Bibliometric Models for Management of an Information Store. III. Developing an Empirical Model

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

Based on earlier studies by the author relating to differential demand among items in an information store and to the relation of demand to age of material, this article undertakes to develop an empirical model for predicting the size of an information store necessary to satisfy specified levels of demand. A *modus operandi* for selecting items for retirement or removal with the least adverse impact on effectiveness of the store is suggested.

Control and limitation of the size of an information store depends on existence of necessary and sufficient conditions. In previous articles [1,2], this author has developed theoretical bibliometric models for each of two necessary conditions: (1) differential demand for individual items and (2) decline in demand with increasing age of materials. Although they establish the possibility of developing selection and retention programs which would limit the size of the store, or at least reduce the net rate of its growth, the models do not guarantee the sufficiency of conditions.

In this article, we shall attempt to consider this sufficiency and the potential for limiting the size of the information store using empirical rather than theoretical models.

In any process which requires a certain amount of time for completion and which has a constant rate of input, the number of items in process (N), according to the principles of queuing theory, will continue to increase until that number is equal to the product of the input rate (λ) times the average time in process (τ) : $N = \lambda\tau$. For example, if the decline in demand for an item is such that, on the average, it may be removed (discarded or put in storage) after 40 years, and if the rate of input is 50,000 items per year, the minimum size of the information store would be 50,000 times 40 or 2,000,000 items. At this point, the store will have achieved a steady state, and it may be considered to have maturity. For a mature information store to maintain a steady state at this or any other level, without loss of capacity to meet demand, infusion of new items must continue to offset the losses from obsolescence and eventual removal of old items.

The most important contributions to the literature in discarding lesser used materials include those by Ash [3], Fussler and Simon [4], and Trueswell [5]. Morse [6] and Chen [7] have also added to our understanding of the problems involved.

All these studies use records of prior use of individual items as a means of predicting future use, and of deciding upon retention or removal from the information store. They do not address directly the question of determining the magnitude of a well-selected collection to provide a specified level of success in satisfying demand. Models for this purpose and for indicating the effect of removing low-utility materials will be developed in this article.

Background and Methods

Data from a sample of 385,989 uses of the Ellis Library, the central library of the University of Missouri–Columbia [8], which were used in validation of theoretical models of use distribution and of obsolescence in previous articles, have been used to develop an empirical model for prediction of the number of items (titles) used in any number of circulations (uses), which in turn forms the basis for other predictions.

The number of titles used at intervals of 10,000 uses, as shown in Table 1, are cumulative totals; thus, after 20,000 uses 14,722 titles had been used once or more, of which 8,446 had been used during the first 10,000 uses, and 6,276 were first used during the second 10,000 uses. When the ratio of titles to uses (T/U) is plotted on the Y axis against use on the X axis, the resulting curve (Fig. 1) is rather smooth, with almost imperceptible irregularities

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

TABLE 1. Number of titles used at use intervals of 10,000.

Use[a]	Titles	Uses[a]	Titles
10	8,446	210	92,563
20	14,722	220	95,597
30	20,122	230	98,621
40	25,466	240	101,320
50	30,740	250	104,350
60	35,806	260	106,931
70	40,164	270	109,938
80	44,408	280	112,677
90	48,219	290	115,410
100	52,396	300	118,591
110	56,677	310	121,537
120	60,875	320	123,850
130	65,503	330	125,847
140	69,006	340	128,888
150	72,320	350	131,892
160	75,330	360	134,232
170	79,243	370	136,758
180	83,161	380	139,121
190	86,437	385(989)	139,892
200	89,479		

[a] Times 10^3.

occurring at the end of the fall, winter, and summer terms, the result of the required return and reissue of books.

