A Probabilistic Approach to Automatic Keyword Indexing

Part II. An Algorithm for Probabilistic Indexing

In Part I of this study,* a mixture of two Poisson distributions was examined as a model of specialty word distribution. Formulas expressing the three parameters of the model in terms of empirical frequency statistics were derived, and a statistical measure intended to identify specialty words, consistent with the model, was proposed.

In the present paper, Part II of the study, a probabilistic model of keyword indexing is outlined, and some of the consequences of the model are examined. An algorithm defining a measure of *indexability* is developed—a measure intended to reflect the relative significance of words in documents. The measure is evaluated and is found to consistently produce indexes superior to those produced by another measure which had previously been identified in the literature as producing the best results.

Stephen P. Harter
Library Science/Audiovisual Program (FAO 186)
University of South Florida
Tampa, FL 33620

• Introduction

The 2-Poisson distribution is a mathematical model descriptive of the distribution of specialty words in a technical literature. The model is discussed in detail by Abraham Bookstein and Don R. Swanson (*1*), and in Part I of the present study (*2*). To obtain the model, it is assumed that "level of treatment" t, at which a concept is dealt with in a document, has an effect both on the number of tokens k_t of the word representing the concept in the document and the probability u_t that the document will be found relevant to a request for information concerning that subject. The greater the level of treatment, the larger both k_t and u_t will tend to be.

The model specifically assumes that each word in a technical literature represents a concept which is treated in documents belonging to the literature at exactly two degrees or levels. Within each of the document classes I and II thus defined, the model assumes that documents are equally likely to be found relevant to a request for information on the concept, and that the number of tokens k of the word in these documents is described by Poisson distributions with means λ_1 and λ_2, respectively. Formally, if the proportion of documents belonging to class I is denoted by π, then the 2-Poisson model is defined by the equation

$$P(k) = \pi \frac{e^{-\lambda_1} \lambda_1 k}{k!} + (1-\pi) \frac{e^{-\lambda_2} \lambda_2 k}{k!},$$

where $P(k)$ is the probability that a document contains exactly k tokens of the word.

**Journal of the American Society for Information Science*, 26 (No. 4): 197-206 (1975).

In Part I of this study, results were reported which confirmed the observations of other researchers that specialty words are likely to possess frequency distributions which cannot be described by a single Poisson distribution (*2*). The assumptions underlying the 2-Poisson model as a model of specialty word use were investigated and some of the implications of the model were explored. A suggestion made by John Swets (*3*) and improved by B.C. Brookes (*4*) was modified to produce a statistical measure consistent with the 2-Poisson model for identifying specialty words:

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}}$$

In a test conducted on an experimental document collection, a set of 650 abstracts of the works of Sigmund Freud (*5*), the measure z was found to be relatively successful in this purpose.

In this paper, we consider the problem of making use of the 2-Poisson model for purposes of establishing a criterion for automatic indexing. The general approach we will take is based on decision theory. The relevance of the decision-theoretic approach to problems of information retrieval was apparently first recognized by M.E. Maron and J.L. Kuhns, in their early paper, "On Relevance, Probabilistic Indexing, and Information Retrieval" (*6*). More closely related to the present approach is the paper by John Swets (*3*) and the extensive study by Frederick Mosteller and David L. Wallace (*7*). Swets proposed a general model of an information retrieval system based on decision-theoretic, or Bayesian, concepts, while Mosteller and Wallace used a Bayesian statistical analysis to identify the authors of a number of the Federalist papers whose authorship is in dispute. In papers directly concerned with the processes of indexing and retrieval, Donald Kraft (*8*) and Abraham Bookstein and Don R. Swanson (*9*) have modeled the indexing process in terms of decision theory. The present approach elaborates the cost approach to indexing outlined by these latter authors and examines some of the consequences of this approach.

The models to be developed will be expressed in the notation of probability theory:

- $P(A)$ represents the probability of the occurrence of event A,
- $P(A/B)$ represents the probability of the occurrence of event A, given knowledge of the prior occurrence of event B
- $P(A, B)$ represents the probability of the simultaneous occurrence of events A and B

• A Cost Model for Keyword Indexing

Following Kraft (*8*) and Bookstein and Swanson (*9*), we introduce the notion of the costs of indexing errors. There are essentially two kinds of errors an indexer can make: errors of commission and errors of omission. The indexer can tag a document by a term but find that the document is judged non-relevant to a particular request for information on that subject. Conversely, the indexer can fail to index a document by the term, even though the document would have been judged relevant by the requester.

The value of particular documents to particular requesters cannot be ascertained prior to the act of retrieval. However, we suppose that a requester r is able to assess the *average* costs, c_{1r} and c_{2r} associated with failing to index a document he would have judged relevant and with indexing a document he would have judged non-relevant, respectively. These costs define the retrieval performance expected by requester r. If, for example, the average cost c_{1r} of failing to index a relevant document is large, relative to the average cost c_{2r} of indexing a non-relevant document, then r is willing to tolerate relatively large numbers of non-relevant documents in order to obtain the number of relevant documents he requires. In the language of information retrieval, he is willing to accept a low *precision* as the price of high *recall*.

