Automatic Indexing: An Experimental Inquiry*

M. E. MARON

The RAND Corporation, Santa Monica, California

Abstract. This inquiry examines a technique for automatically classifying (indexing) documents according to their subject content. The task, in essence, is to have a computing machine read a document and on the basis of the occurrence of selected clue words decide to which of many subject categories the document in question belongs. This paper describes the design, execution and evaluation of a modest experimental study aimed at testing empirically one statistical technique for automatic indexing.

1. *Introduction*

1.1. *Some General Remarks Concerning Automatic Indexing.* The term "automatic indexing" denotes the problem of deciding in a mechanical way to which category (subject or field of knowledge) a given document belongs. It concerns the problem of deciding automatically what a given document is "about".

The situation is one in which there is a collection of different documents, each containing information on one or several subjects. Also there exists a set of categories, usually not exclusive nor completely independent, but (we hope) exhaustive in the sense that every document will "fit" into at least one of the given categories. The problem arises because the categories are not defined extensionally. That is to say, a category is not determined by enumerating each and every one of those documents which make up its membership, but, rather, the situation is reversed. Based on some more or less clear notion of the category, we must decide whether or not an arbitrary document should be assigned to it.

To correctly classify an object or event is a mark of intelligence; a mark of even greater intelligence is to be able to modify and create new and more fruitful categories, i.e., to form new concepts. (Perhaps one of the really dominant characteristics of an intelligent machine will be that of creating new categories into which to sort its "experiences".)

Loosely speaking, it appears that there are two parts to the problem of classifying. The first part concerns the selection of certain relevant aspects of an item as pieces of evidence. The second part of the problem concerns the use of these pieces of evidence to predict the proper category to which the item in question belongs. But before examining this way of looking at the problem, let us consider in more detail the problem of classifying *linguistic* entities on the basis of what they *mean* as opposed to the problem of classifying things in general.

1.2. *The Problem of Indexing Information.* It is because words and sentences stand for other things (i.e., they are one step removed from the things and events that they describe) that the problem of indexing information is made even more complex than the problem of classifying nonlinguistic entities. More

* Received, January 1961.

specifically, the problem is so complex not only because words and sentences are one step removed, but also because there is no one-to-one relationship between the individual words and the events that the sentence containing those words describes. Ordinary language does not have explicit rules which prescribe how words and sentences are to be put together to convey various kinds and shades of meaning. There are no known rules which describe those combinations of sentences which refer to, say, thermodynamics or music or switching theory. Nevertheless, we are able to understand the meanings of documents for the most part, and we do manage to index and classify information with a fair degree of success. However, the degree of success is clearly a function of the intelligence and experience of the individual who is doing the indexing.

The problem of indexing brings us to the door of semantics and with it come all of the difficulties involved in an analysis of the concept of meaning. Nevertheless, let us examine a technique of automatic indexing which allows us to by-pass many of the problems concerning meaning and those grammatical functions of language which help to convey specific meanings. This statistical technique involves (1) the determination of certain probability relationships between individual content-bearing words and subject categories, and (2) the use of these relationships to predict the category to which a document containing the words belongs.

1.3. *Implications of Automatic Indexing.* At the risk of getting ahead of ourselves and in view of the obvious information explosion that our scientific and intelligence communities surely face, let us point out what successful automatic indexing could mean. First, we seem to be rapidly approaching the time when along with the printed page there will be an associated tape of corresponding information ready for direct input to a computing machine. This means that as each organization receives its daily incoming documents a machine could read them and route them directly to the proper users. The users could describe their information needs in terms of "standing" requests and on the basis of these a machine could determine how the incoming "take" should be disseminated. Since automatic dissemination is only a special aspect of a mechanized library system, it follows that automatic indexing also would allow incoming documents to be indexed and thus identified for subsequent retrieval.

2. *Method of Attack*

2.1. *Basic Notions.* This approach to the problem of automatic indexing is a statistical one. It is based on the rather straightforward notion that the individual words in a document function as clues, on the basis of which a prediction can be made about the subject category to which the document most probably belongs. The fundamental thesis says, in effect, that statistics on kind, frequency, location, order, etc., of selected words are adequate to make reasonably good predictions about the subject matter of documents containing those words. We do not consider the meanings of individual words—we only look to see what subjects they describe and from such data make predictions. Clearly this is a

very simplfied approach and one which completely ignores the ways in which certain types of words combine to convey information. Nevertheless this approach appears to be a practical attack on the problem of automatic indexing, as this modest experiment will indicate.

