Information Processing & Management, Vol. 13, pp. 79–93. Pergamon Press 1977. Printed in Great Britain

OPERATIONAL ANALYSIS OF LIBRARY SYSTEMS†

FERDINAND F. LEIMKUHLER
Purdue University, West Lafayette, IN 47907, U.S.A.

(*Received* 12 *July* 1976)

Abstract—Mathematical models are developed for describing and optimizing the way libraries organize information sources. By ordering the sources in a more relevant manner, the users' search and retrieval costs are reduced, which leads to an increase in the value and amount of information processed. The cost of organizing collections so as to minimize user effort is dependent on the scatter of relevant sources in the literature and the specification of core classes.

Libraries perform the useful function of collecting and organizing information sources for subsequent search and retrieval by users. The benefit of this service can be readily seen by considering the scatter of relevant information in the stock of information sources and the consequent cost of retrieving pertinent items of information. By grouping the sources into subject classes according to their potential productivity in an optimal way, libraries decrease user cost and increase the level and value of information usage. The benefits should more than offset the cost of organizing collections so as to reduce user effort.

1. INFORMATION SCATTER

The seminal work on information scatter was done by the British librarian BRADFORD[1] in the 1930s. Bradford counted the number of articles in each of all the journals he could find with at least one relevant article on a given topic of interest. This enabled him to rank order the journals in descending order of their productivity, and to divide such a collection into successive zones of equal total productivity. He found that the number of journals in the successive zones tended to increase by a common constant ratio. He believed that this relationship defined a general law of scatter, which has come to be known as Bradford's Law.

> If scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject, and several groups or zones containing the same number of articles as the nucleus, where the numbers of periodicals in the nucleus and succeeding zones will be as $1 : n : n^2 \ldots$ ([1], p. 154).

In reporting his results, Bradford plotted the cumulative number of articles as a linear logarithmic function of journal rank, i.e.

$$R_n = A \log B_n \tag{1}$$

where R_n denotes the cumulative number of articles, or references, in the n most productive journals and A and B are parameters peculiar to specific topics.

VICKERY[2] was the first to point out the inconsistency between eqn (1) and the verbal statement of Bradford's Law. Subsequent studies by LEIMKUHLER[4, 5], BROOKES[3] and WILKINSON[6] have elaborated on this difference. They have shown that the verbal statement leads to the alternative formulation:

$$R_n = A \log (1 + Bn). \tag{2}$$

BROOKES[3] has shown that formulation (1) is a form of Zipf's Law and has called the formulation

†This work was supported in part by NSF Grant SIS-14772.

the Bradford–Zipf model. More recently, MORSE[7] has shown that if the productivity of journals follows a geometric pattern, the scatter over the most productive 80% of the collection is approximately defined by eqn (1). In two more recent papers, BOOKSTEIN[8, 9] argues that formulations (1) and (2) represent special cases of a family of scatter formulas developed from a generalization of Lotka's Law of author productivity.

2. FORMULATION OF THE BRADFORD MODEL

Let a collection of journals be ranked according to the number of pertinent articles they contain on a subject, and then be divided into zones of equal total productivity. Let n_j denote the number of journals in the first j zones, and the rank position of the last journal in each zone. Then according to Bradford's Law, the successive zone sizes, $n_j - n_{j-1}$, increase by some constant factor k, such that:

$$\begin{aligned} n_2 - n_1 &= kn_1 \\ n_3 - n_2 &= k(n_2 - n_1) \\ &\vdots \\ n_j - n_{j-1} &= k(n_{j-1} - n_{j-2}) \end{aligned}$$

and, by induction,

$$\begin{aligned} n_2 &= (1 + k)n_1 \\ n_3 &= (1 + k + k^2)n_2 \\ &\vdots \\ n_j &= (1 + k + \cdots + k^{j-1})n_1. \end{aligned} \tag{3}$$

Equation (3) can be rewritten in the geometric form:

$$n_j/n_1 = (k^j - 1)/(k - 1) \tag{4}$$

which implies that

$$j = \log [1 + (k - 1)n_j/n_1]/\log k \tag{5}$$

when eqn (4) is inverted.

Let R_{n_j} denote the total articles or references contained in the n_j most productive articles, and also the total references in the first j zones. Because each zone has the same number of references, then

$$\begin{aligned} R_{n_j} &= jR_{n_1} \\ R_{n_j} &= R_{n_1} \log [1 + (k - 1)n_j/n_1]/\log k \end{aligned} \tag{6}$$

from eqn (5). Equation (6) defines the accumulated references in the $n_1, n_2, \ldots$ most productive journals, as proposed by Bradford. This relationship is readily generalized to all journals by letting R_n denote the total references in the n most productive journals and

$$R_n = R_{n_1} \log [1 + (k - 1)n/n_1]/\log k. \tag{7}$$

Equation (7) implies eqn (6) and both depend on n_1, the size of the first zone.

Since the choice of the first zone is arbitrary, it is convenient to define a standardized model in terms of the single most productive journal, as follows:

$$R_n = R_1 \log [1 + (k_1 - 1)n]/\log k_1 \tag{8}$$

where R_1 denotes the number of references in the most productive journal, and k_1 defines the minimum number of additional journals needed to obtain R_1 additional references. Equation (8) implies eqn (6), since if a collection with the pattern of (8) is divided into equally productive zones, then the total references in the first j zones is the same as (6) where

$$k = 1 + (k_1 - 1)n \tag{9}$$

defines the change in the Bradford constant.