For each word, let the costs c_{1r} and c_{2r} be averaged over the set of requesters r:

$$c_1 = \frac{1}{n}\sum_{r=1}^{n} c_{1r}; \quad c_2 = \frac{1}{n}\sum_{r=1}^{n} c_{2r}.$$

Costs c_1 and c_2 are the average cost of failing to index a relevant document and the average cost of indexing a non-relevant document, respectively.

In the discussion that follows, the symbol (w) denotes the phrase "the concept named by the term w in the document d." A formal model of keyword indexing arises by supposing that $P(R)$, the probability that a document d will be found relevant to a request for information on (w), can be estimated by an indexer, and that the indexer acts so as to minimize expected user losses (*8,9*). This criterion can be expressed by the rule: Index d by w if and only if:

$$P(R) \cdot c_1 > P(\bar{R}) \cdot c_2 \ . \qquad (1)$$

A number of factors can be identified as affecting the value of the probability $P(R)$. These fall into two broad classes:

(i) properties of the documents in the collection with respect to (w);
(ii) properties of the term w.

The first set of factors is concerned with the extent to which a document treats a subject, both in an absolute sense and as compared to the other documents in the library. For example, if the primary subject of a document is (w), then d is much more likely to be judged relevant to the request w than documents which deal

with (w) only in a peripheral fashion. On the other hand, a document dealing with (w) in a minor way may be relatively likely to be found relevant to a request w if no documents treat (w) as a major subject.

The second set of factors which may affect the value of $P(R)$ is concerned with the semantic "fuzziness" of w. We take a semantically fuzzy word to be a word which possesses a number of distinct referents, or which occurs freely in many linguistic environments, or whose reference class is generically broad. An index user making a fuzzy request w is less likely to find retrieved documents relevant to his request than if w were semantically "sharp." This idea has been discussed in considerable detail by F. W. Lancaster (*10*).

Inequality (1) may be simplified by substituting $[1-P(R)]$ for $P(\bar{R})$ and solving for $P(R)$. The criterion thus becomes: "Index d by w, if and only if:

$$P(R) > c, \tag{2}$$

where $c = \frac{c_2}{c_1 + c_2}$. c is a single measure of the level of retrieval effectiveness required by the user population.

• Implications of the Model

Possibly the most commonly used measures of retrieval effectiveness are the dual measures of recall—the proportion of relevant documents which are retrieved—and precision—the proportion of retrieved documents which are relevant. In general, an empirical "tradeoff" has been observed to exist between recall and precision. Lancaster has discussed the effect of varying the level of exhaustivity of indexing on retrieval performance (*10*). We now prove the relationship between exhaustivity of indexing and recall and precision observed by Lancaster and others:

> *Result (i):* As exhaustivity of indexing is increased, expected recall increases and expected precision decreases.

The following notation is used:

N denotes the number of documents in the library,

X denotes the number of documents in the library expected to be found relevant to the request w,

Y denotes the number of documents in the library which are indexed by the term w,

Z denotes the number of documents in the library indexed by the term w and expected to be found relevant to the request w.

Let R_j denote the event: Document d_j is found relevant to a randomly chosen request w. Then

$$X = \sum_{1}^{N} P(R_j).$$

Label the Y documents indexed by w by $d_1, \ldots, d_Y$. Then

$$Z = \sum_{1}^{Y} P(R_j).$$

By inequality (2) above, our indexing criterion can be expressed by the inequality $P(R) > c$. Expected recall is just Z/X. Now, if exhaustivity of indexing is increased by decreasing the value of c, Y increases or remains the same. Thus, Z will increase or remain the same. Since the value of X is unchanged, expected recall increases, or remains the same.

Expected precision P is given by

$$P = \frac{1}{Y} \sum_{1}^{Y} P(R_j).$$

Now increase the exhaustivity of indexing so that Y' documents are indexed by w, where $Y' > Y$. The new precision P' is

$$P' = \frac{1}{Y'} \sum_{1}^{Y'} P(R_j).$$

For each of the newly indexed documents d_j, as a result of having increased the exhaustivity of the indexing, $P(R_j)$ is less than, or equal to, the value of $P(R_j)$ for each of the documents indexed previously by w. Hence, the new mean P' is less than or equal to the old mean P, and expected precision decreases or remains the same.

Result (i) provides theoretical support to the often observed empirical fact that recall and precision are inversely related, and cannot be maximized simultaneously by either increasing or decreasing exhaustivity of indexing. However, as Cyril Cleverdon has observed, the "fundamental law" between recall and precision may not necessarily hold in real retrieval situations (*11*). With respect to the present model, the reason is that indexing is *probabilistic* in nature. In real retrieval situations, a sample of requesters is drawn, often of size $n = 1$, on the basis of which the values of recall and precision are calculated.

Thus, even though $P(R_i) > P(R_j)$ for two documents d_i and d_j, it may be the case that d_j is found relevant by a particular requester while d_i is not. It is only when relevance judgments are obtained by a "large enough" sample of requesters that we can expect the inverse relation between recall and precision to hold for that sample.

