Probabilistic Models for Automatic Indexing

Abraham Bookstein
Don R. Swanson
Graduate Library School
University of Chicago
Chicago, Illinois 60637

• Introduction

The purpose of a "document retrieval system" is to select, from a relatively large collecion of documents, a manageable number that is likely to satisfy an expressed need for information. To accomplish this task the system needs some kind of representation of each document in the collection. Associating a set of index terms with each document provides a form of representation that facilitates the creation of an efficient retrieval mechanism. In the process of searching, the system typically computes the degree of match between these terms and a corresponding set of terms derived from a request. This degree of match provides the basis for deciding whether a document should or should not be retrieved. Thus the procedure by which a document is indexed determines the retrieval capability of the system and ultimately the costs and benefits to its users.

The process of indexing is a very complex one, as suggested by the variety of approaches available when this process is automated. Implicit in all of these approaches are assumptions, not always well defined, as to how occurrences of words relate to the content of the documents in which they occur. The purpose of this paper is to develop a model of word occurrences that explicitly relates the subject of documents to the pattern of occurrences of words within these documents. Such a model can then form the basis of an indexing algorithm, but we defer such considerations to later papers; one approach to indexing that uses in part the models developed here is being developed and investigated by Stephen Harter (*1*) in a doctoral study.

This paper is developed in two stages. The first stage describes an experiment that explores properties of the class of words that are not useful in conveying subject meaning and distinguishes them from those classes of words that do convey subject meaning to various degrees. In particular, we study the clustering properties of these words; the analysis is based on statistical properties alone, and techniques are introduced that may be of value in other areas of information science. On the basis of the results of this experiment, a model of word occurrences is introduced and discussed. Later papers by us and by Harter will apply this model to indexing.

Salton (*2*), in his recent review of automatic indexing, notes the lack of linguistic models to guide the planning of automated information retrieval systems. The probabilistic models we present here offer one approach towards such a goal. We further hope that this effort will lead to deeper insight into the question of why some words in text are perceived as being useful as index terms while others are not.

• Background

Some of our results extend, and tend to confirm, similar findings reported by others during the past decade. At a symposium in 1964, Sally F. Dennis (*3*) reported results from a study of 2,649 documents in the law literature. She tested a number of statistical discriminators for distinguishing content from non-content words against the impressionistic judgment of a group of people; a clear correlation is evident. The subsequent work of Damerau (*4*) compares words which he judges to be good index terms with other words in text. The comparison is based on three statistical measures suggested by Edmundson and Wyllys (*5*) and a fourth measure suggested by C.T. Abraham. The latter measure, which was the best, derives from the probability that the observed within-document frequency could be described by a Poisson distribution. Later work by Curtice and Jones (*6*) established that co-occurrence data have valid discriminatory powers: content-bearing words tend to occur in a more constrained environment, having fewer co-occurring words, than do non-content words. Stone and Rubinoff (*7*)

found that the ratio of variance to total frequency was the best discriminator and presented graphs comparing the frequency distribution of specialty words with those of nonspecialty words. Carroll and Roeloffs (*8*) compared several frequency criteria with the judgment of panels of indexers for selection of keywords from documents. They found that the statistical ranking agreed better with the consensus of human judgment than did the judgment of individual indexers. Batty (*9*) has written a brief but informative review of the field.

• Clustering and Indexing

The specific ideas explored in this paper are based on the following observation. Consider a set of several hundred equal length abstracts of scientific or technical articles on various subjects (*10*). Suppose we knew that there were a number of occurrences of some "specialty" word such as "lasers." We might reasonably expect these occurrences to cluster in a relatively few abstracts that are on the topic of lasers; that is, these abstracts should have more occurrences of "lasers" than can be expected if this word were distributed randomly over the collection of abstracts. If we consider on the other hand the occurrence of a non-subject word such as "obtain," the same argument would not apply; for this word it seems likely that the occurrences would be distributed as if they were simply scattered at random among the abstracts. An abstract would not be expected to be about "obtain" in the same way that it might be about "lasers." Thus it is plausible to expect that non-subject words would tend to cluster less than subject words.