Equation (8) is an idealized model of a collection of journals obeying Bradford's law of scatter. In order to distinguish such *models* of collections from the collections themselves, it is useful to use different symbols for the parameters R_1 and k_1, such as B for R_1 and $b = k_1 - 1$, whereby eqn (8) becomes:

$$R_n = B \log (1 + bn)/\log (1 + b). \tag{10}$$

Note that B is a *scale* parameter and b is a *dispersion* parameter.

If two collections have the same b value and different B values, then their respective productivity measures are directly proportional to the ratio of their B values. If two collections have the same B or R_1 values but differ in b, then the number of additional journals needed to get R_1 more articles is proportional to b, since

$$n_2 = b + 2. \tag{11}$$

In general, collections with high b values are more compact, i.e. articles are less evenly dispersed or scattered in different journals.

3. ESTIMATION OF THE DISPERSION PARAMETER

Estimates of the parameters B and b for a model of a particular reference collection might be made by merely counting the number of articles in the first few most important journals, and letting B and b equal R_1 and $n_2 - 2$, respectively. In general, n_2 is not likely to be an integer; but it could be estimated by determining the number of journals, J, needed to accumulate *at least* $2R_1$ articles, and then letting

$$n_2 = J - (R_J - 2R_1)/(R_J - R_{J-1}) \tag{12}$$

where R_J is the total number of articles in the first J journals such that the largest integer in R_J/R_1 is 2.

Another easy way to estimate b is to follow Bradford's approach and divide a ranked collection into zones of equal productivity. If n_1 and n_2 denote the rank positions of the last journals in the first two zones, then

$$\begin{aligned} R(n_2)/R(n_1) = 2 &= \log (1 + bn_2)/\log (1 + bn_1) \\ 1 + bn_2 &= (1 + bn_1)^2 \\ b &= (n_2 - 2n_1)/n_1^2. \end{aligned} \tag{13}$$

Since a collection can be divided into an arbitrary number of zones, it is likely that different b estimates could be obtained for the same collection when using the above method.

Bradford thought that if ranked collections were divided into three zones the ratio of the number of journals in the first and second zones would be constant for all collections of a similar type. However, by rearranging eqn (13), Bradford's ratio is equal to:

$$(n_2 - n_1)/n_1 = bn_1 + 1 \tag{14}$$

and therefore depends on both the number of journals in the first zone and the parameter b. Comparison of this ratio value for different collections would only be meaningful if the number of

journals in the first zone of each collection is the same. Equation (14) may explain the rather wide variations in ratio values which Bradford reported.

In order to develop estimation procedures which are more consistent with the observed journal productivity of a given collection, it is necessary to first separate the scale and dispersion parameters as far as possible, and to use estimators that make maximum use of the data available. Consider a ranked collection of m journals, then by (10)

$$R_m = R_1 \log(1 + bm)/\log(1 + b)$$

and

$$R_n = R_m \log(1 + bn)/\log(1 + bm)$$

or

$$R_n/R_m = \log(1 + bn)/\log(1 + bm) \tag{15}$$

where eqn (15) defines the cumulative productivity in relative terms, i.e. the cumulative *proportion* of total productivity in a collection of m journals.

For ease in subsequent computations, it is useful to carry this normalization process one step farther and to define $x = n/m$ as the proportional equivalent of journal rank, and let

$$F(x) = R_n/R_m = \log(1 + bmx)/\log(1 + bm). \tag{16}$$

Equation (17) is a *normalized* version of eqn (10), in which $F(x)$ denotes the proportion of total articles contained in the most productive x proportion of a collection.

The function $F(x)$ in eqn (17) has the form and properties of a continuous probability distribution function over the interval $0 \leq x \leq 1$. The density function $f(x)$ and the mean value $\bar{x}$ of $F(x)$ are defined as follows:

$$f(x) = (1 + bm\bar{x})/(1 + bmx) \tag{17}$$

$$\bar{x} = \int_0^1 [1 - F(x)]\,dx = \frac{1}{\log(1 + bm)} - \frac{1}{bm}. \tag{18}$$

Because of the correspondence between $F(x)$ and R_n, $\bar{x}$ provides a good way to estimate bm or b. Note that

$$F(x = n/m) = R_n/R_m = \sum_{j=1}^{n} r_j \Big/ \sum_{j=1}^{m} r_j \tag{19}$$

where r_j is the number of articles in the journal with rank j, and

$$\bar{n} = \sum_{j=1}^{m} j r_j \tag{20}$$

is a good estimator of $\bar{x}m$. By setting $\bar{x} = \bar{n}/m$, it is possible to use eqn (18) to find a numerical value for bm or b. Unfortunately, the inversion of eqn (18) is a bit tedious but results can be found by iterative methods.

When using observations of journal productivity to estimate b, caution should be exercised in using data on journals which contain only one or two relevant articles. For reasons which are developed below, it would be better to limit the observations to journals with higher productivity and a higher assurance that all journals at or above that level of productivity have been included in the observations. Thus, in choosing a value for m in eqn (15) a balance must be made between setting it too small so as to exclude data and too large so as to include questionable data.

4. ESTIMATION OF THE SCALE PARAMETER

Problems in estimating the scale parameter, B, in eqn (10) can be demonstrated by considering a rearrangement of eqn (10), whereby

$$B = R_n \log(1 + b)/\log(1 + bn). \tag{21}$$

Using eqn (21) and a given value for b, the parameter B can be estimated from any or all observed values of n and R_m, including the first and the last ranked journal.