Various families of equations, both algebraic and transcendental, were tested for possible fits of the curve. The general appearance of the curve, and the distribution of probability of use developed in ref. 1, suggest a rectangular hyperbola, the equation of which is $XY = K$. But this is impossible since Y (the ratio T/U) when multiplied by uses (X) equals T, which can be a constant only when all titles in the store have been used. However, if Y, which can never exceed 1, is raised to some power greater than 1, the value of the product XY can be a constant. This is the basic form of Lotka's law of distribution of scientific productivity [9]. Such a modification of the basic hyperbolic equation may be made to fit observed results except for low values of X. If the equation is further modified by adding a constant K to X, so that the equation becomes

$$(X + K)Y^{\theta} = K + 1, \qquad (1)$$

an acceptable fit may theoretically be obtained for all values of X. Dividing by $X + K$ and extracting the θ root.

$$Y = \left(\frac{K+1}{K+X}\right)^{1/\theta}.$$

The number of titles (T_u) included in a use sample (U) is YU; thus,

$$T_u = U\left(\frac{12,101}{12,100 + U}\right)^{0.28694}. \qquad (2)$$

An iterative computer program for determining the values of K which best fit the observed data was developed (it will be supplied without charge upon application to the author). Needed for input is the number of titles used at each of a number of levels of use. The program alters the values of K and θ until the sums of the squared deviations of predicted title usage from observed title usage can be reduced no further. From the experimental data the best fit occurred for $K = 12,100$ and $\theta = 3.485$. Equation (2) now becomes

$$T_u = U\left(\frac{12,101}{12,100 + U}\right)^{0.28694}. \qquad (3)$$

The goodness of fit is shown in Figure 2.

For use values below 10,000, computed values Y are slightly lower than observed values; for values of U from 20,000 to 100,000, the computed values of Y are slightly higher than those observed, and for values of U from 100,000 to 300,000, the computed values of Y are again slightly low; after 300,000 they cross again. These variations, which in any case are minor, appear to result from random variations and from the seasonal fluctuations referred to above, rather than from any defect of the model.

The significance of the values of the constant and of the exponent is not immediately definable in theoretical terms. But from application to other information stores, of similarly derived equations, it appears that the size of the exponent is related to the breadth and potential size of the information store, and the size of the constant to the range of probable rates of use. Further research may clarify the relationship.

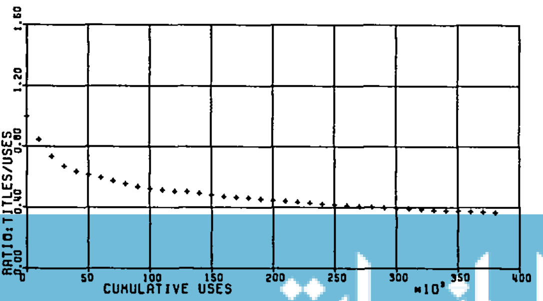

FIGURE 1. Ratio of titles used to total uses bv cumulative uses

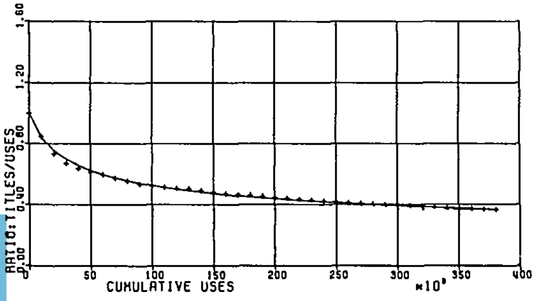

FIGURE 2. Goodness of fit of empirical mode

If, by applying eq. (2), use is extrapolated far enough, T_u can assume values beyond the current size of the information store, which is possible only if the size of the store is increasing sufficiently. In the situation in which eq. (2) was derived, new titles were being added at the rate of 50,000–60,000 per year. Assuming that the observations continue for ten years (about 4,000,000 uses), and the addition of new titles to the store also continues at a similar rate, the titles used after 4,000,000 uses would be projected as 756,507 (Fig. 3); only 52,872 previously unused titles will have been used during the tenth year. Since this is within the range of growth, it appears that the system can operate as if it has unlimited (although finite) size.