A second explanation for the occasional failure of the inverse relationship between recall and precision to be obtained in practice can be traced to the model's assumption that an indexer makes use of the probability $P(R)$ associated with a document. In actual practice, it is only the indexer's estimate of $P(R)$

which is available, an estimate which can be affected by a host of variables.

The theoretical result that recall and precision cannot be maximized simultaneously suggests that an optimal indexing strategy is one which strikes a balance between these conflicting goals. Formally, we propose the following:

> *Definition:* An optimal indexing strategy is a strategy which, for each value of expected recall, achieves the maximum possible value of expected precision.

We now prove:

> *Result (ii):* Our simplified indexing strategy is an optimal strategy.

Consider the criterion: Index d by w if and only if $P(R) > c$. The result of applying this criterion is that the value of $P(R)$ is at least as great for all documents indexed by w as it is for the documents not indexed by w. We show that every indexing strategy possessing this property is an optimal strategy.

Denote by the letter A any indexing strategy which has the result: for every document d_j indexed by w and every document d_k not indexed by w,

$$P(R_j) \geqslant P(R_k). \tag{3}$$

Suppose that the result of applying an indexing strategy A is that Y_A documents are indexed by w. Suppose that Y_B documents are indexed by another, possibly identical strategy B, and that the expected levels of recall Rec_A and Rec_B obtained by the two strategies are equal. Each of the strategies A and B defines a set of indexed documents. Rank the documents belonging to each of these sets according to the value of $P(R)$ associated with each document. Label the documents in each set according to these rankings, so that

$$P(R_j^A) \geqslant P(R_k^A)$$

and

$$P(R_j^B) \geqslant P(R_k^B), \tag{4}$$

for all $j \leqslant k$ for which these quantities are defined. Conditions (3) and (4) guarantee that $P(R_j^A) \geqslant P(R_j^B)$ for all j for which these quantities are defined. Therefore:

$$\sum_1^{Y_A} P(R_j^A) > \sum_1^{Y_A} P(R_j^B).$$

But we assumed that $Rec_A = Rec_B$. Since $P(R_j) \geqslant 0$ for all j, we must have $Y_B > Y_A$. Hence,

$$Pre_A = \frac{X\ Rec_A}{Y_A} \geqslant \frac{X\ Rec_B}{Y_B} = Pre_B,$$

and the proof is complete.

• Extension of the Model of Keyword Indexing

The 2-Poisson model of specialty word distribution assumes the existence, for each word w, of two classes of documents, each class homogeneous with respect to (w). These classes, labelled I and II, have been interpreted as representing levels of treatment of (w) in documents belonging to the two classes. In terms of the 2-Poisson model, the *a priori* probability that d is a member of class I and class II is given by $P(d\epsilon\text{I}) = \pi$ and $P(d\epsilon\text{II}) = 1-\pi$. Then

$$P(d\epsilon\text{I}, k) = P(d\epsilon\text{I}) \cdot P(k/d\epsilon\text{I}) = \pi \frac{e^{-\lambda_1}\ \lambda_1 k}{k!}, \text{ and}$$

$$P(d\epsilon\text{II}, k) = P(d\epsilon\text{II}) \cdot P(k/d\epsilon\text{II}) = (1-\pi) \frac{e^{-\lambda_2} \lambda_2 k}{k!}.$$

We assume that every document d is a member of either class I or class II. Then

$$P(d\epsilon\text{I}, k) + P(d\epsilon\text{II}, k) = P(k).$$

The probability that d is a member of I conditional on the fact that w occurred k times in d is therefore given by

$$P(d\epsilon\text{I}/k) = \frac{P(d\epsilon\text{I}, k)}{P(k)} = \frac{\pi\, e^{-\lambda_1}\ \lambda_1 k}{\pi\, e^{-\lambda_1}\ \lambda_1 k + (1-\pi)\, e^{-\lambda_2}\ \lambda_2 k}. \tag{5}$$

For a particular document d and a word w occurring in d a total of k times, the probability $P(d\epsilon\text{I}/\ k)$ that d is a member of class I with respect to the concept named by w is given by equation (5). Note that $P(d\epsilon\text{I}/\ k)$ is a function of k, the number of occurrences of w in d, and the overall frequency distribution belonging to w, which determines the value of π, λ_1, and λ_2.

Let the symbols u_1 and u_2 refer to the probability that a member of document classes I and II respectively will be found relevant to the request w. The overall probability $P(R)$ that a document will be found relevant by a requester can be expressed as

$$P(R) = u_1 \cdot P(d\epsilon\text{I}/\ k) + u_2 \cdot P(d\epsilon\ \text{II}/k). \tag{6}$$

Substituting $P(d\epsilon\text{II}/k) = 1-P(d\epsilon\text{I}/k)$ into (6) and then into the previous inequality (2), and simplifying, we obtain the indexing criterion: Index d by w if and only if

$$P(d\epsilon\text{I}/k) > \frac{c - u_2}{u_1 - u_2}. \tag{7}$$

For notational convenience, denote by the symbol α the quantity

$$\frac{c - u_2}{u_1 - u_2}$$

The number α is an overall measure of the potential *effectiveness* of w as an index term, taking into consideration both the stated requirements of system users and the efficacy of the term w as a request term. Clearly small values of α indicate highly effective words, while large values of α indicate ineffective words.