A basic problem in choosing estimators is the possibility of systematic bias in observations of R_n. For low values of n, the most productive journals may be limited by physical restrictions on the number of relevant articles they can publish. For large values of n, many of the relatively less productive journals may contain only one or two occasional articles, some of which are likely to be missed in a search. Furthermore, the function in eqn (10) describes a continuous decline in incremental productivity and not the discrete jumps which may be particularly troublesome at higher values of n.

Another problem is that, by definition, R_n is accumulative and dependent on observations at $n-1$, $n-2$, etc. It is better to base an estimate on the independent variables, r_n, where

$$\begin{aligned} r_n &= R_n - R_{n-1} \\ &= [B/\log(1+b)][\log(1+bn) - \log(1+bn-b)] \\ r_n &= [B/\log(1+b)] \log[(1+bn)/(1+bn-b)]. \end{aligned} \tag{22}$$

By inverting eqn (22), i.e.

$$B = r_n \log(1+b)/\log[(1+bn)/(1+bn-b)] \tag{23}$$

independent estimates of B can be obtained from the observations of journal productivity, and a specified value of b.

The computation of an estimate of B can be done efficiently by fitting the regression of R_n on $\log(1+bn)/\log(1+b)$ or r_n on $\log(1+bn)/\log(1+b)\log(1+bn-b)$. The correlation coefficient provides an indication of the appropriateness of eqn (10) for describing journal productivity.

5. ESTIMATION OF TOTAL JOURNALS

The problem of estimating the total number of relevant journals and references on a particular topic is closely related to the problem of estimating the scale parameter B. BROOKES[3] and WILKINSON[6] develop estimates of the totals by computing the value of n for which $r_n = 1$, i.e. finding the rank position of a journal which yields one article or reference. More generally, eqn (22) can be solved for n, i.e.

$$n = [1-(1+b)^{-r_n/B}]^{-1} - 1/b \tag{24}$$

which could be used to estimate n for various values of r_n.

An interesting approximation to (24) can be found by rearranging eqn (10) as follows:

$$\begin{aligned} R_n &= B \log n(b + 1/n)/\log(1+b) \\ &= B[\log n + \log(b+1/n)/\log(1+b). \end{aligned} \tag{25}$$

Then, as n becomes very large, say $n = N$, $1/n$ becomes insignificant and

$$R_N = [B/\log(1+b)] \log bN. \tag{26}$$

Equation (26) is a simple linear function of the logarithm of N, as in eqn (1). This is the same relationship that Bradford observed when plotting his data and BROOKES[3] called the "Bradford–Zipf" model because of its similarity to Zipf's Law.

When eqn (26) is used in an analogous way to eqn (22),

$$\begin{aligned} r_N &= R_N - R_{N-1} = B[\log N - \log(N-1)]/\log(1+b) \\ r_N &= B \log[1/(1-1/N)]/\log(1+b) \\ r_N &= B[(1/N) - (1/2N^2) + \cdots]/\log(1+b) \end{aligned} \tag{27}$$

where the higher order terms in N can be ignored so that

$$r_N = B/N \log(1+b) \tag{28}$$

or

$$N = B/r_N \log(1+b). \tag{29}$$

Equation (29) is a simple, approximate estimator for large values of N as a function of r_N.

In particular, if N_1 denotes the number of journals with at least one article, then

$$\begin{aligned} r_{N_1} &= 1 \\ N_1 &= B/\log(1+b) \end{aligned} \tag{30}$$

where N_1 can be estimated directly from B and b. Furthermore, eqn (26) could be rewritten in the approximate form

$$R_n = N_1 \log(1+bn) \tag{31}$$

where N_1 is the total number of journals which contain one or more articles and N_1 is a relatively large number.

Estimates of the number of journals with 2 or more, and j or more articles per journal can be determined by

$$\begin{aligned} N_2 &= B/2 \log(1+b) = N_1/2 \\ N_j &= N_1/j \qquad\qquad N_1 \text{ large} \end{aligned} \tag{32}$$

and the number of journals with j articles is approx.

$$N_j - N_{j+1} = N_1/j(j+1). \tag{33}$$

Thus, approximately one-half of the journals with one or more articles have 2 or more articles, while the other half have only one article per journal. Only one-sixth of the N_1 journals have two articles, one-twelfth have three articles, and one-twelfth have four articles. These numbers indicate that there are relatively many journals with very low productivity when a collection attempts to include all references on a topic. The use of N_1 as an estimate of this total is somewhat arbitrary, since it cuts off the tail of the function where journal productivity is approximately equal to

$$r_N = N_1/N \tag{34}$$

for large values of $n = N$.

Equations (10), (26) and (31) apply to infinitely large values of n; and as n goes to infinity, R_n goes to infinity, and $R_\infty - R_n$ has an infinitely large value also. Thus, the model does not define in a finite way the number of references in the tail-end of the ranked journals. Only by truncating the function at some point $n = m$ is it possible to define a finite amount of productivity for journals with rank n or more. Truncated models are used in eqns (15) and (16). Although the use of N_1 as the point of truncation is arbitrary, it probably is as good a choice as any. N_1 can be easily estimated from B and b in eqn (30), and N_1 is a useful parameter for describing the pattern of productivity for journals with very large rank positions.