2.2. *Words and Predictions.* Given this approach to automatic indexing, two problems present themselves, viz., the selection of clue words and the prediction techniques relating clue words and subject categories. Concerning the selection of clue words, how shall we decide which words convey the most information, how many different words should be used, etc.? Clearly, certain content-bearing words such as "electron" and "transistor" are better clues than logical type words such as "if", and "then", etc. On the other hand, those unambiguous content-bearing words that occur very rarely are inefficient clues simply by virtue of their rarity. Is there a systematic way of selecting a best class of clue words? Again, in the case of the methods of prediction, we see that there are many kinds of relationships that exist between clue words and subject categories. The selection of an optimal prediction method also involves a number of difficult problems.

In order to clarify matters before considering these questions in more detail, consider the following way of talking about clue words and prediction methods for automatic indexing. The basic objects in this universe of inquiry are the class of documents under consideration. These objects (documents) have properties, viz., the clue words that they contain. The properties are the observables in our universe and we take measurements on them. Thus a measurement is a list of the kinds of properties that an object has. (More sophisticated measurements would provide information about the frequency, distribution, order, etc., of the properties of our objects.) The information supplied by the measurement when properly formulated constitutes the evidence. For example, an evidence statement might assert: "The document D_1 contains clue words W_1, W_5, and W_{11}." Statistical data relating clue words and subject categories constitute hypotheses. Here again, we have several levels of hypotheses—each more complex and sophisticated than the next. Finally, on the basis of the evidence and the hypotheses we can make *predictions* of the following kind: "The document D_1 belongs to the category C_7 with a probability p." Given the basic notions, let us now see what is involved in testing this statistical approach to automatic indexing.

2.3. *The Empirical Test.* First a corpus of documents was selected and indexed using a set of subject categories created for the purposes of the experiment. Then a subclass of those words that occurred in the original corpus was selected. These constituted the clue words. Once the documents were "sorted" into their respective categories, the statistical correlation between the clue words and the subject categories was determined—that is to say, a clue word vs. category matrix was obtained in which the entry i, j represented the number of ith clue words that appeared in those documents, which were indexed under the jth category. Another and different class of documents was obtained and using the statistical data gathered initially, a machine was programmed to index automatically the documents in question. The design, execution, results and evaluation of this test are examined in the following sections.

3. *Nature of the Corpus*

3.1. *First Remarks.* The first problem was that of selecting a suitable collection of documents on which to carry out the experiments. One would prefer short, clearly written, interesting documents and preferably ones where the range of the subjects described by the documents is not too heterogeneous. That is to say, if some of the documents discussed music and others discussed glass blowing, painting, thermodynamics, etc., it would be rather easy to automatically index them since the subjects are so "far" apart. To adequately test the proposed method the degree of discrimination required should be relatively difficult.

A suitable collection was found in the March, 1959, issue of the *IRE Transactions on Electronic Computers.*[1] Starting in that issue the PGEC inaugurated a new literature abstracting service and carried more than one hundred abstracts of current computer literature. (See Appendix A for a typical abstract.) Thus, the corpus contained abstracts (hereafter referred to as "documents") in the computer field in general, but on a wide variety of different aspects of computer design, applications, theory, mathematics, components, programming, etc.

3.2. *Some Statistics.* The complete corpus consisted of 405 different documents and was divided into two groups which we called group 1 and group 2. Group 1 contained those 260 abstracts that appeared in the March and June issues of the 1959 PGEC and it was the basis for the statistical data necessary to make the subsequent predictions. Group 2 consisted of those 145 abstracts which appeared in the September 1959 issue of the PGEC and was not even looked at during the preparatory phase of the experiments. Once the data from group 1 had been collected, the notions were tested by having a machine index the documents of group 2.

Every word in each of the 260 documents of group 1 was key punched. There was a total of over 20,000 word occurrences and the average number of words per document was 79. There were 3,263 different words contained in the documents of group 1—ranging from words that occurred only once to words that occurred several hundred times.

4. *The Categories*

4.1. *Initial Remarks.* Given the corpus, the next step was to index the documents so as to determine how the clue words and subject categories were correlated. Here again a number of questions arise, e.g. how many categories should there be; how fine should the discrimination within subjects in the general field of computers be, etc. The PGEC had provided with its abstracts a classification system of 10 major categories, five of which were subdivided into about three subcategories each. However, it was decided to use a different and finer set of 32 categories. (The list of 32 subject categories is shown in Appendix B.)