• Automatic Keyword Indexing

If the numbers c, u_1, and u_2 were known, for all w in a document d, then an index set for d would be completely determined. In what follows, it is assumed that these numbers are not known. However, we assume that we have knowledge of the frequency distributions of the words used in the document collection of interest. We will then show, with the appropriate approximating assumptions, that our model of keyword indexing can guide us to do automatic keyword indexing.

We begin by noting that the probabilities u_1 and u_2 can be related to the average frequency of occurrence λ_1 and λ_2 in document classes I and II, by two hypotheses suggested earlier; that both sets of parameters are directly related to the level of treatment associated with documents in these classes. Specifically, a relatively large difference between λ_1 and λ_2 is assumed to imply a substantial difference in the extent to which (w) is treated in document classes I and II. This difference in turn is assumed to be reflected in the probabilities u_1 and u_2 that documents belonging to the two classes will be found relevant to a request; $(u_1 - u_2)$ will be relatively large. On the other hand, if λ_1 is relatively near λ_2, it is assumed that the subject is dealt with in much the same way in document classes I and II and hence that u_1 is near u_2; $(u_1 - u_2)$ will be relatively small.

The value of c associated with the word w can also be related to values of λ_1 and λ_2 associated with the word. If, for some word w, λ_1 is near λ_2, a high recall can be obtained only by retrieving all, or nearly all, the documents in the collection; such a word fails to distinguish two distinct classes of documents. We hypothesize that literature searchers, being aware of this, will tend to request only high precision searches for the word w. That is, c_{1r} will tend to be relatively small in comparison to c_{2r}, and

$$c = \frac{c_2}{c_1 + c_2}$$

will be relatively large. However, if λ_1 is much larger than λ_2, no conclusions can be drawn as to the probable value of c.

In summary, we hypothesize that the effectiveness

$$\alpha = \frac{c - u_2}{u_1 - u_2}$$

that is associated with a word w is related to the *degree of overlap* between the populations I and II defined by the 2-Poisson distribution. If λ_1 is near λ_2 and the overlap between the populations is large, c will tend to be relatively large and $(u_1 - u_2)$ will tend to be small. Thus α will tend to be large. If λ_1 is substantially larger than λ_2, and the overlap between populations is small, then while the value of c is unpredictable, $(u_1 - u_2)$ will tend to be relatively large, and α will tend to be small.

In Part I of this study, a measure of effectiveness based on the degree of overlap between populations I and II was proposed:

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}} \quad .$$

The measure z was found to be successful in separating specialty words from non-specialty words (*2*). Because small values of α are associated with highly effective words, the general dimensions of α are roughly those of the number ($-z$). For purposes of automatic indexing, based on our assumption that c, u_1, and u_2 are not known, we replace α in inequality (7) by ($-z$). Thus we obtain the decision rule: Index d by w if and only if

$$P(d\epsilon \mathrm{I}/k) + z > 0. \qquad (8)$$

From inequality (8) we define our measure of *indexability*, that is, our measure of the relative significance of w in d, as the number $\beta = P(d\epsilon \mathrm{I}/k) + z$. Measure β is composed of two components. The first, $P(d\epsilon \mathrm{I}/k)$, is an estimate of the relative level of treatment of the concept w in the document d, while the second, z, is an estimate of the overall effectiveness of w as a potential index term. To do automatic indexing, we rank the words in d by the values β.

• Index Term Weighting

In the present model, index terms are weighted by the values $\beta = P(d\epsilon \mathrm{I}/k) + z$. That weighting in general is of value has received experimental support in the information science literature (*12, 13, 14, 15, 16*). An approach to term weighting originally conceived by H. P. Luhn is to weight index terms in a document by k, within document frequency of occurrence of the word in the document (*17*). Gerard Salton and his colleagues have reported the results of experiments in which terms weighted in this way result in significantly superior retrieval performance over unweighted terms (*12, 13*).

Karen Sparck Jones has taken a different approach to term weighting (*14*). Sparck Jones considered terms

occurring infrequently in the document collection to be more valuable than frequently occurring terms. Her general conclusion was that significantly superior retrieval results were obtainable by weighting by "collection frequency."

Salton and C.S. Yang and, in a separate study, Sparck Jones, have reported results of comparisons between weighting by collection frequency and weighting by k, within document frequency, with somewhat mixed results. Sparck Jones concluded that weighting by collection frequency "leads to material performance improvement in quite different document collections," while Salton and Yang found that both procedures led to improved performance, but that "the results concerning the best procedure to be followed differ from collection to collection (*15, 16*)."

Each of the two approaches to index term weighting is consistent with the model of indexability outlined in the present study, under certain conditions. First, consider within document frequency, k. Suppose that there exist two terms w_i and w_j, for which $\pi_i \approx \pi_j$, $\lambda_{1i} \approx \lambda_{1j}$, and $\lambda_{2i} \approx \lambda_{2j}$, and that w_i and w_j occur k_i and k_j times respectively in a document d. Then $\beta_i \geq \beta_j$ if and only if $k_i \geq k_j$. That is, other factors being roughly equal, within document frequency k is a valid measure of the relative significance of w_i and w_j in d.