The cumulative number of articles in all journals with j or more articles can be estimated with the aid of eqns (26), (30) and (32) as approx.

$$\begin{aligned} R_{N_j} &= N_1 \log bN_j \\ &= R_{N_1} - N_1 \log j. \end{aligned} \tag{35}$$

6. OPTIMAL SEARCHING METHODS

Ordinary classification systems divide collections into zones of more and less relevant materials, and in effect define sub-collections which contain the most relevant sources on a subject. A good criterion for making such divisions in a collection is the effect it will have on the effort required to search the collection for particular queries, and a good measure is the expected search length.

If $p(n)$ denotes the probability that the nth journal examined will successfully terminate a search and provide the desired information, then the expected search effort can be defined as

$$\begin{aligned} E(n) &= \sum_{n=1}^{m} np(n) \\ 0 &< p(n) \leq 1 \\ \sum_{n=1}^{m} p(n) &= 1. \end{aligned} \tag{36}$$

The last equation states that the information is contained in the collection. The partial sum

$$P(n) = \sum_{j=1}^{n} p(j) \tag{37}$$

defines the cumulative probability of finding the sought item in the nth journal.

If the collection is searched in order of decreasing probability, i.e. $p(n) \geq p(n+1)$, then $E(n)$ achieves a minimum value and $P(n)$ is maximized for each value of n. This is the most efficient way to search a collection. However, it may be necessary or desirable in many circumstances to merely divide a collection into two zones of more and less probability. The imputed probability is the same for every member of each zone, and any search sequence within a zone is equally optimal.

When a collection is divided into two zones of size n_1 and $m - n_1$, the expected search length is

$$\begin{aligned} E_2(n_1) &= \sum_{n=1}^{n_1} nP(n_1)/n_1 + \sum_{n=n_1+1}^{m} n[1 - P(n_1)]/(m - n_1) \\ &= [m + n_1 + 1 - mP(n_1)]/2. \end{aligned} \tag{38}$$

Furthermore, if $p(n) \geq p(n+1)$, then $P(n_1)$ increases at a decreasing rate as the size of the first zone increases, and

$$\begin{aligned} \Delta E_2(n_1) &= E_2(n_1) - E_2(n_1 - 1) \\ &= [(1/n) - p(n_1)](m/2) \\ &< 0 \quad \text{if} \quad p(n_1) > 1/m \\ &> 0 \quad \text{if} \quad p(n_1) < 1/m. \end{aligned} \tag{39}$$

Thus, $E_2(n_1)$ is minimized when $p(n_1)$ is approximately equal to $1/m$, or n_1 is approximately equal to n_1^0 where

$$p(n_1^0) = 1/m. \tag{40}$$

Here, n_1^0 is a guide for dividing a collection in an optimal manner; and, in effect, determining the optimal size of sub-collections or core collections on a specific topic.

7. OPTIMAL BRADFORD-TYPE CORE COLLECTIONS

The Bradford-type journal collection is one that conforms to the model in eqn (10) with scatter parameter b and scale parameter B. Let the probability that a particular journal contains some desired information be proportional to the productivity of the journal as measured by eqn

(10), then

$$P_n = R_n/R_m = \log(1 + bn)/\log(1 + bm) \tag{41}$$

where m is the total number of relevant journals and $P_m = 1$.

The Bradford-type collection is arranged in order of decreasing probability and if searched in that way the expected search length is

$$E(n) = [bm/\log(1 + bm)] - 1/b. \tag{42}$$

If $m = N_1$ as defined in eqn (30), it is possible to evaluate $E(n)$ in terms of b and B.

If the collection is divided into two zones according to the optimality rule in eqn (40), then

$$\begin{aligned} p(n_1^0) &= 1/m = [R_{n_1^0} - R_{n_1^0 - 1}]/R_m \\ 1/m &= \log[(1 + bn_1^0)/(1 + bn_1^0 - b)]/\log(1 + bm) \\ (1/m)\log(1 + bm) &= \log[1 - 1/(n_1^0 + 1/b)]^{-1} \\ 1/(n_1^0 + 1/b) &= 1 - (1 + bm)^{-1/m} \\ n_1^0 &= [1 - (1 + bm)^{-1/m}]^{-1} - 1/b. \end{aligned} \tag{43}$$

Equation (43) can be evaluated in terms of b and B by letting $m = N_1$ and using eqn (30).

A simplified approximation of eqn (43) can be obtained by letting

$$\begin{aligned} (1 + bm)^{-1/m} &= 1 - (1/m)\log(1 + bm) \\ n_1^0 &= m/\log(1 + bm) - 1/b \end{aligned} \tag{44}$$

and

$$\begin{aligned} x_1^0 = n_1^0/m &= [1/\log(1 + bm)] - 1/bm \\ &= 1/\log bm \quad \text{approx.} \end{aligned} \tag{45}$$

for relatively large values of m and bm. Here, x_1^0 is the proportion of the total relevant journals that should be included in an optimal special collection.

Optimal core collections for various values of m are computed using the approximate eqns (44) and (45) in Table 1. This table indicates that over a very wide range of values for m and b the percent of m that should be included in the core is confined to a rather tight range of about 10–20%. The size of the core is largely determined by the value of m.