From a point of view closer to decision theory, we see that "clustering" might well be intimately related to indexing. If a word is to be indicative of the subject matter of a document, then its occurrences must serve to distinguish that document from those documents not about the subject indicated by that word. However, such a distinction could not be made by a word whose occurrences are distributed randomly among documents in a collection. That is, the occurrence of such a word could not in that case convey any information that would serve to distinguish one document from another; regardless of what the document was about, that word would be just as likely to occur. Thus the property of clustering to a degree exceeding whatever could be accounted for by a random distribution of word occurrences must be intrinsic to the concept of indexing.

To express these ideas in operational terms, if we were to pick out from an entire collection of abstracts or documents those words whose occurrences are more clustered than could be reasonably accounted for by chance, we could then test whether they are more suitable as index terms than those words that exhibit less clustering. Although there are conceptual difficulties in judging the indexing suitability of words, as a practical matter people *do* compile useful indexing vocabularies and in so doing they distinguish words that they believe to be index terms from those that are not. We shall first compare an index vocabulary derived on the basis of the formal statistical criteria developed here with a similar vocabulary based on human judgment as reflected by an independently compiled index; then we shall pursue the ideas expressed above to derive a more general model for word distribution.

• The Occupancy Problem

We shall next define a model in which the various occurrences, or tokens, of a given word are scattered at random into abstracts or documents. We are interested in the extent to which the distribution of words as observed in actual abstracts departs from such a model, the degree of departure to be expressed in terms of a "clustering factor." This clustering factor will be used to establish a *relative* ranking of words for the purpose of discovering a possible relationship to indexing suitability.

Let us consider, then, the following probabilistic model describing how R tokens of a word may be distributed over A documents: we assume a set of processes in which each of the A abstracts independently receives tokens of a given word according to a Poisson process. We run these processes for the length of time needed so that the expected number of terms each document will receive is equal to R/A; the effect of variable abstract length as it might influence the *a priori* probabilities is dealt with more fully in the aforementioned study by Harter. The probability, $P(k)$, that a document receives k terms is accordingly given by:

$$P(k) = \frac{1}{k!}\left(\frac{R}{A}\right)^k e^{-R/A} \qquad (1)$$

While the total number of tokens distributed in all the abstracts might not be exactly R, it is governed by the following Poisson distribution having R as its mean value:

$$Q(K) = \frac{R^K}{K!} e^{-R}, \qquad (2)$$

where $Q(K)$ is the probability that K tokens in all, of the given word, are distributed.

$P(m,k)$, the probability that exactly m abstracts receive k tokens in a total of A "trials," is exactly given by the binomial distribution:

$$P(m,k) = \binom{A}{m} P^m(k)(1 - P(k))^{A-m} \qquad (3)$$

Provided $P(k)$ is small, and $AP(k)$ is of reasonable magnitude, then the standard Poisson approximation to the binomial distribution yields:

$$P(m,k) = \frac{L^m}{m!} e^{-L} \qquad (4)$$

with $L = AP(k) = \frac{A}{k!}\left(\frac{R}{A}\right)^k e^{-R/A}$

Similarly we can derive the exact occupancy distribution; we define $P'(s_0,s_1,s_2,\ldots s_R)$ as the probability that there will be exactly s_l abstracts receiving l tokens. It is given by a multinomial distribution over the A abstracts:

$$P'(s_0,s_1,\ldots s_R) = \frac{A!}{(s_0!s_1!s_2!\ldots s_R!)}\left[e^{-\lambda s_0}(\lambda e^{-\lambda})^{s_1}\left(\frac{\lambda^2}{2!}e^{-\lambda}\right)^{s_2}\ldots\left(\frac{\lambda^R}{R!}e^{-\lambda}\right)^{s_R}\right] \qquad (5)$$

$$= \frac{A!}{(s_0!s_1!s_2!\ldots s_R!)}\left(\frac{R}{A}\right)^R \frac{e^{-R}}{[2^{s_2}(3!)^{s_3}(4!)^{s_4}\ldots(R!)^{s_R}]}$$

where $\lambda = \frac{R}{A}$.