The argument used by Sparck Jones in support of weighting index terms by a function of collection frequency was that document/request matches on nonfrequent terms should be treated, for retrieval purposes, as being more effective, or valuable, than matches on frequent terms. This view is consistent with the present model, just if a document d treats two subjects (w_i) and (w_j) to roughly the same extent, that is, if $P(d\epsilon \mathrm{I}_i/k_i) \approx P(d\epsilon \mathrm{I}_j/k_j)$. Then $\beta_i \geq \beta_j$ if and only if $z_i \geq z_j$. Where Sparck Jones measures the effectiveness of an index term by its statistical specificity—by its overall collection frequency—we use the overlap z between the 2-Poisson populations I and II.

We now briefly consider two other methods of term weighting, suggested by H.P. Edmundson and R.E. Wyllys (*18*). Following Edmundson and Wyllys, we use the notation:

f_i = relative frequency of w_i within $d = k_i/L_d$

r_i = relative frequency of w_i in general use $= F_i/L_c$.

Edmundson and Wyllys proposed the measures $m_1 = f/r$ and $m_2 = f \cdot r$ as being especially promising measures of indexability.

Consider first the measure m_1, and two words w_i and w_j. Restricting our attention to a particular document d, the ranking produced by m_1 is equivalent to that produced just by k/F, since L_d and L_c are constant for all words of the document. Thus $m_{1i} > m_{1j}$ if and only if $k_i/F_i > k_j/F_j$, or $k_i/k_j > F_i/F_j$. Expressing the ranking produced by m_1 in this way suggests that far too much weight is placed on the ratio of total frequencies F_i/F_j. For example, in the experimental document collection, the important term "dream" has overall frequency F_i greater than 400. This means that "dream" would have to occur in an abstract more than 10 times in order to be ranked above any word with overall frequency less than 40. Conversely, all words in a document d with overall frequency F_i exactly equal to 1 would occur at the very top of the ranked list, since $1 = 1/1 = k_i/F_i \geq k_j/F_j$, for all w_j. But, in general, such words have no particular utility for purposes of indexing. We believe that these considerations constitute strong *a priori* grounds for questioning the adequacy of m_1 as an effective measure of indexability.

We now consider the measure m_2. For a particular document d, the ranking produced by m_2 is equivalent to that produced by $k - F \cdot L_d/L_c$. In the experimental document collection, $L_d < 450$ for all documents and $L_c \approx 145{,}000$. Then $F \cdot L_d/L_c < 1$ if and only if $F < 322$. Thus, except for very high frequency specialty words ("dream," "ego," "psychoanalysis," and "sexual"), the ranking produced by $m_2 = f \cdot r$ is equivalent to ordering first by within document frequency k and then inversely within each of the k classes thus formed by overall frequency F. The measure m_2 is thus roughly equivalent to simple within document frequency k.

In their study comparing several *ad hoc* measures of significance, John Carroll and Robert Roeloffs found that results obtained by simple within document frequency k were superior to the measure m_2, but statistically significantly so in only one of three indexing trials (*19*). Both of the measures k and m_2 were found to be significantly superior to all other measures the authors tested, including m_1, in all trials.

We conclude the study by subjecting the β-measure to an empirical test.

• Methodology

To evaluate the quality of an index produced by the measure β, we will compare it to a human-prepared index of the same document. Admittedly, to assume a human-prepared index as our ideal involves a conceptual difficulty, dozens of writers have noted a conspicuous lack of consistency among indexers. An alternative approach would be to perform a series of retrieval tests with user provided queries. Such an approach involves a problem of at least equal magnitude as that of inter-indexer consistency—the gathering and evaluation of "relevance judgments." It has been shown that improperly taken relevance judgments can seriously affect measures of retrieval performance (*20,21*). The simplest method of evaluating an automatically produced index is just to compare it to a human—prepared index of the same document, keeping in mind that this norm, while not necessarily "good,"

is the most objective method which can simply be defined.

Formally, we define the human assigned index set S to an abstract d to be the set of all words w_i occurring in d which appear in a comprehensive human-assigned index to the works of Sigmund Freud (*22*) as the first word of an entry to the paper of which d is a summary. Thus, if the index entries assigned to a particular document, as reported in (*22*), were "boys–beating phantasies in," "masculine and feminine," "beating–phantasies," and "phantasies–masochistic," the index set S for that document was taken to be "boys," "masculine," "beating" and "phantasies."

The effectiveness of β_i as a measure of indexability is a function of its success in "retrieving" the members of S from the set of all words in d. An automatically produced index to a document d would perfectly simulate the corresponding human-assigned index S to d, if the members of S were to be ranked at the very top of the β -list.

We are interested not only in how well β functions as an identifier of human-assigned index terms, but also in how β compared to some of the other measures of indexability which have been suggested in the literature. For the reasons outlined in the previous section, we were content to compare indexes prepared by the measure β with indexes prepared by within document frequency k, taking the human-prepared index set S as a norm. The relative success of β and k in "retrieving" elements of S was expressed in terms of recall and precision.