Concurrent with growth of the store is obsolescence, the reduction of probability of use of each title. As indicated in ref. 2, the reduction is two-factor negative exponential, with variations between subjects, forms of materials, and individual titles. It appears that the mean rate of decline in use for large broad-based collections over long periods (more than 50 years) is in the neighborhood of 0.045 (4.5%) per year. An infusion of about 55,000 new titles each year into a collection of 1,200,000 would be necessary to offset the decline through obsolescence and thus to maintain stable capacity.

If no intervention in this process of receding usefulness and of stock replenishment occurs, the index of differential demand (as defined in ref. 1) will continue to increase, and the proportion of the collection necessary to produce any specified part of the use (up to about 95%) will decline. Trueswell has suggested [5,10] that the most productive 20% of the collection produces between 60 and 70% of the use, and that between 25 and 30% of the collection can produce 80% of the use. The figures in Table 2 suggest that unless the collection is much more homogeneous (the index of differential demand is lower) than likely to be found in most large research libraries, less than 15% of the collection will be required to produce 80% of the use. The significance of this phenomenon is that relatively large numbers of the least used items can be removed without significant loss of capacity to serve.

TABLE 2. Percent of collection providing selected percentages of use (hyperbolic distribution of demand).

Percent of Use	Percent of Collection
10	0.004
20	0.016
30	0.052
40	0.16
50	0.5
60	1.4
70	4.
80	12.
90	34.
95	59.
99	90.
99.5	95.

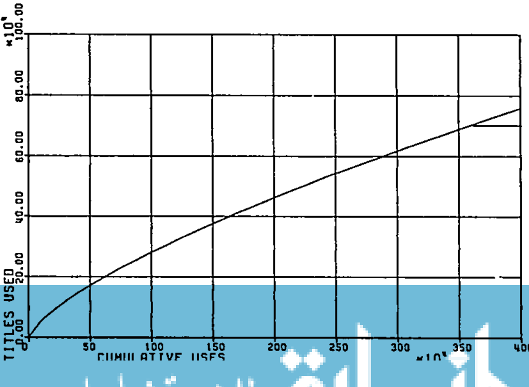

FIGURE 3. Expected titles used projected for ten years

Optimal Size of Information Store

From data thus far developed, we are not able to estimate the size of an optimally selected collection needed to provide for any level of adequacy. The breadth of subject coverage is the most important component; an engineering library need not be as large as an entire university library. From eq. (2) we may arrive at a practicable, though not optimal, basis for making such an estimate. Since selection of items for inclusion in the store would be dependent upon evidence of use, the collection would include items of low probability of use in proportion tc their actual use, along with those in high demand.

The adequacy of a collection may be considered as the probability that any request for an item can be fulfilled. This is the complement of the probability (P_v) that any subsequent request will be for an item not previously requested, hence not in the collection. This probability can be determined from the slope of the curve [eq. (1)] at any point:

$$P_v = \frac{T_{u(2)} - T_{u(1)}}{U_2 - U_1}. \qquad (4)$$

For example, if at $U = 40{,}000$, $T_u = 26{,}311$, and at $U = 50{,}000$, $T_u = 31{,}272$, the probability that, on any use between the 40,000th and 50,000th a new title will be used is

$$\frac{31{,}272 - 26{,}311}{50{,}000 - 40{,}000} \text{ or } \frac{4{,}961}{10{,}000}.$$

This is 4,961 new titles per 10,000 uses, which is equivalent to 0.4961 new titles per use. If computations are sufficiently accurate, the best value of the slope results from using an interval of one use. The equation now becomes

$$P_v = T_u - T_{u-1}. \qquad (5)$$

At 50,000 uses, $T_u = 31{,}272.4682$; at 49,999 uses, $T_u = 31{,}271.98725$. Subtracting we find that the rate is 0.48095, which is the value of P_v. The probability that a collection of 31,272 titles selected on the basis of observed use will include the next request of use is (1.0 − 0.480952 = 0.51015) 51.015%.