But equation (5) allows for R to vary. The probability, $P(s_0,s_1,\ldots s_R)$, that the words be distributed as above, subject to the condition that

$$\sum_{i=1}^{R} i s_i = R,$$

a fixed number, is given by $P'(s_0,s_1,s_2,\ldots s_R)/Q(R)$, or:

$$P(s_0,s_1,\ldots,s_R) = \frac{R!A!}{A^R(s_0!s_1!s_2!\ldots s_R!)[2^{s_2}(3!)^{s_3}(4!)^{s_4}\ldots(R!)^{s_R}]}. \qquad (6)$$

Equation (6) agrees with the equation derived by Von Mises (*11*) to answer a parallel question for the classical occupancy problem, in which exactly R tokens are distributed at random into A urns. The asymptotic form of the classical problem, for large R and A, leads to the same expression as equation (4). A derivation of this asymptotic form can be found in Feller (*12*). The reason that our results agree is because $Q(R)$ has a mean of R and, for large R, a relatively small standard deviation; thus the final configuration of words resulting from the process we defined above should be very similar to that examined by Von Mises and by Feller; indeed, in the conditional case, leading to equation (6), they should be identical. Our approach may in fact be seen as an alternative and possibly simpler treatment of the occupancy problem discussed by Von Mises and by Feller.

• Measures of Clustering

We have already mentioned our expectation that words suitable as index terms tend not to be distributed as though at random, so it is of interest to define a measure of how far any given distribution of tokens departs from randomness. We could then relate such a measure to the suitability of a word as an index term in the following manner.

If $\tilde{x}$ denotes the random variable that indicates a measure of "clusteredness" for the tokens of a given word, and if we suppose that $g(x)$ is the probability that the degree of clustering resulting from a random distribution would equal or exceed an observed value, x, then this probability can be used as a basis for ranking all words in the vocabulary. We should emphasize that while $g(x)$ can, in some sense, be viewed as an inverse measure of term concentration, this is not its primary function. Rather, for such a measure, $\tilde{x}$, it indicates how unlikely it is that any particular value is achieved or exceeded by a given word. It is quite possible that two words have identical values on a clustering measure, but, because of values of other parameters, in one case that degree of clustering can be explained as a chance occurrence while in the other it cannot. Those words with the smallest value of $g(x)$ would be most "remote" from a random distribution.

A number of choices for $\tilde{x}$ are available. We shall here consider two of them, denoted by $\tilde{x}^I$ and $\tilde{x}^{II}$. Probably the simplest plausible form of $\tilde{x}$ can be taken to be:

$$x^I = s_2 + 2s_3 + 3s_4 + \ldots = R - (A - s_0). \qquad (7)$$

That is, this choice of $\tilde{x}$ is simply the difference between the total number of tokens and the total number of abstracts which they occupy. A possible criticism of the above measure is that it would not be sensitive to what may be a very significant tendency for a few documents to "attract" a large number of tokens; a configuration in which one document contained five occurrences and three documents contained only one occurrence of some given word would produce the same value for the measure as four documents containing two occurrences each. An alternative measure correcting what would seem, intuitively, to be a weakness is suggested by the following analogy.

Suppose we imagine that the various occurrences of a word have a tendency to "attract" each other, in a way loosely analogous to the forming of chemical bonds. If so, we can think of a "bond" as connecting each pair of occurrences of the same word within a given abstract. In that event, the choice

$$x^{II} = \sum_j \binom{j}{2} s_j = s_2 + 3s_3 + 6s_4 + \ldots$$

representing the total number of bonds occurring among all tokens of a particular word, would appear to be a reasonable measure of clustering.

We shall explore further the choices $\tilde{x}^I$ and $\tilde{x}^{II}$ as given above.

• Word Distribution in Abstracts

The text selected for calculations of word distributions consisted of the *Abstracts of Standard Edition of the Complete Psychological Works of Sigmund Freud*, edited by Carrie Lee Rothgeb, published by the National Institute of Mental Health. The average length of an abstract is around 250 words, with 90% falling in the range of 90 to 330 words. The total number of abstracts is 650.

Since words with very few occurrences could not be expected to "cluster" even if there were a tendency to do so, the subset of words with 14 or more, but fewer than 125, occurrences was selected for study. This subset consisted of 990 words, each form of each word—including singular/plural distinctions—being separately counted.

The text of the 650 abstracts, and a word index to the text, were available on magnetic tape. The index did not include 220 frequently occurring words which were used as a "stop-list"; this latter group of words was not studied. The stop-list was compiled prior to, and independently of, this study; the words were selected on the basis of human judgment as being useless as indicators of subject content. Many words *not* on the stop list also give the same impression of uselessness for indexing; these are focal to our study.