As described in Part I, a computer program, named CALC, was written to calculate estimates of π_i, λ_{1i}, and λ_{2i} for each word type w_i occurring three or more times in the experimental document collection, from the frequency distribution of w_i (*2*). Using these parameter estimates, values of β_i were calculated for each w_i and for several values of k. The output of CALC was in the form of an alphabetically ordered printed list of the 4000 word types. For each word, the value of β_i for $k = 1, 2, \ldots, 6$ was displayed in tabular form. Calculations were not performed for values of k greater than 6 because, in a sample run, $P(d \epsilon I_i / k=6)$ was greater than .999 for all w_i in the sample. Hence for $k > 6$, $P(d \epsilon I_i / k) \approx 1$ for all w_i.

A second computer program, COUNT, was written, using as input a magnetic tape containing the texts of the 650 abstracts in the document collection. The output of COUNT was, for each abstract d, an alphabetically arranged printed list of the words w_i and the number of tokens k of w_i in d.

Thirty-eight documents were chosen at random, using a table of random units. The 38 documents were indexed according to the two measures k_i and β_i. Indexing by k_i was done by simply checking, for each document, the output of COUNT and ranking the words in the document by k_i. To index a document according to the measure β_i, a manual check was performed for each word type in the document. From the output of COUNT, the value of k_i corresponding to w_i in the document was noted. Then, the value of β_i corresponding to w_i and k_i was looked up in the output of CALC and recorded. The words in the document were then ranked by β_i *.

• Results

Results for document number 645 are presented in Table 1. These results are expressed graphically in Figure 1. A recall/precision curve is displayed, reflecting the success of the measures β_i and k_i in "retrieving" members of S.

The overall results for the 38 documents were evaluated in three ways. The first was by means of a sign test. The measure β_i was taken to be superior to k_i as a measure of indexability for a document d if the recall/precision curve belonging to β_i possessed: (i) for some value of recall, a higher value of precision than that for k_i; and (ii) for every value of recall, a value of precision no lower than that for k_i. A similar definition was taken for "k_i is superior to β_i for the document d." According to these criteria, for example, β_i is superior to k_i for document 645. The results of the sign test were: for 26 documents, β_i was found to be superior to k_i; in the remaining 12 documents, neither measure could be judged superior to the other. The binomial probability that this result would have obtained by chance is less than 10^{-7}.

The overall indexing results were evaluated in a second way by calculating the mean value of precision, over the 38 documents, for fixed values of recall, for each of the two measures. These values were read from each of the 38 recall/precision graphs. These composite recall/precision curves are displayed in Figure 2. The curves indicate that measure β_i provides substantially superior retrieval results to k_i, except at the lowest values of recall where the difference between the two measures is not marked.

A final approach to the evaluation of β_i and k_i as measures of indexability is summarized in Figure 3. Figure 3 displays a 95 percent confidence interval on the value of μ_d = precision$_\beta$ - precision$_k$, for fixed values of recall. It was assumed that μ_d is a random variable. For a fixed value of recall and a particular document d, the difference precision$_\beta$ - precision$_k$ was regarded as a sample value of μ_d for that value of recall. A 95 percent confidence interval for each of the several values of recall was then constructed in the usual way. Figure 3 combines these data into a single picture and indicates that μ_d differs significantly from zero at all levels of recall greater than 0.1.

*While the COUNT/CALC lookup procedure was performed manually, it was a purely mechanical operation. It was felt that the use of the computer for this step was not economically justifiable for the purposes of a small experimental test.

Table 1. An Automatic Index for Document 645

k	Word Type	β	Member of S?
4	ego	3.220	x
2	repression	2.495	x
1	superego	2.482	x
1	id	2.406	x
2	impulse	2.307	
2	secondary	2.279	
2	neurosis	2.212	x
2	instinctual	2.033	x
1	satisfaction	1.922	
1	obsessional	1.919	x
1	paranoia	1.882	x
1	narcissistic	1.840	
1	repressed	1.748	x
3	symptom	1.742	x
3	struggle	1.697	
2	behavior	1.690	
1	initial	1.686	
1	world	1.675	
1	forms	1.600	
2	line	1.594	
1	attempt	1.545	
1	symptoms	1.534	x
1	nature	1.503	
1	illness	1.490	
1	external	1.489	x
1	presence	1.459	
1	portion	1.455	
1	representative	1.455	
1	demand	1.410	
1	way	1.403	
1	remains	1.397	
1	fact	1.376	
1	act	1.372	
1	part	1.368	
1	rule	1.365	
1	important	1.355	
1	follows	1.343	
1	direction	1.333	
1	capacity	1.302	
1	comes	1.239	
1	less	1.208	
1	otherwise	1.167	
1	2	1.161	
1	obtain	1.157	
1	defensive	1.155	x
1	followed	1.152	
1	valuable	1.130	
1	adopted	1.118	
1	advantage	1.118	
1	second	1.111	
1	contradictory	1.111	
1	gradually	1.111	
1	presents	1.111	
1	gain	1.104	x
1	prolonged	1.097	
1	springs	1.097	
1	results	1.093	
1	expressions	1.079	
1	friendly	1.079	

