On the Expected Gain from Adjusting Matched Term Retrieval Systems

R. H. Shumway
Westat Research, Inc., Bethesda, Maryland

A file adjustment procedure based on maximizing the Bayes expected gain is proposed for matched term retrieval systems. The expected gain and its probability distribution are derived as a function of: (1) the prior proportion of omitted terms, and (2) the coefficient of separation between two distributions corresponding to values of an adjustment statistic. An example evaluates the gain parameters for a typical information retrieval system.

Introduction

A number of papers [1–5] have been directed towards the problem of developing transformations or adjustments to be applied to term-adjusted files. Generally the term associations are used to generate a set of adjusted codings which improve retrieval by leading one more quickly to the relevant documents. However, while many empirical evaluations have been made based on file adjustments made on experimental data, theoretical investigations into the amount that one could reasonably expect to gain in retrieval effectiveness from such procedures have been notably lacking. (An exception is [7].)

It is the intention here to provide a possible basis within a decision-theoretic framework for evaluating the gain which might be expected for some file adjustment procedures. The basic approach, as in [7], is to consider only adjustments which correct for term omissions using the empirical result that the relative frequency of incorrectly applied indexed terms is negligible [6]. With this restriction we may limit our attention to developing an approach for deciding whether or not a term should be adjusted upward. This binary decision can be formulated in Bayesian terms with the probability of a user adjusting a term upwards playing the part of a prior probability. We use a measure which associates with each document (term) a measure of its association (mismatch) with the query. Our definition of gain is the amount that the measure of mismatch can be increased for irrelevant documents or decreased for relevant documents by making a set of corrections for underindexing. A procedure for adjustment is chosen which is optimal in the sense that it maximizes the gain and this gain is tabulated for various values of the system parameters. Finally we compute the probability distribution of the gain along with the positive gain probability. Thus, for binary adjustment procedures which assign either a 0 or 1 to the corrected indexing we may evaluate the gain for systems in which the basic parameters can be measured.

Sponsored in part by the Air Force Office of Scientific Research of the Office of Aerospace Research and partly by the US Department of Commerce, Patent Office

Theoretical Considerations

We shall use the formulation of Bryant et al. [6] as a basis for the theoretical development. In this case the original term indexed file is regarded as a dxt matrix of zeros and ones, say c_{ij}, with c_{ij} taking the value 1 if the jth term pertains to the ith document and 0 otherwise. We consider a set of requests or queries expressible as a matrix q_{jk} where q_{jk} is assigned a value of 1 if the searcher regards the presence of the jth term as important in the kth query and 0 otherwise. Hence, a measure of mismatch between the mth document and the kth query can be defined as:

$$r_{mk} = \sum_j (c_{mj} - q_{jk})^2. \tag{1}$$

If the c's and the q's are either 0 or 1, eq. (1) reduces to the number of mismatched terms between the mth document and the kth query. This measure of mismatch gives one the option of asking for the absence of certain terms as well as their presence. Note that in eq. (1) the summation is, in general, performed over a subset of terms which are of interest to the searcher.

We suppose now that the original indexings c_{ij} are not indexed correctly or at least they are not indexed from the point of view of the searcher or ideal user who might prefer to have assigned some different coding u_{ij}. We assume, as in [6], that underindexing represents the major type of error in the file and adjusts only terms originally indexed with a 0. Let u_{ij} (0 or 1) be the value that the ideal user would assign. Suppose that it is not feasible to correct all the term indexings c_{ij} with the ideal user and that the correction is to be made on the basis of some statistic $T = T(c_{11}, c_{12}, \cdots, c_{dt})$ computed from the other unadjusted codings. We do not consider the method (associative or otherwise) for generating this statistic but regard it as being characterized by the two conditional probability distributions:

$$\begin{aligned} F_0(x) &= P(T \leqq x \mid u_{ij} = 0) \\ F_1(x) &= P(T \leqq x \mid u_{ij} = 1). \end{aligned} \tag{2}$$

The first distribution function F_0 gives the probability distribution for the statistic T when the adjusted term should be 0 while F_1 gives the distribution of the statistic when the adjusted term should be 1. Figure 1 shows the possible forms which the density functions f_0 and f_1 corresponding to the distributions given in (2) could take. Our procedure for assigning a user indexing will be a binary decision scheme which assigns $u_{ij} = 1$ for $T > K$ and assigns $u_{ij} = 0$ for $T \leqq K$ since we shall presume that the statistic T chosen should be high when $u_{ij} = 1$ and low when $u_{ij} = 0$. The assigned user value will not always be identical to

the correct user indexing, so that to avoid confusion we will denote this assigned user indexing by b_{ij}.