4.2. *Some Statistics.* Once the list of 32 categories was created, each one of the 260 documents was carefully read and "sorted" into one or more of the categories. In the majority of instances a document was indexed under a single cate-

[1] *IRE Transactions on Electronic Computers*, Vol. EC-8, No. 1. Published by the Professional Group on Electronic Computers.

gory, but in about 20% of the cases a document was indexed under two categories and, in only a few cases was it necessary to index a document under three categories. In no case did a document fall into more than three categories. There were 37 documents indexed under the most "popular" category and as few as two in the least "popular" category.

5. *Problem of Clue Word Selection*

5.1. *Initial Remarks.* Some simple rules were formulated to help the key punching. For example, any group of words in quotes was considered a single word. Any expression containing a hyphen was considered a single word (e.g., "analog-to-digital"). The singular and plural forms of a word (e.g., "circuit" and "circuits") were considered two distinct words. Two different spellings of the same word (e.g., "analogue" and "analog") were considered different words, etc.

The 55 most frequently occurring logical type words (e.g. "the", "of", "a", etc.) accounted for 8,402 of the total (20,515) occurrences. Thus, less than 2 per cent of the words accounted for over 40 per cent of the total occurrences. Clearly these words in isolation provide no information about the subject content of a document and they were immediately rejected as candidates for the clue word list. The most frequently occurring nonlogical type words were considered next. This list contained words such as "computer", "system", "data", "machine", etc. They also were rejected as possible clue words because it was felt that they were too "common" to be clues for the specification of subject content within the general field of computers.

All those words that appeared only once or only twice in the entire corpus were then listed. Of the total 3,263 different words, 2,120 or 65 per cent occurred less than three times in the 260 documents. Although they might be good clues to indicate the content of a document, to use them as clues would be inefficient because they occur so rarely. (Furthermore, because of the small numbers, statistics on them would be unreliable.) Thus, they also were rejected as possible clue words. This left just over 1,000 different words—words with neither a very high nor very low relative frequency of occurrence.

A listing was made showing the number of times each of these 1,000 words occurred in those documents belonging to category 1, category 2, etc. In this way one could scan the list and for each word see whether or not it "peaked" in any of the 32 categories. If a word did peak it was felt that the word would be a good clue. If the distribution was flat for a given word (i.e., it did not have a peak in at least one category), then it was rejected as a good clue.

5.2. *Clues and Predictability.* Let us digress for a moment to consider the criteria of adequacy for a clue word. Clearly one such criterion should be that the word leads to a good prediction of the correct subject content of a document. Consider, prior to looking at any of the words in a document, the uncertainty that exists as to the category to which the document in question belongs. This uncertainty is represented by the so-called a priori probability distribution of

the categories and can be measured by Shannon's expression for entropy; viz.,

$$H = -\sum_{j=1}^{32} P(C_j) \log_2 P(C_j)$$

where $P(C_j)$ is the a priori probability that an arbitrary document will be indexed under the jth category. Given that a particular word, say W_i, does occur in a document, by how much does the initial uncertainty (as represented above) change? The new uncertainty is represented by the following expression

$$H' = -\sum_{j=1}^{32} P(C_j \mid W_i) \cdot \log_2 P(C_j \mid W_i),$$

where $P(C_j \mid W_i)$ is the probability that if the ith word occurs in a document, the document belongs to the jth category. Thus the amount of uncertainty that is removed can be determined by the difference between H' and H. Therefore, given two words W_1 and W_2, W_1 is a better clue if its occurrence in a document removes a greater amount of the initial uncertainty than would the occurrence of W_2. (But one must consider, in addition, the a priori probabilities for W_1 and W_2.)

If we begin to consider the dependence relationships between individual words, these considerations would soon become very involved as would the corresponding computations necessary to provide an ordered list of words ranked by their "goodness" as clues. No such sophisticated considerations were given. Rather, by inspection a word was admitted to the clue word list if its distribution peaked in at least one category. An attempt was made to find at least one word to peak in each of the 32 categories. In this way, 90 different words were selected and these constitute the class of allowable clue words. (See Appendix C for the list of clue words.)