Table 1. (Continued)

k	Word Type	β	Member of S?
1	faces	1.079	
1	adopts	1.079	
1	isolated	1.079	
1	organized	1.079	
1	restoration	1.069	
1	interminable	1.069	
1	refuse	1.069	
1	character	1.041	
1	regarded	0.710	

• Summary and Conclusions

The research was prompted by the observation made by a number of writers that non-specialty

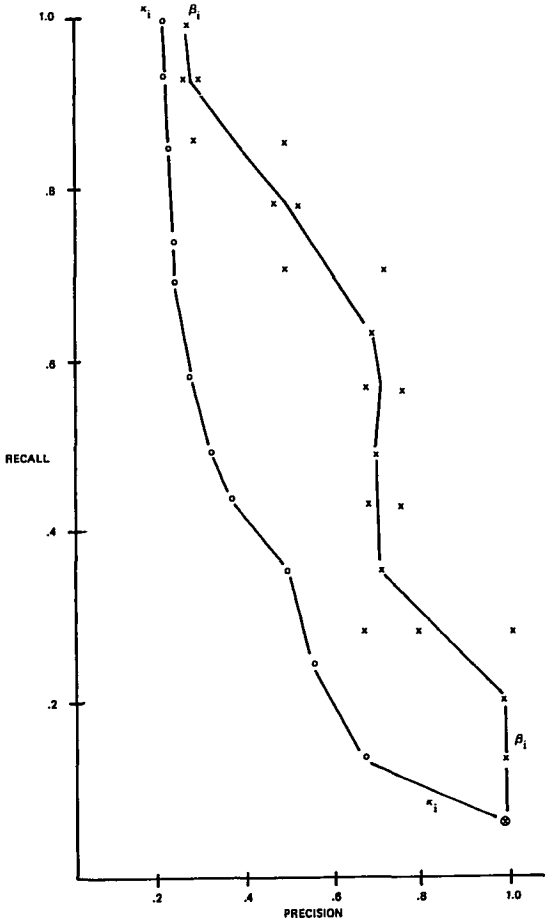

Fig. 1. Indexing Results for Document Number 645

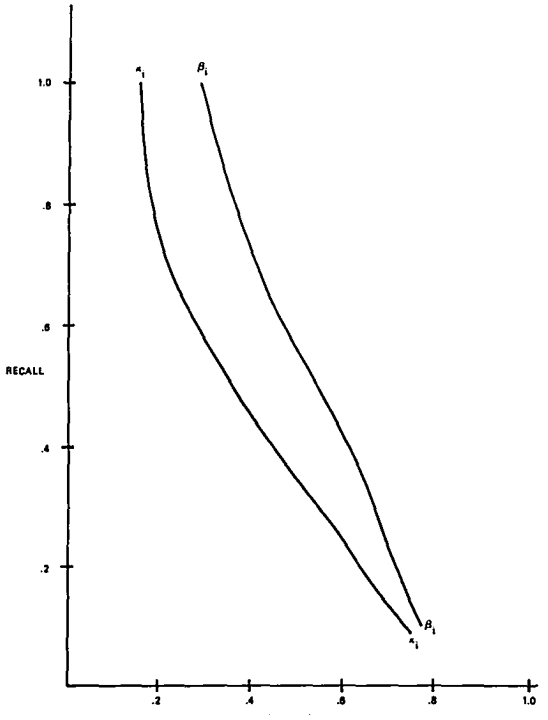

Fig. 2. Composite Indexing Results for 38 Documents.

words, words which possess little value for indexing purposes, tend to be distributed at random in a collection of homogeneous documents. In contrast, specialty words tend not to be so distributed.

In Part I of the study, the 2-Poisson model of specialty word distribution was examined in detail. The 2-Poisson hypothesis assumes that for every document d in a collection, a specialty word w_i is treated at one of exactly two degrees or levels. By assumption, the document collection is thus partitioned into two subsets, I_i and II_i. Tokens of w_i are assumed to occur in I_i and II_i at the rates λ_{1i} and λ_{2i}, respectively. Formulas expressing the parameters in terms of the first three sample moments, calculated from the within document frequency distribution of w_i, were derived. A measure z intended to separate specialty words from non-specialty words, consistent with the 2-Poisson model, was proposed and evaluated.

In the present paper, a probabilistic model of keyword indexing was proposed. The model was shown to imply, with appropriate approximating assumptions, the indexing criterion: Index d by w_i, if and only if

$$P(d\epsilon I_i/k_i) + z_i > 0,$$

where

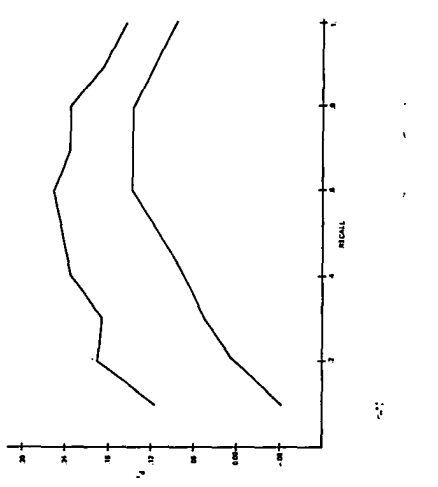

Fig. 3. 95 Percent Confidence Intervals on μ_d, for Selected Values of Recall.