Equation (1) indicates that the measure of mismatch is also influenced by the query indexing through the parameter q_{ij} which may take the values 0 or 1. Hence, the identities and values of a number of parameters associated with a single term may be arranged as in Table I. (In subsequent discussion of single-term values the subscript ij is omitted.) The library coding c is always 0 since errors of overindexing are being neglected. In order to proceed further with the analysis of Table I, some assumptions are needed about the joint probability distributions of b, c, u, and q and we assume that the user indexing u and the query are independent of each other and that the query q is independent of the adjusted coding b. Hence, the expected gain for a single term search is expressed as:

$$E(G) = \sum_{b,u,q} G(b, u, q)P(b \mid u)P(u)P(q) \tag{3}$$

where $G(b, u, q)$ is some appropriate gain function defined for each b, u, and q. The conditional distribution of b given u is determined by the decision point K in Figure 1 for:

$$\begin{aligned} P(b = 0 \mid u = 0) &= P(T \leqq K \mid u = 0) = F_0(K) \\ P(b = 1 \mid u = 0) &= P(T > K \mid u = 0) = 1 - F_0(K) \\ P(b = 0 \mid u = 1) &= P(T \leqq K \mid u = 0) = F_1(K) \\ P(b = 1 \mid u = 1) &= P(T > K \mid u = 1) = 1 - F_1(K). \end{aligned} \tag{4}$$

We also take the densities of u and q to be given as binomial with parameters v and Q, respectively. If the values

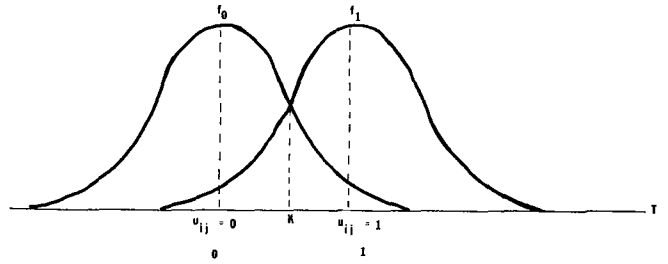

FIG. 1. Distribution of the statistic T for $u_{ij} = 0,1$ respectively with $\sigma = .5$

of the parameters are examined, it is clear that the measure of mismatch and hence the ranking is influenced in a predictable way by the adjustment procedure. Our values of the gains filled in from columns (1), (2), and (3) of Table I reflect these considerations. For example, in the first row the desired contribution to mismatch $(u - q)^2$ is 0 with the contribution to mismatch without adjustment $(c - q)^2$ also being 0. The adjusted mismatch is 1 which is in error, contributing a gain of -1. The reader may easily convince himself that the other gains are reasonable and that positive gains tend to reflect a favorable adjustment of the mismatch and hence, the ranking. Then, using Table I and eqs. (3) and (4) with the binomial assumption on u and q leads to:

$$\begin{aligned} E(G) = &-(1 - Q)(1 - v)(1 - F_0) \\ &+ (1 - Q)v(1 - F_1) - Q(1 - v)(1 - F_0) \\ &+ Qv(1 - F_1) = v(1 - F_1) - (1 - v)(1 - F_0) \end{aligned} \tag{5}$$

which is maximized by choosing a value K such that:

$$\frac{f_1(K)}{f_0(K)} = \frac{1 - v}{v}. \tag{6}$$

If the probability densities f_0 and f_1 are known or a discrete approximation is available, we may solve for K using eq. (6) and then substitute into eq. (5) to determine the maximum expected gain. For example, if the densities f_0 and f_1 can be regarded as being approximately normal with means 0 and 1 respectively and common variance σ^2, eq. (6) yields:

$$K = \frac{1}{2} + \sigma^2 \log \frac{1 - v}{v} \tag{7}$$

with the maximum expected gain per term represented in (5) as a function of σ^2 and v. In this case the mean separation is unity so that the value of σ represents a *coefficient of discrimination* in the sense that a larger σ is associated with an increased difficulty in discriminating between $u = 0$ and $u = 1$.