In summary, the procedure suggested is to estimate m by assuming $m = N_1$ and using eqn (30). This provides an estimate of m in terms of b and B, which can be estimated from

Table 1. Optimal core collection sizes

Total journals (m)	Core size and % for different scatter		
	$b = 0.1$	$b = 1$	$b = 10$
50	31	13	8
	62%	25%	2%
100	43	22	15
	43%	(20%)	(15%)
500	127	80	59
	25%	(16%)	(12%)
1000	220	145	109
	(22%)	(15%)	(11%)
5000	800	588	463
	(16%)	(12%)	(9%)
10,000	1450	1087	870
	(15%)	(11%)	(9%)

observations of the more productive journals in a field. The estimate of m can then be used in eqns (44) and (45) or (46) to give estimates of the optimal dimensions of the core collection.

8. GENERALIZED SEARCH MODEL

A more general model for evaluating the implications of zone searching can be developed by considering the continuous version of the Bradford model, eqn (17). Figure 1 shows the application of this model to the results of a very thorough search of 1282 journals for 9810 useful papers on the thermophysical properties of materials made by the CINDAS group at Purdue University[10]. The data indicates a good fit to the continuous Bradford model with a b value of 0.21 or $bm = 269$. The value of b was estimated by first computing $\bar{n} = 15.75$ for the 50 most productive journals, using eqn (20), and then setting $\bar{x} = \bar{n}/m = 0.315$. By substituting this value of $\bar{x}$ in eqn (18), a value of $50b = 10.5$ or $b = 0.21$ was obtained by iteration.

In Fig. 2, the collection is optimally divided into two search zones so as to minimize the

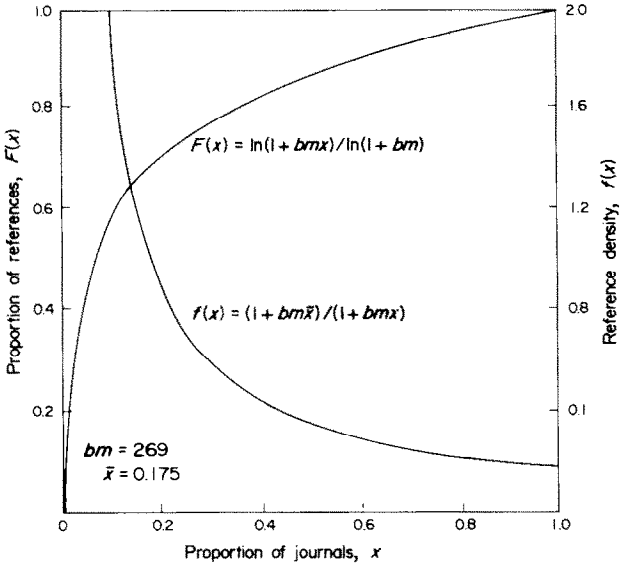

Fig. 1. Bradford distribution of the relative productivity of 1282 ranked journals yielding 9810 papers on thermophysical properties of materials identified in Ref. 10.

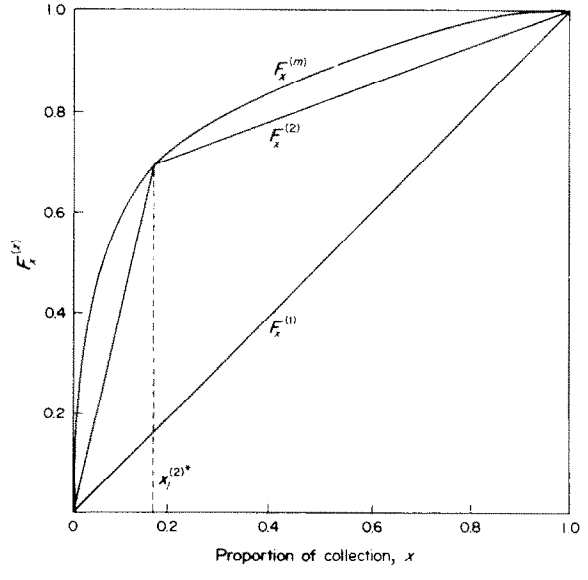

Fig. 2. Search functions $F_x^{(z)}$ defining the cumulative probability of success for z-zone searches of a collection with Bradford parameter $bm = 269$.

expected search effort according to the decision rule

$$f(x_j) = (F_{j+1} - F_{j-1})/(x_{j+1} - x_{j-1}), \qquad j = 1, 2, \ldots, z-1$$
$$F_0 = 0 = x_0 < x_j < x_z = F_z = 1 \tag{46}$$

where $F_j = F(x_j)$ defines the proportion of total productivity in the most productive fraction, x_j, of the collection; $j = 1, 2, \ldots, z-1$ denotes the zones; and x_j is the fraction of the collection in the first j zones, $x_0 = 0$, $x_z = 1$.

The decision rule (46) is developed in Ref. 5 by setting the partial derivatives equal to zero for the function

$$E_z(x_1, x_2, \ldots, x_{z-1}) = \frac{1}{2}\sum_{j=1}^{z} (x_j + x_{j-1})(F_j - F_{j-1}) \tag{47}$$

$$E_1 = 0.5 > E_z > \bar{x} = E_m \tag{48}$$

where $E_z(x)$ defines the expected search effort with z search zones. If only one zone is used the entire collection is searched randomly and the relative search effort is $E_1 = 0.5$, i.e. on the average half of the collection is examined per search attempt.