$$P(d\epsilon I_i/k_i) = \frac{\pi_i e^{-\lambda_{1i}} \lambda_{1i}^k}{\pi_i e^{-\lambda_{1i}} \lambda_{1i}^k + (1-\pi_i) e^{-\lambda_{2i}} \lambda_{2i}^k},$$

and

$$z_i = \frac{\lambda_{1i} - \lambda_{2i}}{\sqrt{\lambda_{1i} + \lambda_{2i}}}.$$

Thus, according to the model, *indexability* is a function both of the effectiveness of a word w as a potential index term and of the relative extent to which (w) is treated in the documents in the collection. Our approach to automatic keyword indexing was to rank the word types w_i in a document by the index weights

$$\beta_i = P(d\epsilon I_i/k_i) + z_i.$$

The β-measure was tested by indexing 38 documents and by comparing the results to human-assigned indexes and to indexes compiled by ranking the w_i in each document by k_i, simple within document frequency. The measure β_i was found to produce consistently superior results to those produced by k_i.

Acknowledgment

The U.S. Office of Education supported an early phase of this work under Title II-b, grant number HEW OEG-0-0-230002-1010(320). I am indebted to Donald R. Swanson and Abraham Bookstein for their invaluable contributions to our many discussions. Much of the general philosophical approach that I have followed, as well as several more specific ideas, originate from these dialogues. I would especially like to express my gratitude to Donald R. Swanson, my principal advisor, without whose encouragement and support this project would not have been completed.

References

1. **Bookstein, A.** and **D.R. Swanson,** "Probabilistic Models for Automatic Indexing," *Journal of the American Society for Information Science*, 25 (No. 5): 312-318 (1974).
2. **Harter, S.P.,** "A Probabilistic Approach to Automatic Keyword Indexing: Part I. On the Distribution of Specialty Words in a Technical Literature," *Journal of the American Society for Information Science*, 26 (No. 4): 197-206 (1975).
3. **Swets, J.,** "Information Retrieval Systems," *Science*, 141: 245-250 (1963).
4. **Brookes, B.C.,** "The Measures of Information Retrieval Effectiveness Proposed by Swets," *Journal of Documentation*, 24: 41-54 (1968).
5. **Rothgeb, C.L.** (ed.), *Abstracts of the Standard Edition of the Complete Psychological Works of Sigmund Freud*, Washington, DC: National Institute of Mental Health (1972).
6. **Maron, M.E.** and **J.L. Kuhns,** "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of the Association of Computing Machinery*, 7: 216 (1960).
7. **Mosteller, F.** and **D. Wallace,** *Inference and Disputed Authorship: The Federalist*, Reading, Massachusetts: Addison-Wesley (1964).
8. **Kraft, D.,** "A Decision Theory View of the Information Retrieval Situation: An Operations Research Approach," *Journal of the American Society for Information Science*, 24: 368-376 (1973).
9. **Bookstein, A.** and **D.R. Swanson,** "A Decision Theoretic Foundation for Indexing," *Journal of the American Society for Information Science*, 26 (No. 1): 45-50 (1975).
10. **Lancaster, F.W.,** "MEDLARS: Report on the Evaluation of Its Operating Efficiency," *American Documentation*, 20: 119-142 (1969).
11. **Cleverdon, C.,** "On the Inverse Relationship of Recall and Precision," *Journal of Documentation*, 28 (No. 3): 195-201 (September 1972).
12. **Salton, G.,** "Automatic Text Analysis," *Science*, 168: 335-343 (1970).
13. **Minker, J., E. Peltola** and **G.A. Wilson,** "Document Retrieval Experiments Using Cluster Analysis," *Journal of the American Society for Information Science*, 24: 246-260 (1973).
14. **Sparck Jones, K.,** "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentation*, 28 (No. 1): 11-21 (1972).
15. **Salton, G.** and **C.S. Yang,** "On the Specification of Term Values in Automatic Indexing," *Journal of Documentation* 29 (No. 4) 351-372 (December, 1973).
16. **Sparck Jones, K.,** "Index Term Weighting." *Information Storage and Retrieval* 9: 619-633 (1973).
17. **Luhn, H.P.,** "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development* 1 (No. 4) 309-317 (October, 1957).
18. **Edmundson, H.P.** and **R.E. Wyllys,** "Automatic Abstracting and Indexing—Survey and Recommendations," *Communications of the Association for Computing Machinery*, 4: 226-234 (1961).
19. **Carroll, J.** and **R. Roeloffs,** "Computer Selection of Keywords Using Word-Frequency Analysis," *American Documentation* 20: 227-233 (1969).
20. **Harter, S.,** "The Cranfield II Relevance Assessments: A Critical Evaluation," *Library Quarterly*, 41: 229-243 (1971).
21. **Swanson, D.,** "Some Unexplained Aspects of the Cranfield Tests of Indexing Performance Factors," *Library Quarterly*, 41: 223-228 (1971).
22. **Klumpner, G.** (comp.), *Computer Compiled Cumulative Index of the Standard Edition of the Complete Psychological Works of Sigmund Freud*, Chicago Psychoanalytic Research Group (1970).