The adequacy of collections of selected sizes is shown in Table 3. A collection of 30,000 titles can, in the environment served by our experimental library, provide 51% of the requests for use; an increase to 50,000 titles provides only 9% more use, or approximately 60%. To increase adequacy another 10%, to 70%, requires that the collection be doubled to 100,000 titles; the next 10% increase, to 80%, 300,000 titles. To raise adequacy from 90 to 95% requires a fivefold increase in collection (8,000,000), and any further improvements require astronomical increases.

TABLE 3. Probability that any request can be filled from collection of selected sizes.[a]

Titles (*N*)	Adequacy (%)
30,000	51
50,000	60
100,000	70
300,000	80
500,000	84
800,000	87
1,600,000	90
4,000,000	93
8,000,000	95

[a]Based on empirically derived projections.

A *Modus Operandi*

After having ascertained the probable size of a mature information store required to provide the desired level of satisfaction, procedures must be developed for removing the least productive items to balance new acquisitions. Assuming that the distribution of items by demand (probability of use) is hyperbolic, the removal of the least productive 5% of the store will reduce use by less than 0.5%. A weeding procedure might, through failure to identify the *least* useful items, result in a loss of as much as 3% of the use. The procedure adopted must therefore be designed to hold this loss as near the theoretical minimum as possible.

The criterion for retention will probably be the expected future use of an item above some established threshold, let us say 0.05 uses per year (once in 20 years). Tests must be developed for identification of items to be retained or removed, and procedures must be based on these tests. Two factors to consider in setting up the tests and procedures are as follows.

Limits of Error of Decision

We need to recognize that identification of the items with least potential for future use can not be achieved without error. Two types of error may occur: Type I errors identify as *true* cases which are in fact *false*; thus an item might be selected for retention when its probable use is below the threshold. Type II errors identify as *false* cases which are in fact *true*; thus an item might be scheduled for discard when its probable use is above the threshold for retention. This type of error is most serious and must be guarded against.

Simplicity of Test

Since the operation of segregating items to be removed from those to be retained will be performed by persons of limited experience and training and under constraints of time to make the decision, the test to be applied to each item must be objective and easy to apply. Two types of test are most likely to be used:

(a) A record of prior use of the item during some observed period of time: this is the type used in most of the cited studies. This type of test is subject to great danger of type II error. Since occurrence of a use may be considered as a random process, these occurrences form a Poisson distribution, giving us a basis for estimating errors of decision. Let us consider as a test criterion the observed use of an item in the preceding year. An item with an expected average use of once per year will be observed to be used exactly once during the year only 37% of the time; more than once 26% or the time; and will *not* be used at all 37% of the time. An item with an expected use of once in ten years (twice the criterion rate) will have been used during the previous year less than 10% of the time, so that the test will fail more than 90% of the time (type II error).

Lengthening the observation period to the *five* prior years would improve the error rate. An item with an expected use of once per year will now be erroneously discarded only 7% of the time; but one with a rate of once in ten years will still be erroneously discarded 39% of the time. If the observation period is lengthened sufficiently to reduce type II errors to a satisfactory level, the validity of the test may be lost by the process of obsolescence.

(b) Application of criterion to all items in a statistically defined class: By analyzing use of all items in the store by identifiable characteristics such as subject, form of material, or language, with age the dependent variable, classes which on the average meet the test criterion may be defined. In using this type of test we must be cognizant that some items in the class will be above the criterion threshold, and that their removal will constitute type II errors. But knowing that the distribution of items by rate of use in the class is hyperbolic, we can assess the limits of danger from erroneous removal of the relatively few items above the threshold.

If we combine both types of test, first applying the statistical class test, then the observed prior use test to the selected classes only, both types of errors, but particularly type II, can be reduced to acceptable levels. Thus, if statistical analysis indicates that books in biology in the English language more than 25 years of age are each used on the average less than once in 20 years, this class would be considered eligible for discard. Classes with similar characteristics might be books in mathematics more than 50 years of age and in sociology in non-English languages more than 20 years old.

From studies of hyperbolic distribution of use within a class, we are fairly certain that two-thirds of the use will be produced by about 5% of the items (Table 2). Thus, if a significant portion of these items can be identified by the observed use test, we can eliminate most of the type II error inherent in the statistical class test.