The above results pertain to single-term searches only and it would be useful to extend the results to a search

TABLE I. SYSTEM PARAMETERS FOR FILE ADJUSTMENT

q c u b	(1) $(u - q)^2$	(2) $(c - q)^2$	(3) $(b - q)^2$	Gain
0 0 0 1	0	0	1	−1
0 0 0 0	0	0	0	0
0 0 1 1	1	0	1	1
0 0 1 0	1	0	0	0
1 0 0 1	1	1	0	−1
1 0 0 0	1	1	1	0
1 0 1 1	0	1	0	1
1 0 1 0	0	1	1	0

Col. (1): desired contribution to mismatch
Col. (2): contribution to mismatch without adjustment
Col. (3): contribution to mismatch with adjustment

$q \equiv$ query indexing
$u \equiv$ user indexing
$b \equiv$ adjusted indexing
$c \equiv$ original indexing

TABLE II. PROBABILITY DISTRIBUTION OF GAIN FOR A SINGLE-TERM SEARCH

G	*Probability distribution* P_G
−1	$(1 - v)(1 - F_0)$
0	$F_0 + v(F_1 - F_0)$
1	$v(1 - F_1)$

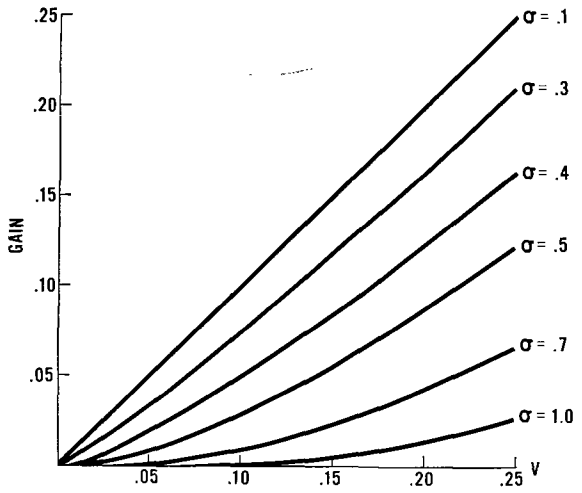

FIG. 2. Maximum expected gain as a function of v, the prior probability of adjustment

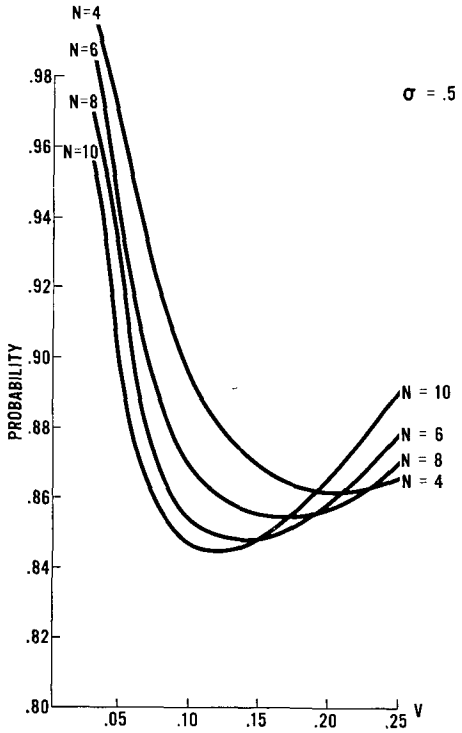

FIG. 3. Probability of positive gain for N-term searches $\sigma = .5$

involving N terms. In addition, we are interested not only in the expected maximum gain but also in the exact or approximate probability distribution of the gains. The gain density for a single-term search can be written down immediately from Table I and is reproduced in Table II.