From the index tape, the occupation distribution $(s_0, s_1, s_2, \ldots s_R)$ for each word was computed, each such distribution corresponding to a value of $\bar{x}^I$ and $\bar{x}^{II}$. A table, computed from equation (6), was prepared, showing all distributions for which $P(s_0, s_1, s_2, \ldots s_R)$ was greater than 10^{-8}. The set of probabilities thus obtained was sorted in increasing order of the x's, and these values were then cumulated. Thus, each entry in the resulting table represented the probability of realizing, on the basis of a random distribution of tokens, any given value of $\bar{x}^I$ and $\bar{x}^{II}$ or higher; that is, the probability of a greater degree of clustering occurring as the result of a random distribution of tokens.

To illustrate, consider the word "CHARACTER," which occurred 75 times in the 650 abstracts. There were 57 single occurrences ($s_1 = 57$), six double occurrences ($s_2 = 6$), two triple ($s_3 = 2$), and no instances of more than three occurrences per abstract. Thus $\bar{x}^I = 10$. The probability that the exact distribution 57, 6, 2, 0, 0 .... would occur by chance, given that the total number of occurrences is 75, is .00049975. The probability that $\bar{x}^I$ will have the value 10 (which can be realized by any of the following sets of s_2, s_3, s_4: 8:1:0, 10:0:0, 6:2:0, 7:0:1, 4:3:0, 5:1:1, 3:2:1, 2:4:0, 4:0:2, 2:1:2, 1:3:1, 0:5:0, 0:2:2, 1:0:3) is .0036755. The probability corresponding to s_5 and higher is negligible. The cumulated probability $g^I(10)$ is .00509173, where the superscript, I, indicates that $\bar{x}^I$ is the random variable being considered.

Tables of these probabilities, computed from equation (6), were prepared for $A = 650$ and for most values of R between 4 and 100. Similar computations were carried out for $\bar{x}^{II}$ and $\bar{x}^I$.

The difficulty of computing $g(x)$ prompted us to seek a simplification. The simple expression, given in (*16*), can be compared with the various arbitrary measures investigated by others, for example the nine measures computed by Dennis. However the measures we use derive from the probability of attaining a certain cluster value, rather than from the cluster value itself; in this respect it differs from the various cluster measures investigated by Dennis and by others.

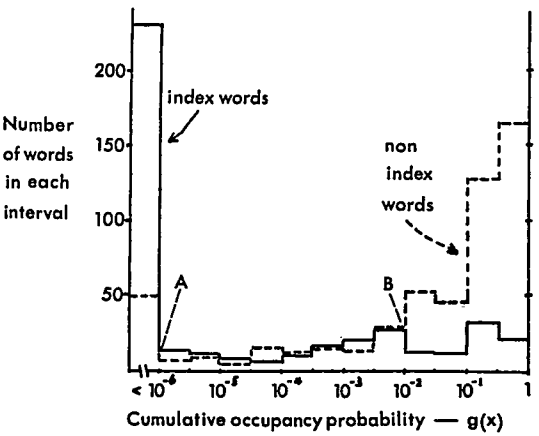
• Comparison with Published Index

It was of interest next to determine whether the clustering tendency was related to some kind of impressionistic judgment of "subject-indicativeness." The problem of course was to obtain, in some way, human judgment as to the suitability of each word for an indexing vocabulary. For the purposes of this research we considered that such judgment was implicitly available in the form of the manually prepared index to each of the twenty-three volumes of the *Standard Edition* of Freud's works. The use of these indexes has been facilitated by the recent compilation, by Dr. George H. Klumpner (*13*), of a single cumulative index (*14*). Each word on the aforementioned list was then checked in order to ascertain whether it appeared as an entry-word in the cumulative index (*15*). The criterion by which words are judged as being good index terms is simply their appearance in a published index. Thus, subjective judgments by project personnel are not utilized, but an external and independent source is chosen. Although such choices and judgments were made without knowledge of actual word-occurrence distributions in text striking regularities are discernible. Such a criterion for "suitable" index terms does not permit direct inferences as to "retrieval effectiveness," since no retrieval experiments were performed. In the last analysis, of course, retrieval effectiveness must be the basis of judgment of good indexing. Yet published indexes have been used and found useful for more than a century, and we think it of considerable interest to demonstrate a statistical model which more or less reproduces the same end result as does human judgment in distinguishing index terms from non-index terms.