If two zones are used and the collection has a Bradford distribution, the optimal size of the first zone is $x_1^* = \bar{x}$, as defined in eqn (18); and the expected search effort E_2^* is defined by

$$E_2^* = [\bar{x} + 1 - F(\bar{x})]/2. \tag{49}$$

The lowest search effort is obtained by use of an "m zone" search or search all m items in the rank order of relative productivity. For a Bradford-type collection, this method has an expected search effort of $E_m^* = \bar{x}$.

When $bm = 269$ in eqn (18), $\bar{x} = x_1^* = E_m^* = 0.175$; and from eqn (49), $E_m^* = 0.24$. This indicates that going from a "one zone" random search to an optimal two zone search would reduce the search effort more than 50%. By going on to an "m zone" search only reduces the random search effort an additional 14%. Put another way, the best that could be done is to reduce the search effort from 0.50 to 0.175 by going from random searching to rank order searching. However, a simple two zone search could be used whereby the most productive 17.5% of the collection is searched randomly before randomly searching the balance of the collection. This type of search would reduce the search effort to 0.24 and accomplish 74% of the maximum possible reduction in search effort. These results are plotted in Fig. 3 for various values of the Bradford parameter bm.

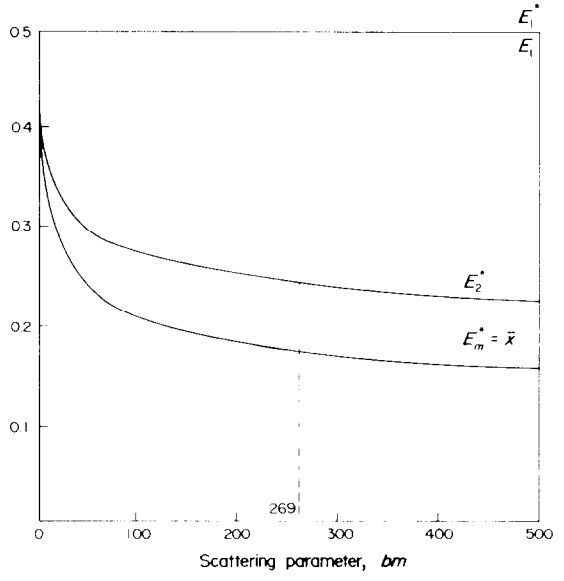

Fig. 3. Optimal expected search effort E_z^* for z-zone searching as a function of the Bradford parameter bm.

Figure 3 indicates that the relative advantage of optimal two zone searching holds steady over a fairly large range of values for bm.

In general Fig. 3 suggests that a two zone system with an initial search zone of about 20% of the collection is a highly cost-effective way to conduct searches. If a search is confined to the more productive region of a collection so that bm is relatively small, then the optimal size of the first zone is greater than 20% of this smaller collection and the benefit of two zone searching is not as large. However, if bm is unusually large, the optimal size of the first zone is less than 20% of the collection and the benefit of two zone searching is increased. In particular, if the collection is heavily diluted with irrelevant material, then the gains from two zone searching can be exceptionally high, as shown in Fig. 4, where the shaded area represents the reduction in search time, when compared with effort required by random search which is the area above the diagonal line. The area between the curved line and the shaded area is the additional reduction possible by an m zone search.

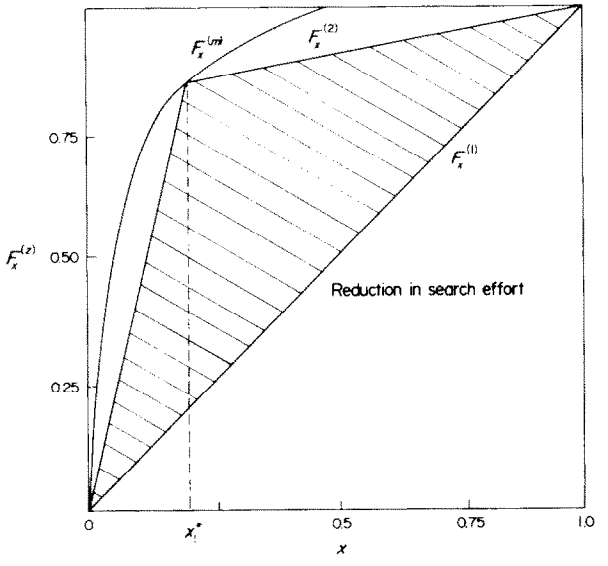

Fig. 4. One and two zone search when the collection contains irrelevant material.

9. COST OF INFORMATION

The above arguments indicate how substantial reduction in user search cost can be obtained by systematically searching a collection in zones of decreasing average productivity. The larger part of these savings can be obtained by using a simple two-zone search, with the first zone containing 10–20% of the more productive items in the collection. A search technique of this kind requires some prior organization of the collection to classify core and non-core items.

The results indicate that an appropriate cost function would have the form

$$C(I) = c_0 + c_1 I \tag{50}$$

where c_0 is the initial classification cost, c_1 is the marginal retrieval cost and I is the amount of information processed. Note that c_0 could include such initial costs as acquisition and storage as well as classification, i.e. typical library costs. Furthermore, these costs are likely to be proportional to the size of the collection, i.e.

$$c_0 = c_2 I_{\max}. \tag{51}$$

From the user viewpoint, such cost is a sunk cost and not marginal cost. However, it could be made marginal, if all or part is recovered by a lump sum user access charge, or a unit charge on the information processed, or by a combination of both kinds of charges.

