas for equal probability, eqs. (4) and (5).

Equation (4) now becomes for each segment

$$T_{u(i)} = N_i - N_i P_{v(i)} \tag{7}$$

and eq. (5) becomes for each segment

$$P_{v(i)} = \left(\frac{N_i - 1}{N_i}\right)^{u/g}. \tag{8}$$

The total number of titles used in any use sample is obtained by summation of the segments:

$$T_u = \sum_{i=1}^{g} (N_i - N_i P_{v(i)}). \tag{9}$$

Let us first consider the hypothesis that the frequency distribution of items by demand rate (λ) is exponential; eq. (6),

$$P_\lambda = ae^{-a\lambda},$$

yields the probability of occurrence (frequency) of titles with a demand of λ. The expected circulation (C) for titles of any rate of demand is the product of rate and frequency (λP_λ). The equation for circulation then becomes

$$C_\lambda = \lambda(ae^{-a\lambda}). \tag{10}$$

The area under the curve, hence the total circulation, is the integral of eq. (10):

$$C = \int_{\infty}^{0} \lambda ae^{-a\lambda} = \frac{a(e^{-a\lambda})\,(a\lambda + 1)}{a^2} = \frac{a\lambda + 1}{ae^{a\lambda}}. \tag{11}$$

When the information store is divided into segments of equal yield the proportion of titles in each segment may be obtained by a two-step process: (1) by evaluation of the definite integrals, the boundaries of each zone (values of λ) may be obtained; (2) when these values are applied to eq. (10), the probable frequency of occurrence of items in each segment is obtained. When divided into five segments of equal yield, the one with the highest demand consists of 0.05 (5%) of the items; the least productive segment consists of 0.55 (55%) of the items (Table 1).

When the estimated number of items used from all segments, using eq. (9) for use samples from 10,000 to 380,000 (corresponding to the empirical data against which the results are to be tested), it was found that the estimates far exceeded the observation. At $U = 10{,}000$, $T_u = 9{,}920$ as compared with the observed 8,446. The disparity increases with larger-use samples. The comparison is shown graphically in Figure 1.

The divergence is so great that the hypothesis of exponential frequency distribution of titles by demand rate must be rejected.

Hyperbolic Distribution

We shall next consider the hypothesis that the distribution of titles by demand rate is hyperbolic. Two classic statements of hyperbolic distributions have strongly influenced bibliometrics. The first was by Zipf [9], who in 1934 pronounced his law of frequency in the use of words. The law states that if the words used in a universe of discourse are ranked according to the frequency of use, the product of rank (r) and frequency (f) is a constant ($rf = c$). This he pointed out is the equation of a rectangular hyperbola. There have been, since 1960, numerous attempts to interpret and improve the Zipf equation and to apply it to other bibliometric phenomena.

The second statement of the hyperbolic distribution was by Bradford [4], who, also in 1934, propounded his law of scattering. It stated that if the "journals publishing

TABLE 1. Proportion of titles by segment of equal yield for selected distributions of titles.

	Distribution Hyperbolic			
Segment	$K = 82$	$K = 12{,}365$	$K = 1{,}835{,}000$	Exponential
1	0.012	0.0003	0.00005	0.05
2	0.036	0.0021	0.00012	0.086
3	0.085	0.0157	0.00236	0.129
4	0.232	0.1137	0.04731	0.183
5	0.634	0.8645	0.95021	0.552

articles in a subject are ranked according to the number of articles published, and then are divided into zones of equal yield (number of articles) the number of journals in each zone will vary by the ratios $1:N:N^2:N^3$." If the rank of journals is plotted on the X axis, using a logarithmic scale, and the cumulative yield of articles on the Y axis using the arithmetic scale, the resulting curve (called a bibliograph) will, after an initial upward curve, become linear.

Numerous articles interpreting and modifying the Bradford statement have been published, including one by Brookes [10] which states that the slope of the linear portion of the curve (when using natural logarithms) indicates the total number of articles in the universe. Numerous other articles have discussed the relation of the Zipf and Bradford distributions.