As a result of the aforementioned lookup process, 440 words out of the 990 were determined to be entry words in the cumulative index. Singular-plural variants were accepted as equivalent, but otherwise different forms were taken as distinct words. The words that appeared as entry terms in the cumulative index will be referred to here as "index words," and all others as "non-index words."

As described in the preceding section, the values of $g(x)$ were computed for each word for each of the cluster measures $\bar{x}^I$ and $\bar{x}^{II}$.

The graph in Figure 1 shows separately the distribution of index words and non-index words over the values of the cluster probability, $g(x)$. The difference between the two groups is striking. A similar graph for $\bar{x}^{II}$ is not shown since it differs negligibly from that for $\bar{x}^I$. Note that one can divide the abscissa of Figure 1 into three portions, such that the left-most portion contains about five times as many index words as non-index words, the right contains five times as many non-index words as index words, and the center about equal numbers of the two. The central portion, between the points *A* and *B* of the figure, contains approximately 25% of the total number of words (*16*).



Distribution of Cumulative Occupancy Probabilities For Index Words and for Non-Index Words

Clearly index words tend to be significantly more "clustered" than non-index words. This result confirms and extends similar findings of Dennis (*3*), Damerau (*4*), Curtice and Jones (*6*), Stone and Rubinoff (*7*), Carroll and Roeloffs (*8*), and others.

The implication of the dissimilarity between the two distributions shown in Figure 1 is that one can specify a formal or computable criterion for distinguishing between index and non-index words. This criterion can be expressed in terms of a "cutoff" value of $g(x)$, below which a word is taken to be a suitable index term and above which it is not. The fact that the two distributions overlap implies that any cutoff will be imperfect. For example, if all words for which $g(x) < 10^{-2}$ are selected as index entries, then 69% of those so chosen will be actual index terms in the particular experiment reported, and of all actual index terms on the list, 82% will have been selected. There is of course a complementary relationship between these two measures as the cutoff is varied. At a cutoff of 10^{-5}, the two numbers are 79% and 60%, respectively. Note that in any group of words selected at random from the list, 440/990 or 44% would be expected to be actual index terms.

Without going quite so far as to suggest that people actually judge the usefulness of a word for indexing purposes by means of some hidden intuitive sense of its statistical clustering properties, we note that the vocabulary selected for a published index bears a remarkable resemblance to a vocabulary compiled purely on the basis of such statistical properties. A case might eventually be made for *defining* indexing suitability on the basis of statistical clustering criteria, and so dispensing with human judgment altogether in compiling indexing vocabularies.

• A Model for Index Term Distribution

It may be inferred from Figure 1 that most words are much more clustered than would result from being placed at random among all 650 abstracts. We next propose a model for word distribution that generalizes the pure random model and which might be able to account for the observed distribution of content bearing words.

We have already noted the relation between a word's indicating content and that word tending to cluster. For if the word does cluster, then its appearance in a document gives us some information that allows us to distinguish that document from other documents in a manner soon to be made precise. Only if the word occurs randomly can we draw no conclusions from an appearance of the word in a document; in that case the documents are homogeneous with regard to that word.

The model we shall now introduce extends these ideas to words that do distinguish classes of documents, so that the collection of documents is no longer homogeneous with regard to these words. We shall assume, however, that the documents can be broken up into subclasses of documents such that each subclass *is* homogeneous with respect to the appearance of a word. In that case, if we restrict ourselves to a single subclass, the appearance or lack of appearance in a document of the word defining the subclass is attributable to random fluctuation and conveys no information distinguishing that document from other documents in that subclass. However, the various subclasses are distinguished by occurrences of the word.

We shall here associate the various degrees to which the classes tend to attract a word with the various degrees to which the documents in these classes are about that word. This tendency to attract words is a latent property in our model in that it cannot be measured directly. It does however have measurable consequences, which we shall explore in this and ensuing papers.

The first consequence of this model is that we can generalize the earlier distribution formulas for word

occurrences so as to include content-bearing words; we now assume that for documents in class i words arrived in a Poisson stream, with λ_i the expected number of words in a document. If we knew that a document is in class i, then the probability that there be k occurrences of that word is given by

$$\frac{\lambda_i^k}{k!} e^{-\lambda i} ,$$

the familiar Poisson distribution. In fact, we do not know to which class a document belongs; let us assume that π_i is the probability that it belongs to class i. Then $f(k)$, the probability that a randomly chosen document contain k occurrences of the word, is given by

$$f(k) = \sum_i \pi_i \frac{\lambda_i^k}{k!} e^{-\lambda i} \tag{8}$$

In the earlier model we had only a single class, $\pi = 1$, and $f(k)$ was accordingly given by a single Poisson distribution; thus we see that the present model is a generalization of the first model.