If, for example, statistical analysis shows that the total use of classes with an average rate of use less than once in 20 years produces 3% of the total use, we know that this is the upper limit of the effect of discarding. Using a five-year observation period we can identify 65% of the items with rates equal to the average for the entire collection, and 99% of those with truly high probable use, say twice a year. By application of both tests, the loss from discard can be reduced to no more than 1%, or twice the theoretical optimum, which in most cases ought to be acceptable.

Conclusion

We have, in this and in the two preceding studies, examined from several perspectives the distribution of use among the items in an information store and have determined that:

(1) there is a differential probability of use among the items for which an index of differential demand can be computed;

(2) there is, in a statistical sense, a decline in use with increasing age of material;

(3) the variability of probability of use among the items in the store and the magnitude of the store are the most important, but not the only, factors determining the number of titles used in any use sample;

(4) it is possible to develop an empirical model which will predict satisfactorily the number of titles included in a use sample of any specified size;

(5) extrapolation from this model may be used to estimate the adequacy of a store of any specified size, within the environment of the original observations;

(6) it is not possible to generalize, from information developed, to other environments without repetition of the observations in that environment.

Results with usable accuracy may, however, be obtained with samples much smaller than used in this study, perhaps 20,000 observed uses would suffice. The results of the analyses undertaken suggest that, for large academic libraries serving a broad spectrum of knowledge, the differential probability of use between the lowest and highest may be expressed as a ratio of the general magnitude of 1:25,000; i.e., the most frequently used title will have a probability of use approximately 25,000 times that of the lowest.

By computation of the probability that any succeeding use, after any specified number of titles have been used, will be of a title already used, we can estimate the adequacy of a selected store of an arbitrarily chosen size. In the environment observed, a collection of 1,600,000 titles should provide 90% of all demands upon it; 4,000,000 titles would be required to provide 93%, 8,000,000 would be required for 95%, and any further improvement would require increases in collection size beyond practicability.

If addition of high-probability-of-use items to the collection is sufficient to balance the loss of productivity resulting from obsolescence, the removal of lowest-probability items equaling the loss by obsolescence will not reduce significantly the potential to fulfill requests. If additions are not equal to the loss by obsolescence, the adequacy of the collection will decline regardless of removal; and if additions are in excess of essential replacement to balance loss of obsolescence, the collection will grow regardless of removal.

References

1. Parker, Ralph H. "Bibliometric Models for Management of an Information Store. I. Differential Utility Among Items." *Journal of the American Society for Information Science.* 33: 124–128; 1982.
2. Parker, Ralph H. "Bibliometric Models for Management of an Information Store. II. Use as a Function of Age of Material." *Journal of the American Society for Information Science.* 33: 129–133; 1982.
3. Ash, Lee. *Yale's Selective Book Retirement Program.* New Haven, CT: Archon Books: 1963.
4. Fussler, Herman; Simon, Julian. *Patterns in the Use of Books in Large Research Libraries.* Chicago: Univ. Chicago P.; 1969.
5. Trueswell, Richard W. "A Quantitative Measure of User Circulation Requirements and Its Possible Effect on Stack Thinning and Multiple Copy Determinism." *American Documentation.* 16(1): 20–25; 1965.
6. Morse, Philip M. *Library Effectiveness.* Cambridge, MA: MIT Press; 1968.
7. Chen, Ching-Chih. *Applications of Operations Research Models to Libraries.* Cambridge, MA: MIT Press; 1976.
8. Parker, Ralph H. *A Stochastic Analysis of Books Circulated from Elmer Ellis Library, 1972–1973.* Columbia, MO: University of Missouri; 1974: pp. 37–45.
9. Lotka, Alfred J. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Science.* 16(2): 317–323; 1926.
10. Trueswell, Richard W.; Turner, Stephen J. "Simulating Circulation-Use Characteristic Curves Using Circulation Data." *Journal of the American Society for Information Science.* 30(2): 83–87; 1979.