6. *The Prediction Method*

6.1. *First Remarks.* Again, the type of inference with which we are concerned is the following: A document is selected and a machine looks to see which of the selected clue words are contained in that document. On the basis of the occurrence of the clue words, the computer makes an inference as to the subject category to which the document in question belongs. Thus, the inference is inverse transition from evidence to hypothesis; the calculus of probability provides us with the proper schema for computing the values. In order to clarify the nature of the inference, consider the case where a document, say D_1, contains one and only one of the clue words, say W_1. Given W_1, what is the probability that D_1 belongs to each of the categories C_1, C_2, C_3, $\cdots$, C_{32}? Its value is computed according to the following expression:

$$P(C_j \mid W_i) = \frac{P(C_j) \cdot P(W_i \mid C_j)}{P(W_1)} \tag{1}$$

$P(C_j)$ is the so-called a priori probability that a document will be indexed under

the jth category and $P(W_1 \mid C_j)$ is the probability that if a document is indexed under the jth category it will contain word W_1. For any W_1, the denominator of (1), $P(W_1)$, is a constant and hence (1) may be rewritten as follows:

$$P(C_j \mid W_1) = k \cdot P(C_j) \cdot P(W_1 \mid C_j) \tag{2}$$

where k is a scaling factor.

6.2. *Independence and Exclusiveness.* In the case where a document has two different clue words, say W_1 and W_2, then the inference schema is the following:

$$P(C_j \mid W_1 \cdot W_2) = \frac{P(C_j) \cdot P(W_1 \mid C_j) \cdot P(W_2 \mid C_j \cdot W_1)}{P(W_1) \cdot P(W_2 \mid W_1)}. \tag{3}$$

Again the denominator is constant and (3) can be rewritten as follows:

$$P(C_j \mid W_1 \cdot W_2) = k \cdot P(C_j) \cdot P(W_1 \mid C_j) \cdot P(W_2 \mid C_j \cdot W_1). \tag{4}$$

Assuming that, relative to a given category, any two clue words are independent, then (4) reduces to

$$P(C_j \mid W_1 \cdot W_2) = k \cdot P(C_j) \cdot P(W_1 \mid C_j) \cdot P(W_2 \mid C_j), \tag{5}$$

where k is another scaling factor. Clearly this independence assumption is false in the sense that

$$P(W_k \mid C_j \cdot W_j) \neq P(W_k \mid C_j); \tag{6}$$

nevertheless, to facilitate (although degrading) the computations, we can make the independence assumption.

Thus, in the general case where a document contains different clue words, $W_k, W_m, \cdots, W_s$, compute the probability that the document belongs to the jth category as follows:

$$P(C_j \mid W_k \cdot W_m \cdot \ldots \cdot W_s) = k \cdot P(C_j) \cdot P(W_k \mid C_j) \cdot P(W_m \mid C_j) \cdot \ldots \cdot P(W_s \mid C_j). \tag{7}$$

Call the values of the left-hand side of (7) "attribute numbers". Thus for each document obtain 32 attribute numbers—one for each of the 32 categories. Again, in (7), k is a scaling factor so that

$$\sum_{j=1}^{32} P(C_j \mid W_k \cdot W_m \cdot \ldots \cdot W_s) = 1. \tag{8}$$

Before proceeding, let us consider expression (8) which implies, in effect, that our subject categories are mutually exclusive and exhaustive. They are exhaustive in that every document must belong to at least one category, but the categories are not mutually exclusive. A document may be indexed under more than one subject category. In spite of this fact, and again to facilitate the computations, the categories are treated, in fact, as exclusive.

6.3. *Question of Estimates.* We now come to the problem of obtaining or estimating the values of those probabilities that are needed in order to compute the

probability that given the occurrence of a set of clue words, the document belongs to the jth category.

Estimate the value of $P(C_j)$ as follows: Count the number of index entries that are made under the jth category and divide this by the total numer of index entries.

Estimate the values of $P(W_i \mid C_j)$ as follows: Count the number of occurrences of the ith word which belong to documents that were indexed under the jth category. Count the total number of clue word occurrences in all documents belonging to the jth category. The ratio is our estimate of $P(W_i \mid C_j)$.

7. *Test and Results*

7.1. *First Remarks.* The stage has now been set to test our basic notions: viz., whether or not a computing machine can correctly index documents on the basis of the occurrence of selected clue words in the document. We have already discussed the key notions, the experimental corpus, the prediction methods, the relevant statistics, etc. We can now describe the test.