We can now state more precisely how occurrences of a word in a document give information regarding the class to which that document belongs. Before knowing how often the word w occurs in a document, the probability that the document belongs to class i is π_i. If we know that the document contains k occurrences of w, then the probability that the document belongs to class i, denoted by $Pr\{i|k\}$, is given by

$$P_r\{i|k\} = \frac{\pi_i \lambda_i^k e^{-\lambda i}}{\sum_i \pi_i \lambda_i^k e^{-\lambda i}} i ; \tag{9}$$

this is an instance of Bayes' theorem. The effect of knowing how many times w occurred in a document is thus to modify the initial probability that the document belongs to a given relevance class with regard to w.

To relate the relevance classes to relevance judgments, we shall need parameters giving the probability that a patron requesting documents about w will find a document in class i relevant; the idea that relevance judgments are essentially probabilistic is very much in the spirit of the classic paper by Maron and Kuhns (*17*). We denote the probability that a document in class i will be found relevant by r_i. Combining the above results implies that if a person requests documents about w and is given a document with k occurrences of w, then the probability that the document is found to be relevant is given by

$$P_w(k) = \sum_i r_i P_r\{i|k\} = \frac{\sum_i r_i \pi_i \lambda_i^k e^{-\lambda i}}{\sum_i \pi_i \lambda_i^k e^{-\lambda i}} \tag{10}$$

Thus the observable assessments of relevance can be predicted by this model. Similarly this model can be used to predict the recall and precision of search strategies based upon word occurrences.

To proceed further, it is necessary to specify in greater detail the number of relevance classes and the values of the parameters. Several models were considered.

The simplest model that appears promising, and which has been tested by S. Harter, can be referred to as the "two Poisson model." We here assume that the distribution of words is given by

$$f(k) = \pi \frac{\lambda_1^k}{k!} e^{-\lambda_1} + (1-\pi) \frac{\lambda_2^k}{k!} e^{-\lambda_2} \tag{11}$$

Harter evaluated the parameters π, λ_1, and λ_2 by fitting this distribution to the data by the method of moments, and found that the resulting fit was good for a large majority of the words. The assumption behind this model is that the documents being investigated can be divided into two classes, those about a topic and those peripherally about the topic. The degree to which a document is about the topic determines the magnitude of λ_i.

Examination of words that the above model does not adequately describe suggests that a somewhat more realistic model might divide the documents into three classes:

1. those unrelated to the word (association strength zero), and for which the average number of occurrences of the word is negligibly small;
2. those which deal peripherally with the subject or concept named by the word (association strength small);
3. those which deal *centrally* with the subject or concept named by the word (association strength large).

This particular model, with four parameters, $(\pi_1, \pi_2, \lambda_1, \lambda_2)$ offers some computational difficulty, but appears promising enough to deserve further investigation.

The multiple Poisson distribution may, of course, be generalized to a continuous association strength, $\pi(\lambda)$. The negative binomial distribution, which was examined by Mosteller and Wallace (*18*), is of this form for a suitably chosen $\pi(\lambda)$.

• Summary

In this paper we studied the pattern of occurrences of words in text as part of an attempt to develop formal rules for identifying those indicative of content and thereby suitable for use as index terms. The significance of this work lies in its potential use in more efficient fully automatic retrieval systems, and possibly even for providing insight into the intellectual process of indexing.

This work begins with the observation that content-bearing words are concentrated in fewer documents

than non-content-bearing words. This concept is made more precise and tested in a corpus made up of abstracts to the works of Freud. It was found that most non-content words have distributions much closer to what would result from a random process than is the case for words useful for indexing.

A probabilistic model was proposed which, with a suitable fitting of parameters, could account for the occupancy distribution of most words, both index terms and non-index terms. The parameters take quite different values for the two classes. In this model each abstract was considered to receive word occurrences in a Poisson process. Abstracts can then be divided into classes, such that all abstracts within a given class receive word occurrences at the same average rate. The appearance of a particular number of occurrences of some word within an abstract then serves to give information, in a Bayesian sense, on the class membership of that abstract.

