RESEARCH IN AUTOMATIC GENERATION OF CLASSIFICATION SYSTEMS

Harold Borko, Ph.D.
System Development Corporation
Santa Monica, California

INTRODUCTION

This paper is concerned with the organization of information, in the form of documents, for efficient storage and retrieval. By documents we mean books, technical reports, articles, memoranda, letters, photographs, data facts, etc.—all forms of memory file organization ranging from documents in a library to data in a real-time command-and-control system. Therefore, the implications of this work are applicable to a field broader than the concerns of the ordinary library.

In actual practice, most of the information retrieval research has been concerned with document files because the most highly organized collection of documents in existence today is the library, and in doing research on methods of organizing information, one must compare the adequacy of proposed new techniques with existing library methods. Procedures which will improve information storage and retrieval in a library will probably be sufficiently powerful to help improve other methods of file organization.

PURPOSES OF DOCUMENT CLASSIFICATION

The reason for maintaining a collection of documents is to have an available store of information and to be able to retrieve desired information rapidly and with confidence. The value of classification is that it increases efficiency in locating this desired information. If we tried to locate a book on a particular subject in a library that did not use any system of classification, we would have to spend a long time reading the titles and authors of several thousand books before we could find the one for which we were looking. If we knew the author, and the books were arranged alphabetically by author, we could locate the book quickly. On the other hand, if we didn't know the author, but knew the subject content of the book, we would want the books arranged by subject category in order to search the file efficiently. Finally, if all we knew was that the book we sought was a big black one which we could recognize, we would like the files arranged by color. The point being made is that there are various ways of organizing a file, and whether or not a particular method of file organization is efficient depends upon the search strategy. Furthermore, no one method of file organization would be equally efficient for all search questions. This is an important, if obvious, point and one which is often overlooked.

The central theoretical problem of classification as a method of organizing documents is that only one principle at a time can be utilized for gathering items together. This principle can be alphabetic arrangement by author, color coding based on the binding of the book, subject classification, or and other scheme—as long as only one principle is used at a time.

Since a document collection is a store of information, it is usually desirable to organize this store according to subject matter. By establishing clearly demarcated groups, or classes, of documents on related topics, the number of documents to be scanned can be reduced to reasonable proportions. This, in essence, is the purpose of classification. A classification system is a scheme for organizing a mass of material into groups so that related objects are brought together in a systematic fashion. Objects in one group are selected so as to be more like each other than objects in any other group. However, before this aim can be realized, two questions must be answered:

1) How many classes shall be established?
2) What shall be the measure of similarity and, hence, what is the principle to be used in determining class membership?

If one is interested in automatic procedures, one has to answer a third question, namely:

3) What principles of classification are most amenable for use in an automated document classification system?

All three of these problem areas are being studied, and some results are already available.

DEVISING A CLASSIFICATION SCHEDULE

The classification of knowledge is not a new problem. Even in ancient times, man sought to organize information of the world around him into categories for efficient retrieval. What is new, perhaps, is the application of mathematical techniques to the classification problem. The older forms of classification, from ancient times through the Dewey Decimal System, were attempts to impose logical subdivisions on the whole field of knowledge. The surprising thing is not that these systems were imperfect, but rather that they succeeded as well as they did. Melvil Dewey first proposed his Dewey Decimal System in 1876, and it is still in extensive use. Now, as a result of new inventions and accelerated research, the traditional boundaries between the sciences are breaking down. It is time to reexamine the concept of classification, to go back to basic principles and to study the various methods of deriving a classification system.

Factor Analysis—Borko

In 1958, Tanimoto[12] published a theoretical paper on the applications of mathematics to the problems of classification and prediction. Specifically, he pointed out how the problems of classification can be formulated in terms of sets of attributes and manipulated as matrix functions. An actual application of matrix mathematics to the analysis of a collection of documents was made by Borko[2] in 1961. The aim of this study was to determine whether it was possible to derive a reasonable classification schedule for a collection of documents by factor analysis,[6] a mathematical technique which enables one to isolate the underlying variables in a domain of events. This method has been used by psychologists to determine the underlying variables of intelligence, personality, creativity, ability, etc.

In Borko's classification study, factor analysis was used to discover the relationship of key content words as they are used in psychological literature. Approximately 600 psychological abstracts were selected for study. These were key punched in their entirety, and by means of a computer program called FEAT[9] (Frequency of Every Allowable Term), a frequency count was made of all words, and 90 tag terms were selected for further analysis. These data were arranged in the form of a matrix consisting of 90 terms and 618 documents. A portion of this matrix is reproduced in Table 1. The number in each cell represents the number of times a given word occurred in a particular document. Table 1 shows that the term "child (children)" did not occur in document number 74, occurred twice in document number 307, and three times in document number 374. The term "level (s)" occurred once in document-numbers 74, 626, and 674 and did not occur in the other documents in the example.

Words	Case (s)	Child (ren)	Factor (s)	Level (s)	Psychology (ical)	School (s)
Abstract # 74	0	0	1	1	1	0
307	1	2	0	0	0	1
321	0	2	1	0	1	0
575	1	2	0	0	1	0
626	0	1	0	1	0	2
647	1	2	0	0	1	0
653	4	1	0	0	1	0
674	1	3	0	1	0	0

Table 1. A Portion of the Document Term Matrix

Based upon the data in the document-term matrix, one can compute the degree of association among the terms as a function of their occurrence in the same set of documents. A measure of this association is the correlation coefficient. This is a decimal number which varies from +1.000 to —1.000. A +1.000 would mean a perfect correlation, namely, that every time word X occurred, word Y appeared in the same document; a zero correlation would indicate no relationship; and a negative correlation would mean that if the word X occurs in a document, then word Y is not likely to occur.