The test was performed in two quite separate and distinct parts, viz., on group 1 and on group 2. We would expect that better predictions could be made on that population from which the statistics were obtained and, to jump ahead briefly, that is how it turned out. We have made the assumption that our estimates of the values of $P(C_j)$ and $P(W_i \mid C_j)$ would allow us to predict fairly successfully on the uncontaminated population and, as we shall see, the results confirmed this initial state of confidence. We can now present the results separately in the following sections.

7.2. *Results on Population 1.* It turned out that in the initial group of 260 documents, 12 documents contained none of the 90 clue words, and hence no automatic indexing was possible for these 12 documents. Also there was an error preventing one of the remaining documents from being automatically indexed—this left 247 documents. Using the rules described in Section 6, the computer printed out a list of categories for each of the 247 documents and with each category the corresponding value of the attribute number. The categories in each list were then ranked according to the attribute numbers. We then asked the following question: What is the probability that a correct category will appear at the top of an output list? That is to say, what is the probability that a correct category will have the highest attribute number? In 209 of the 247 cases, the category with the greatest attribute number in each output list was a correct category. Thus, the probability[2] that if a category appears at the top of a list, it is a correct category is $209/247 = 84.6\%$. It is most interesting to note that, as one might expect, the more clue words in a document, the better the automatic indexing. Some detailed results are shown in Table I.

The data in Table I show that if a document has at least two clue words, then

[2] Strictly speaking the values that are obtained are only estimates of probabilities.

TABLE I

Number of clue words in documents	Number of such documents	Number of cases where correct category heads output list	Probability that the category on an output list with the greatest attribute number is a correct one
1	37	18	48.7%
2	33	28	84.9%
3	54	46	85.2%
4	45	41	91.1%
5	46	44	95.7%
6	19	19	100.0%
7	9	9	100.0%
8	4	4	100.0%
	247	209	84.6%

the probability that the category with the greatest attribute number is a correct one is 191/210 = 91%.

Now let us look at the results in some further detail; in the following cases consider only those documents which contain at least two clue words. Of the 210 documents with at least two clue words, 157 were indexed under exactly one category, 50 indexed under exactly two categories, and only three indexed under exactly three categories.

In 143 cases where a document was indexed under only one category, that category correctly appeared at the top of the output list computed by the machine; i.e., the correct category had the greatest attribute number. Thus, the probability that if a document is indexed under exactly one category and has at least two clue words, a machine will correctly index it is 143/157 = 91.1%.

In 45 cases where a document was indexed under exactly two categories, at least one of these categories appeared at the top of the list. And in 39 of these cases *both* of the first two categories that were listed were the correct categories. Thus, the probability that if a document contains at least two clue words and is indexed under exactly two subject categories, then the first two categories on the output list will be the correct ones is 39/50 = 78%.

In all three cases where a document was indexed under exactly three categories, the machine printed out the three correct categories in the first three positions.

7.3. *Results on Group 2.* Before turning to the analysis of the documents from group 2, it is important to keep in mind that no prior statistics on any aspects of these documents were obtained. Not only did the machine have no statistical data concerning the documents from group 2, but in the design and preparation of the experiment (an example is the selection of the clue word list) these documents were not even examined, thus avoiding any possible bias. Only after the tests of group 1 were completed were the documents of group 2 read and indexed. Once the documents were assigned to their proper subject categories, then the clue words from each document were fed to the computer, which performed the automatic indexing using the statistics gathered from the documents of group 1.

One modification was made on the computing procedure, and it is necessary to describe this before continuing.[3] Consider again the prediction schema for the general case when clue words W_k, W_m, $\cdots$, W_s are in a given document and the machine is to compute $P(C_j \mid W_k \cdot W_m \cdot \ldots \cdot W_s)$;

$$P(C_j \mid W_k \cdot W_m \cdot \ldots \cdot W_s) = k \cdot P(C_j) \cdot P(W_k \mid C_j) \cdot P(W_m \mid C_j) \cdot \ldots \cdot P(W_s \mid C_j). \quad (9)$$

Since the right-hand side of (9) is a product of probabilities, if at least one of them is zero, the value of the entire expression goes to zero. If, for example, $P(W_k \mid C_s) = 0$, then the value of the attribute number, $P(C_s \mid W_k \cdot W_m \cdot \ldots \cdot W_s)$ is 0. Furthermore, since the estimates of $P(W_i \mid C_j)$ were obtained by determining the relative frequency of the ith word in the jth category for the very small corpus of 260 documents, in many cases the value of $P(W_i \mid C_j)$ was zero, i.e., the clue word-category matrix had many empty entries.