This author has undertaken a somewhat different approach in testing the hypothesis of hyperbolic distribution of titles in an information store, applying but modifying both Zipf's and Bradford's. The first modification is that if the number (f) of items (or titles), with probable rate of use of λ or greater, rather than the rank of an item, is multiplied by λ, the product is a constant K ($f\lambda = K$). Since any number of items can conceptually have the same rate, one logical problem faced by Zipf, when there are numerous items with one use, is obviated. If $K = 100$, $f\lambda = K = 100$; when f is set at 1, $\lambda = 100$; when $f = 100$, $\lambda = 1$; when $f = 2$, $\lambda = 50$; etc. The constant K, therefore, indicates both the maximum probable demand rate, and the number of items with the rate of 1 or more.

To simplify the computations involved in applying formulas, let us introduce the concept of a standardized hyperbolic unit, which consists of a set of K items with probable rates of use from 1, the minimum by definition, to the maximum K. An information store may consist of one or more standardized hyperbolic units.

The second modification in approach is to replace Bradford's use of the logarithmic scale by the harmonic scale. Although Bradford did not characterize his distribution as hyperbolic, others have done so by noting the similarity to the Zipf distribution.

An essential feature of the hyperbolic distribution is that the two factors (X, Y) are reciprocals; if one is linear, the other forms a harmonic series (H_s). The sum of the harmonic series ($1 + 1/2 - 1/3 + 1/4 + \cdots + 1/N$) can be approximated by the algorithm

$$H_s = 0.5772156649 + \log_e N + \frac{1}{2N} - \frac{1}{12N^2} + \frac{1}{120N^4}. \quad (12)$$

The cumulative harmonic values can be used on a scale for representing cumulative frequency of titles replacing Bradford's use of the logarithmic scale.

If, following Bradford, the number of items in a standardized hyperbolic unit at each level of demand (λ) is plotted cumulatively on the X axis using the cumulative harmonic scale, and the use of these items is plotted cumulatively on the arithmetic Y axis, the resultant curve will be a straight line (Fig. 2). When the frequency distribution of items by rate of probable use is hyperbolic, a series of points on the X axis which divides the harmonic scale into equal segments, if projected vertically to intersect with the curve and then horizontally to the Y axis, will divide total use into equal segments and the number of titles harmonically. The slope of this line will equal the constant K of the basic hyperbolic equation $XY = K$.

Unlike the exponential distribution, which yields the same proportion of titles in segments regardless of changes in the value of a (in the equation $Y = ae^{-ax}$), with a hyperbolically distributed demand rate, the proportions of items in the various segments vary with changes in the value of K ($XY = K$). Table 1 shows the distribution between five segments of equal yield when $K = 82$ ($H_s = 5$), $K = 12{,}365$ ($H_s = 10$), and $K = 1{,}835{,}000$ ($H_s = 15$). The proportion of titles in the most productive segment declines from 0.012 to 0.0003 to 0.00005 ($K = 1{,}835{,}000$). The magnitude of the variation of probability of use among items may be expressed by the index of differential demand which has the value of the constant K (in the equation $f\lambda = K$) applicable to the particular information store.

To test the hypothesis of hyperbolic distribution, computations using eqs. (6), (7), and (8) were used for a number of values of K. For each value, the number of titles expected to be used in each of 10 segments for each of 38 use levels (from 10,000 to 380,000) were computed. The segment results were then summed for each use level and were compared with the observed data. The best fit was found to be $K = 25{,}000$. The results of these computations are shown graphically in Figure 1, comparing the results with the observed data as well as with the exponential hypothesis. At 10,000 uses the hyperbolic prediction is almost 20% below observation, but rapidly approaches observation as the use samples increase in size. At about 250,000 uses the difference is insignificant, and disappears by 300,000 uses.