In an N-term search the gain can range over the integers $-N, -N + 1, \cdots, 0, 1, \cdots, N$. Then, let n_G be the number of terms in the search that produced a single-term gain of G. Then, if the total gain is designated by G_T we may write

$$P(G_T = k) = \sum_{\substack{n_1 - n_{-1} = k \\ n_{-1} + n_0 + n_1 = N}} \frac{N!}{n_{-1}!n_0!n_1!} P_{-1}^{n_{-1}} P_0^{n_0} P_1^{n_1}. \quad (8)$$

For moderate sized N, G_T will be the sum of the individual single term gains and the central limit theorem will apply yielding:

$$P(G_T \leq k) \cong \phi\left(\frac{k - \mu_T}{\sigma_T}\right); \quad (9)$$

$$\mu_T = NE(G), \qquad \sigma_T = \sigma_G(N)^{1/2}$$

and an approximate expression for the probability distribution of the gain. Here $\phi(x)$ denotes the cumulative normal distribution with $E(G)$ and σ_G the mean and standard deviations of the gain as computed from Table II. One measure of possible interest would be $P(G_T > 0)$ or the probability of making a positive gain. We shall henceforth refer to this measure as the *positive gain probability*.

Examples

The measures of effectiveness developed in the preceding section will be quite different for the various adjustment procedures in both the form and separation of the distributions f_0 and f_1 of Figure 1. Empirical data categorizing adjusted and unadjusted terms into correctly adjusted and unadjusted terms and incorrectly adjusted and unadjusted terms, as well as the sample values T of the adjustment statistic will be needed in order to determine the performance characteristics of a particular system. Since the distribution of T is often the distribution of some linear combination of adjacent terms as in adjustment procedures using regression or other associative correction measures, we may frequently assume that it is approximately normal for terms that should have been adjusted as well as for terms that should not have been adjusted. For purposes of simplified computation we shall also assume in this example that the variances are equal in the two populations and that the average separation between f_0 and f_1 has been normalized to one. This allows the use of eq. (7) to determine a cutoff point which maximizes the expected gain. Equation (5) then determines the maximum expected gain as a function of the parameters v and σ^2. Figure 2 shows the expected gain per term in the mismatch measure as a function of the prior proportion of omitted terms v and the spread of f_0 and f_1 denoted by σ^2. Note that we can never gain more on the average than the value of the parameter v. Also, with increasing σ the maximum expected gain goes down while with increasing v the maximum expected gain increases. If the basic parameters remain relatively constant from term to term the expected total gain from an N-term search is N times the expected single-term gain. Note that this expected gain is over terms in a single document which were not indexed in the original file. Hence, in a 20-term search a single document might contain only ten candidates for adjustment. Therefore, using Figure 2 with $v = .22$ and $\sigma = .5$ a maximum expected gain of .10(10) = 1 would be reasonable for documents containing ten terms originally indexed as zero.

In some cases a more interesting and informative measure might be the probability of making a specified gain,

determined from eq. (8) or its approximation (9). The characteristics of the system will determine the particular probabilities which contribute the most as measures of effectiveness. We have chosen to present the probability of making a positive gain $P(G_T > 0)$ in Figures 3 and 4. Note that while the expected gain increases with v and decreases with σ the probability of some gain (positive gain probability) for values of v less than .20 is increased with an increased variance. Hence, in this example, the improvement in expected gain with the decreased σ leads to a slight decrease in the positive gain probability. The phenomenon observed above where the expected gain and positive gain probability seem to work against each other does not cause serious problems since the positive gain probability is uniformly high over the entire range of v. The same conflict characterizes the relation between the gain and the number of terms in the search with expected gain increasing for higher N and the positive gain probability decreasing. If the mean separation between the distributions in Figure 1 is positive we will always have a positive expected gain regardless of the variance σ^2.

As another example, consider the computation of the entire probability distribution of the gain as given by eq. (8). Let us suppose that in making five-term searches it is true on the average that three terms in the documents would not be coded in the unadjusted file. Assume also that the prior probability of omission, v, is .10. Then, for $\sigma = .5$ we use eq. (8) to determine that the probability of gaining 1 is about .12. If we are searching for presence in the query then there is a .12 chance of decreasing the mismatch by 1, which with a total possible mismatch of 5 would lead to a substantial improvement in the ranking. If the prior probability of omission is .20 the chance of a gain of 1 increases to .26. In this case the expected gain and gain probability do not seem to work against one another. It is also clear that the gain probability is a measure of the improvement in the ranking if it is assumed that a documents position in the ranking is determined incorrectly because of omitted terms.