The user marginal cost rate, c_1, is sensitive to the skill and extent with which items are properly classified into core and non-core tiems. If done optimally, the marginal search or

processing cost could be reduced by 50% or more depending on the scattering characteristics of the information processed. In order to gain some insight about how this could affect library performance, it is necessary to measure the value of information.

10. VALUE OF INFORMATION

The marginal economic value of information is what users are willing to pay for incremental increases in information, as measured by a demand curve. There is very little data on information demand curves, however, in a recent report, KING[11] suggests that information demand curves may have the form:

$$V'(I) = v_1 e^{-v_2 I} \qquad v_1 > 0, \qquad v_2 > 0. \tag{52}$$

King's data on the demand for scientific and technical books is reproduced in Fig. 5.

Equation (52) implies that the total value of information has the form shown in Fig. 6, i.e.

$$V(I) = (v_1/v_2)(1 - e^{-v_2 I}) \tag{53}$$

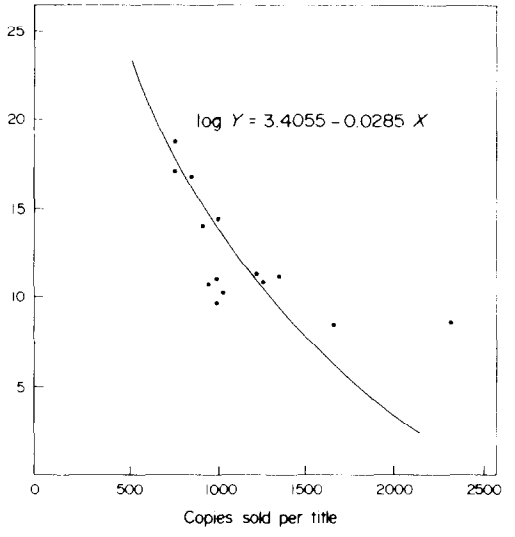

Fig. 5. Price vs demand for scientific and technical books as reported by KING[11].

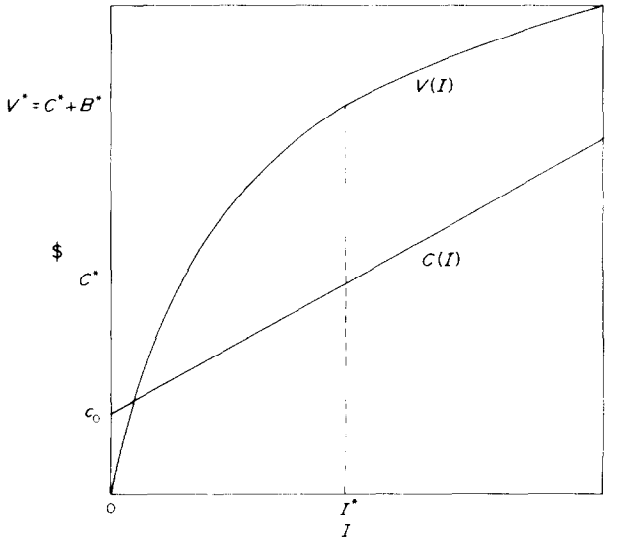

Fig. 6. Value, cost, benefit and information output at equilibrium.

where v_1 is a scale parameter, and v_2 governs the elasticity of demand, defined by:

$$E_d = -V'/IV'' = 1/v_2 I. \tag{54}$$

The demand is generally inelastic and is more inelastic as v_2 increases. A recent study by ZAIS[12] indicates that information demand is inelastic.

11. MARKET EQUILIBRIUM

If the users process information up to the point where marginal value equals marginal cost, then the equilibrium level of information processing is defined by:

$$I^* = (1/v_2)\log(v_1/c_1), \qquad c_1 \leq v_1, \qquad c_1 < 1/v_2. \tag{55}$$

This relationship indicates that the amount of information processed is inversely proportional to the elasticity parameter v_2, i.e. greater elasticity in demand leads to more information processing. This level is also influenced positively by the scale parameter v_1, and negatively by the marginal cost parameter c_1. Both of these latter effects are tempered by the logarithmic nature of the relationships.

The value of the information processed at the equilibrium level is

$$V^* = (v_1 - c_1)/v_2 \tag{56}$$

which is linearly related to c_1. V^* and the equilibrium variable cost are plotted in Fig. 6, where

$$C_V^* = (c_1/v_2)\log(v_1/c_1). \tag{57}$$

The net difference is:

$$B^* = V^* - C_V^* = (v_1/v_2) - (c_1/v_2)(1 + \log v_1 - \log c_1). \tag{58}$$

The amount B^* represents the maximum user surplus made possible by the existence of the library. In principle, therefore, this sum could be tapped for paying all or part of the library overhead cost, c_0. B^* defines the maximum amount available for this purpose, if it is recovered in the form of a lump sum user charge, so as to not affect the equilibrium position I^*.

12. COST AND BENEFIT CHANGES

Increases in the marginal cost by the imposition of marginal fees would decrease the amount of information processed, the value of that information to the user and the benefit of the library transaction, since

$$\frac{dB^*}{dc_1} = -I^* < 0. \tag{59}$$

These effects are shown in Fig. 7. The amount recovered by the marginal fee would always be less than the equilibrium benefit B^* defined above, but it may be sufficient for the purpose of recovering c_0. However, the lump sum charge would appear to be preferred since it could recover more funds without reducing the information activity level.