It is of central interest to determine the minimum number of classes that can account for the occupancy distribution of each word. Though more testing needs to be done it may be concluded that the distribution of a very large majority of words can be accounted for by assuming three or fewer classes.

• Acknowledgment

An early phase of this work received financial support of the National Science Foundation, Office of Science Information Service, grant GN 380 "A Study of Indexing Depth and Retrieval Effectiveness." We are grateful to Dr. George H. Klumpner for helpful discussions and for making available the cumulative index and other materials. Our many discussions with S. Harter concerning his doctoral research were valuable to us; in particular his observations on the dual use of a word, as informing and non-informing, were helpful in leading to the two-Poisson model proposed in this paper.

Notes and References

1. **Harter, S.**, Ph.D. dissertation in preparation, Graduate Library School, University of Chicago.
2. **Salton, G.**, "Automatic Text Analysis," *Science*, 168: 335-343 (1970).
3. **Dennis, S.F.**, "The Construction of a Thesaurus Automatically from a Sample of Text," (in) *Statistical Association Methods for Mechanized Documentation*, (ed.) **Mary E. Stevens** *et al.*, Washington, D.C.: National Bureau of Standards Miscellaneous Publication 269 (1965).
4. **Damerau, F.J.**, "An Experiment in Automatic Indexing," *American Documentation*, 16(No. 3): 283-289 (1965).
5. **Edmundson, H.P.** and **R.E. Wyllys**, "Automation Abstracting and Indexing—Survey and Recommendation," *Communications of the ACM*, 226-234 (1961).
6. **Curtice, R.M.** and **P.E. Jones**, "Distributional Constraints and the Automatic Selection of an Indexing Vocabulary," *Proceedings of the American Documentation Institute Annual Meeting*, 4: 152-156 (1967).
7. **Stone, D.C.** and **M. Rubinoff**, "Statistical Generation of a Technical Vocabulary," *American Documentation*, 19(No. 4): 411-412 (1968).
8. **Carroll, J.M.** and **R. Roeloffs**, "Computer Selection of Keywords Using Word-Frequency Analysis," *American Documentation*, 20(No. 3): 227-233 (1969).
9. **Batty, C.D.**, "Automatic Generation of Index Languages," *Journal of Documentation*, 25: 142 (1969).
10. The theoretical argument applies as well to documents as to abstracts, but the experimental corpus later to be discussed consists of abstracts.
11. **Von Mises, R.**, "Uber Aufteilungs-und Besetzungs-Wahrscheinlicheiten," *Revue de la Faculté des Sciences de l'Université d'Istanbul*, N.S., 4: 145-163 (1939).
12. **Feller, W.**, *Introduction to Probability Theory and Its Applications*, Chap. IV (1968).
13. **Klumpner, George H.**, M.D., (Comp.) *Computer Compiled Cumulative Index of the Standard Edition of the Complete Psychological Works of Sigmund Freud*, Chicago Psychoanalytic Indexing Research Group.
14. The fact that the *cumulation* was performed by computer should not obscure the point that each term in the index represents *human* judgment as to indexing suitability.
15. The inherent suitability of a word, irrespective of the context of any particular use, is what its appearance in the cumulative index is presumed to indicate. Thus the fact that the indexes referred to the *text* of Freud's works, whereas our word list to be tested came from the *abstracts*, is no obstacle unless the vocabulary of the abstracts differs substantially from the vocabulary of the text. In view of the results to be discussed, any such differences did not seem to be important.
16. A ranking of words based on an approximate clustering factor $\frac{N_A \cdot N_0}{\sqrt{N_0(1-N_0/A)}}$, where N_A is the total number of occupied abstracts and N_0 is the expected number in a random distribution, was also computed and led to very similar results with respect to the division of the word list into three portions. This factor is of course much simpler to compute than is the exact cumulative probability; it is suggested by the normal approximation to a binomial distribution.
17. **Maron, M.E.** and **J.L. Kuhns**, "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of the Association of Computing Machinery*, 7: 216 (1960).
18. **Mosteller, F.** and **D. Wallace**, *Inference and Disputed Authorship: The Federalist*, Reading, Massachusetts: Addison-Wesley (1964).