Conclusions

We have developed the expected Bayes gain and the positive gain probability as measures of retrieval effectiveness for file-adjustment procedures. These measures do not depend on the form of the adjustment which has been applied, as it may be any one of a number of the so-called associative schemes. The requirements are that the proposed procedure generate a set of adjustment statistics on a continuous scale and that the correct codings corresponding to these adjustments be available. Then the competing forms of Figure 1 can be plotted and the distributional forms F_0 and F_1 can be estimated. This yields a critical value K which maximizes the expected Bayes gain. The resulting measures of retrieval effectiveness (here the expected gain and the positive gain probability) are expressed in terms of the prior probability that a user would have preferred a different indexing. The computed examples show that it would be useful to examine the parameters in an operating system quite closely to determine the relative benefits of competing adjustment procedures.

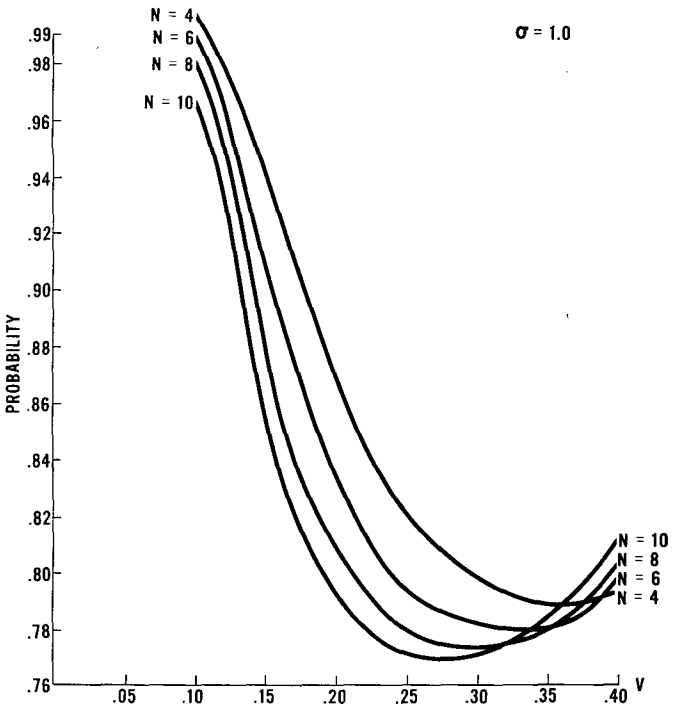

FIG. 4. Probability of positive gain for N-term searches $\sigma = 1.0$

RECEIVED NOVEMBER, 1966; REVISED MARCH, 1967

REFERENCES

1. BAKER, FRANK B. Information retrieval based on latent class analysis. *J. ACM 9* (Oct. 1962), 512–521.
2. BORKO, HAROLD AND BRENICK, MYRNA. Automatic document classification. *J. ACM 10* (April, 1963), 151–162.
3. GIULIANO, VINCENT E. AND JONES, PAUL E. Linear associative information retrieval. In Howerton and Weeks, *Vistas in Information Handling, Vol. 1*, Spartan Books, Washington, D. C., 1963, Ch. 2.
4. MARON, M. E. Automatic indexing: an experimental inquiry, *J. ACM 8* (July, 1961), 404–417.
5. NEEDHAM, B. M., AND JONES, K. SPARCK. Keywords and clumps. *J. Documentation 20*, 1 (Mar., 1964), 5–15.
6. BRYANT, E. C., KING, D. W., AND TERRAGNO, P. J. Analysis of an indexing and retrieval experiment for the organometallics file of the U. S. Patent Office, P. O. 10. Report to U. S. Dept. of Commerce, Patent Office Contr. 6078, Westat Research, Inc., 1963, PB 166 488.
7. BRYANT, E. C., SEARLS, D. T., AND SHUMWAY, R. H. Associative correction for underindexing. Air Force Off. of Sci. Research, Contr. AF49(638)-1484, Nov. 30, 1965, AD 628 191 (submitted 1966).