This hyperbolic model omits two factors present in the empirical sample, so that some deviation is expected. First, the model equates arrival of a user requesting a title with use of that title. We know that, especially in books

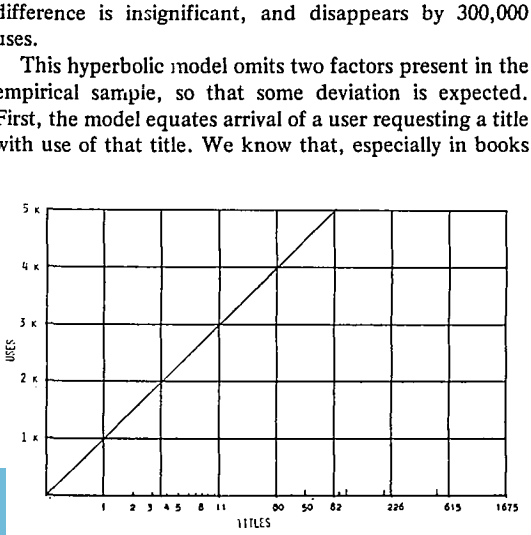

FIGURE 2. Hyperbolic yield.

with high rates of use, there is circulation interference, which results in lost circulation in the more productive segments, consequently a higher representation of titles from the lesser used segments, thus the excess of observed use over prediction is expected. As the use samples increase in size and *all* titles in the middle segments have been used, the effect of circulation interference becomes less pronounced, and the difference between observation and prediction disappears.

Second, the model assumes a static information store, when in fact it is ever-changing. New items are being added and there are losses from attrition and obsolescence; the effect will be discussed in a subsequent article.

Despite the variations between theoretical and observed uses of titles, there is close enough correspondence to consider seriously the hypothesis of hyperbolic distribution of titles by rate of expected use.

Conclusion

Consideration and testing of a number of hypothetical frequency distributions of titles in an information store by demand rate, i.e., probability of use, indicates that the best conformity to empirical data is with the hyperbolic hypothesis. The relationship between frequency of occurrence of titles and the demand is such that the yield (expected total use) of all titles with a specific demand rate will equal the yield of all titles with any other specific demand rate.

The study suggests that the differential rates of demand can be described by an index of differential demand which has the value K in the equation $f\lambda = K$. The larger the value of K, the greater the concentration of use in a small fraction of titles in the information store.

The results also suggest that the value of K, at least in a large general university library, is great enough that in matters of selection and acquisition, of retention or disposal, effective resource management is possible.

References

1. Chen, Ching-Chih. *Applications of Operations Research Models to Libraries.* Cambridge, MA: MIT Press; 1976: see p. 33.
2. Morse, Philip M. *Library Effectiveness.* Cambridge, MA: MIT Press; 1968: see particularly Chap. 4, "Queues and Book Circulation Interference," pp. 54-82.
3. Booth, A. S. "On the Geometry of Libraries." *Journal of Documentation.* 25(1):18-22; 1969.
4. Bradford, Samuel C. "Sources of Information." *Engineering.* 12.. 85ff; 1934; restated in his *Documentation.* London: Crosby Lockwood; 1948.
5. Trueswell, Richard W. "A Quantitative Measure of User Circulation Requirements and Its Possible Effect on Stack Thinning and Multiple Copy Determination." *American Documentation.* 16(1): 20-25; 1965.
6. Ash, Lee. *Yale's Selective Book Retirement Program.* New Haven, CT: Archon Books; 1963.
7. Fussler, Herman; Simon, Julian. *Patterns in the Use of Books in Large Research Libraries.* Chicago; Univ. Chicago P.; 1969.
8. Parker, Ralph H. *A Stochastic Analysis of Books Circulated from Elmer Ellis Library, 1972-1973.* Columbia, MO: University of Missouri; 1974:37-45.
9. Zipf, George K. *Psychobiology of Language.* Boston: Houghton Mifflin; 1935. Also *Human Behavior and the Principle of Least Effort.* New York: Addison-Wesley; 1949 (reprinted Hafner, 1972).
10. Brookes, B. C. "The Derivation and Application of the Bradford-Zipf Distribution." *Journal of Documentation.* 24(4):247-265; 1968.