We should realize at the same time that if the sample size were larger, it is highly probable that the number of non-zero entries in the clue word-category matrix would have increased considerably. For every empty entry in the clue word-category matrix a very small value (viz., .001) was introduced; this was just enough to prevent expression (9) from going to zero but small enough not to disturb the ordering of the values of $P(C_j \mid W_k \cdot W_m \cdot \ldots \cdot W_s)$ as j varies.

There was a total of 145 documents in group 2, twenty of which contained no clue words and another 40 which contained only one clue word. This left us with 85 documents, each containing at least two different clue words. In 44 of these 85 cases the machine printed the correct category at the top of the output list, i.e., that category with the greatest attribute number was the correct category. Thus, the probability that the first category on an output list is the correct one is $44/85 = 51.8\%$. The probability that the machine will print out the correct category in one of the first three positions is $68/85 = 80\%$.

In 66 of the 85 documents which contained at least two clue words, the document in question was indexed under only one category in 16 cases it was indexed under exactly two categories, and in only three cases was it indexed under exactly three categories.

In 33 of the 66 cases where a document belonged to only one category, the machine printed an output list in which the correct category appeared first on the list. Thus, for group 2, the probability that if a document has at least two clue words and belongs to only one category, a machine will correctly index it is $33/66 = 50\%$.[4]

In 5 of the 16 cases where a document belonged to exactly two categories, the first two categories listed by the machine were the correct ones. In 9 of the 16

[3] Paul Baran suggested this important modification

[4] Although 50 percent might seem to be a rather poor score, one should note that the probability of doing as well or better purely by chance is essentially zero

cases, the two correct categories appeared in two of the first three positions on the output list.

Finally, the results for the three documents belonging to three categories are as follows: In one case the correct three categories appeared in the first, second and third positions on the output list. In one case two of the correct three appeared in the first two positions. In one case two of the correct three appeared in the first and third position.

7.4. *Further Discussion.* What conclusions can be drawn from the results of this experiment? Many variables are present and it is difficult to submit a single, precise, unqualified judgment. Qualitatively speaking, the results are surprisingly good—much better, in fact, than one could hope for at the outset. Given a much larger class of clue words (say, 900 instead of 90), and much firmer statistics (say, taken from an initial sample of 2600 instead of 260 documents), and more complete statistical data (namely, the values of $P(W_k \mid C_j \cdot W_i)$, instead of just $P(W_k \mid C_j)$, and better prediction techniques (for example, techniques where we would consider not only the occurrence of clue words, but the frequency with which the clue words occur in a document), then clearly a machine could produce much better predictions. Again, considering the data and methods that were used, the results are encouraging indeed. The surprising thing is not that the computer could place the correct category first on a list of 32 in *only* 33 out of 66 cases, but the fact that it did this well at all.

8. *Closing Remarks*

8.1. *The Notion of Automatic Probabilistic Indexing.* An implicit assumption behind this work has been the one which asserts that either a document belongs to a given subject category or it does not—that there is no middle ground in indexing. In spite of the fact that the present work is based on that assumption, it is felt that this is essentially not the case. The relationship that a document has to a subject category is one that admits of degrees. That is to say, instead of stating that either a document belongs to a given category or not, it would be more realistic to recognize that a document can belong to a category to a degree (i.e., with a weight). Once we allow a weight to be associated with an index, the road is cleared for a radically improved interpretation of the entire library problem.[5] Specifically, such weights, in addition to statistical data on library usage, could be used by a library computer; given a request for information, a statistical inference could be made in order to derive a number (called a "relevance number") for each document, which is a measure of the probable relevance of the document for the requester. The result of a search would be an ordered list of those documents which satisfy the request and ranked according to the relevance number.

Automatic indexing can be replaced by automatic probabilistic indexing. That

[5] This new interpretation of information identification and retrieval, along with an explication of the notion of an index weight and an explication of a comparative concept of relevance is described in "On Relevance, Probabilistic Indexing and Information Retrieval," M. E Maron and J. L. Kuhns, *J. ACM* 7 (1960), 216–244.

is to say, recognize the notion of a weighted index and arrange so that the attribute number can be interpreted as the weight of an index tag.