The reduction in marginal cost by improved library organization would tend to increase the amount of information processed, its value, and the benefit to the users, as in Fig. 8. For example, a 50% reduction in the value of c_1 would establish a new equilibrium where:

$$I_2^* = I_1^* + 0.7/v_2 \tag{60}$$

$$V_2^* = V_1^* - 0.5c_1/v_2 \tag{61}$$

$$B_2^* = B_1^* + 0.15c_1/v_2. \tag{62}$$

The increase in the amount of benefit, $0.15c_1/v_2$, is the amount available to cover the cost of the

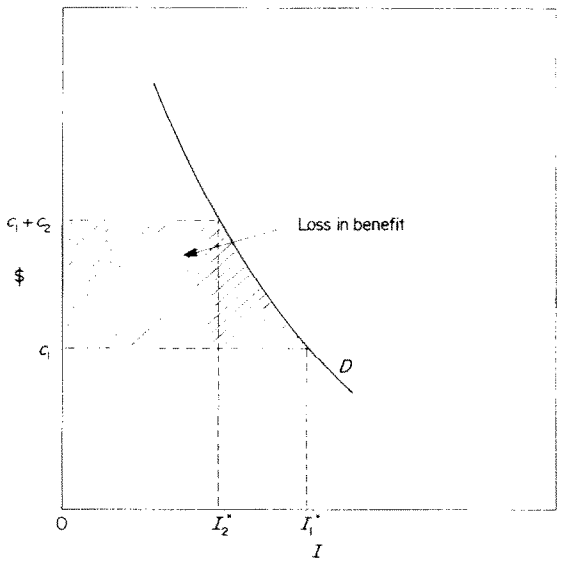

Fig. 7. Loss in benefit due to imposition of marginal fee c_2 on user marginal cost c_1.

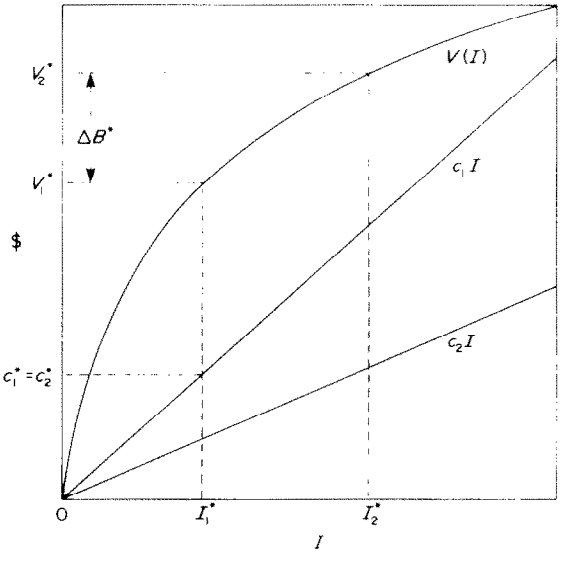

Fig. 8. Increase in value, benefit and output level when marginal cost is reduced 50% from c_1 to c_2.

improved search methods. Users should be indifferent between the payment of a lump sum equal to $0.15c_1/v_2$ for the improved ease of processing information. Yet, more information is processed and greater information value is generated. It should be noted that the changes in I^*, V^* and B^* are all inversely proportional to v_2. With greater inelasticity of demand, there is less beneficial effect of zone searching.

If there is an increase in the initial effort required from the user in order to search in a more systematic way, this cost should be added to the additional system cost when determining whether or not the increased benefits are sufficient. User initial costs probably diminish with frequency of use, but would be significant when users are forced to change their ways of processing information. This is a good argument for subsidizing new systems during the learning period.

REFERENCES

[1] S. C. BRADFORD, *Documentation*. Crosby Lockwood (1948).
[2] B. C. VICKERY, Bradford's Law of scattering. *J. Docum.* 1948, **4**, 198–203.
[3] B. C. BROOKES, Bradford's Law and the bibliography of science. *Nature* 1969, **224**, 953–956.

[4] F. F. LEIMKUHLER, The Bradford distribution. *J. Docum.* 1967, **23**, 197–207.
[5] F. F. LEIMKUHLER, A literature search and file organization model. *Am. Docum.* 1968, **19**, 131–136.
[6] E. A. WILKINSON, The ambiguity of Bradford's Law. *J. Docum.* 1972, **28**, 122–130.
[7] P. M. MORSE, The geometric and the Bradford distributions. Massachusetts Institute of Technology, Operations Research Center (1975).
[8] A. BOOKSTEIN, Patterns of scientific productivity and social change: a discussion of Lotka's Law and bibliometric symmetry. University of Chicago Graduate Library School (1975).
[9] A. BOOKSTEIN, Bibliometric symmetry and the Bradford–Zipf Laws. University of Chicago Graduate Library School (1975).
[10] A. O. CEZAIRLIYAN, P. S. LYKOUDIS and Y. S. TOULOUKIAN, A new method for the search of scientific literature through abstracting journals. *J. Chem. Docum.* 1962, **2**, 86–92.
[11] D. W. KING, *Statistical Indicators of Scientific and Technical Communication*, 1960–1980, Vol. II: *A Research Report*, King Research, Inc., 6100 Executive Blvd., Rockville, Maryland (May 1976).
[12] H. W. ZAIS, *The Pricing of Information: A Model for Selective Dissemination of Information Services.* University of California, Lawrence Berkeley Laboratory (1975).