8.2. *A Final Remark.* On the basis of this very modest experiment we can make the following inferences:[6] It is feasible to have a computing machine read a document and to decide automatically the subject category to which the item in question belongs. No real intellectual breakthroughs are required before a machine will be able to index rather well. Just as in the case of machine translation of natural language, the road is gradual but, by and large, straightforward. If one is willing to collect enough statistical data relating words and categories, and if one is prepared to consider more and more of the relationships that exist between individual words, word combinations, word type, etc., and categories, then one can index by machine with increasing accuracy.

Acknowledgment

It is a pleasure to acknowledge the helpful assistance received from various members of the RAND Computer Sciences Department. In particular, I thank Sharla Perrine for her excellent work in performing the computations. William Nadeau's help was invaluable in sorting and organizing the word occurrence data. Michael Warshaw helped to create the subject categories and to index the documents of population 1. Finally, I benefited from enthusiastic discussions with Paul Baran.

APPENDIX A. A Typical Document

Control Apparatus for a Serial Drum Memory, by D. S. Kamat (Indian Stat. Inst , Calcutta), Electronic Engrg., vol. 30, pp. 634–639; November, 1958. A control apparatus that has been developed and used successfully to obtain design data for a faster track switching device for a serial magnetic drum memory is described. The apparatus generates coded information consisting of a 32-bit word, routes this information to a given location on a memory track or extracts information from a location and stores it in a register, and generates fast switching impulses used both in the selection of a required track and in performing either of the record-reproduce functions. Circuit and performance details of the gates, triggers, switches, and other components are given The apparatus can also be used to study other computer functions.

[6] There is a distinction between a document and an abstract of a document and, strictly speaking, these inferences hold for a library of abstracts. However, these principles for automatic indexing can be applied equally well for document indexing.

APPENDIX B. List of Subject Categories

1. Logical design and organization of digital computers
2. Digital data transmission systems
3. Information theory
4. Intelligent machines and programs
5. Number theory
6. Cybernetics
7. Pattern recognition techniques
8. Digital computer storage devices.
9. Language translation and information retrieval.
10. Digital counters and registers
11. Error control techniques
12. Number systems and arithmetic algorithms
13. Arithmetic units
14. Digital logical circuitry
15. Analog circuits and subsystems
16. Digital switching components
17. Physical characteristics of switching and memory materials
18. Automatic control and servomechanisms
19. Input-output devices
20. Analog-to-digital conversion devices
21. Analog system descriptions
22. Business applications of digital computers
23. Scientific and mathematical applications of digital computers
24. Real-time control system applications of digital computers
25. Digital computer programming
26. Applications and theory of analog computers
27. Simulation applications of computers
28. Glossaries, terminology, history and surveys
29. Numerical analysis
30. Boolean algebra
31. Switching theory
32. Combined analog-digital systems and digital-differential analyzers

APPENDIX C. Selected Key Words

1. Abacus
2. Adder
3. Analog
4. Arithmetic
5. Average
6. Barium
7. Boolean
8. Bound
9. Carry
10. Character
11. Characters
12. Chess
13. Circuit
14. Circuits
15. Code
16. Coding
17. Communications
18. Complexity
19. Compression
20. Control
21. Conversion
22. Counter
23. Decoder
24. Definition
25. Delays
26. Differential
27. Diffusion
28. Division
29. Element
30. Elements
31. Equation
32. Equations
33. Error
34. Expressions
35. Fields
36. File
37. Films
38. Function
39. Functions
40. Generator
41. Information
42. Language
43. Library
44. Logic
45. Magnetic
46. Matrix
47. Mechanical
48. Mechanisms
49. Memory
50. Monte (Carlo)
51. Multiplication
52. Network
53. Networks
54. Numbers
55. Office
56. Parity
57. Plane
58. Printed
59. Process
60. Processing
61. Program
62. Programming
63. Programs
64. Pseudo-Random
65. Pulse
66. Randomness
67. Recording
68. Register
69. Scientific
70. Section
71. Shift
72. Shuttle
73. Side
74. Simulation
75. Solution
76. Speech
77. Square
78. Stage
79. Storage
80. Switching
81. Synthesis
82. Tape
83. Traffic
84. Transistor
85. Transistors
86. Translation
87. Transmission
88. Uncol
89. Unit
90. Wire