The formula for computing the correlation coefficient is as follows:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)\ (\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]\ [N\Sigma Y^2 - (\Sigma Y)^2]}}$$

By applying this formula and computing the correlation between each of the 90 words with every other word (a total of approximately 4000 correlations), one creates the term-term correlation matrix (Table 2). This matrix expresses the actual associations which occurred among selected words in a sample of documents.

These statistical procedures, preparatory to the factor analysis, are important in demonstrating a method for translating a conglomeration of words and documents into a set of vectors which can be processed mathematically. Factor analysis, when applied to the correlation matrix, enables one to determine the basic underlying variables which account for the relations among the words as expressed in the vectors. It enables us to mathematically determine which words are related and form a set; these sets, in turn, are interpreted as classification categories for grouping the original sample of documents.

	Ability	Achievement	Activity	Analysis	Anxiety
Ability		.272	—.028	.048	.080
Achievement	.272		—.026	—.041	.119
Activity	—.028	—.026		—.002	—.025
Analysis	.048	—.041	—.002		.030
Anxiety	.080	.119	—.025	.030	

Table 2. A Portion of the Correlation Matrix

In the experiment just described, the original 90-column matrix was reduced to 10 vectors which accounted for 62% of the total, and it is assumed most all of the common, variance. These vectors were then rotated mathematically to achieve a simpler and more meaningful structure of the hyperspace. They were then interpreted by the investigator as ten classification categories into which the original sample of 618 psychological reports could be grouped—and by implication, all psychological literature.

To illustrate how the factors were interpreted, let us examine the words which had significant loadings on the first factor.

Term #	*Word*	*Factor Loading*
33	girls	.74
10	boys	.73
70	school	.30
2	achievement	.20
63	reading	.18

There were only five words with significant loading. It is fairly obvious that the concept underlying these terms deals with the achievement of boys and girls in school; consequently, this factor was interpreted as academic achievement of boys and girls in school; consequently, in a like manner. These included factors named experimental psychology, social psychology and community organization, school guidance and counseling, clinical psychology and psychotherapy, etc.

Thus, we have arrived at answers to the two questions posed earlier in this paper: How many classes should be established, and what shall be a measure of similarity? The application of factor analysis enables one to determine the number of categories which should be established in order to adequately describe a given sample of documents. Furthermore, it provides a statistical technique, or principle, for measuring the similarity of content based upon the co-occurrence of key content terms.

There are many questions still to be answered before one can decide on the usefulness of this technique for classification. These questions include:

1) Are the categories stable; do they hold from one sample of psychological literature to another?
2) Are the categories valid; can all documents be reasonably classified into these categories?
3) Are the categories useful; do they lend themselves to automated document classification?
4) Is the technique a general one; can it be applied to documents other than psychological reports?

Before reviewing the studies designed to answer these questions, it would be well to first examine some other mathematical techniques for deriving classification schedules.

Clump Theory—Parker-Rhodes and Needham

At the Cambridge Language Research Unit in England, Parker-Rhodes was also interested in classification theory and a mathematical basis for forming classes of documents. Interestingly, he considered the use of factor analysis but rejected it on two grounds, one theoretical and the other practical. From a theoretical point of view, Parker-Rhodes claimed that "the statistical type of technique has its place only after we have discovered whatever classification there may be. For then it is up to the statistician to say how nearly the properties of particular elements of the universe are inferable from a statement of the classes to which each belongs. . . . This is quite a different enterprise from that of finding the classes themselves"[10] (page 4). On the practical side, factor analysis is rejected as being incapable of handling "really large universes."

Having decided to avoid the statistical concept of determining the probability of class membership, Parker-Rhodes restructured the problem in terms of locating clumps "in a Boolean lattice representing all possible subsets of the universe." Within a Boolean lattice there are many ways of defining clumps, and in fact, many different clumps are defined. Without getting involved in details, it can be broadly stated that "members of a clump must be more like each other, and less like non-members, than elements of the universe picked at random" (page 9). Thus we see the relationship between the theory of clumps and the theory of classification. The method used for locating clumps within the lattice remains to be worked out. Initial procedures for clumping are described by Needham.[8] Research aimed at improving and testing these procedures is still going on. However, even now these techniques have been applied to a 346 x 346 matrix which is beyond the capabilities of presently available factor analysis programs.

Latent Class Analysis—Baker

The similarity between document classification and the problems inherent in the analysis of sociological questionnaire data was recognized by Baker. He then proposed an information retrieval system based upon Lazarsfeld's latent class analysis.[1] As Baker points out, "The raw data of documents, the presence or absence of key words, is amenable to latent class analysis without modification of either the analysis or the data. The latent classes and the ordering ratios yielded by the analysis provide the basis for a straightforward means of classification and retrieval of documents" (page 520).

The latent class model assumes that the population—that is, the number of documents in the sample—can be divided into a number of mutually exclusive classes. Usually the number of classes is determined by the investigator, although it is conceivable that this parameter can be determined mathematically. One starts by selecting the key words which characterize each class of documents. Then latent class analysis is used to compute the probability that a document having a certain pattern of key words belongs to a given class.

Baker gives the following example: Let us assume that we have 1000 documents in our file. We are interested in classifying these documents into two classes—those dealing with computer automated instruction and those not directly related to this topic. We select as the key words in our search request the following:

1. computer.
2. automated.
3. teaching.
4. devices.

Each of the 1000 documents are then analyzed to determine whether they contain one or more of the four terms. Sixteen (2^4) response patterns are possible, ranging from ++++ to 0000. A χ^2 test enables one to estimate the latent structure from the observed data. Having obtained a latent structure which fits, one can compute an ordering ratio, which is the probability that a document having a given word pattern belongs to a particular latent class. For example, a document with all four key words present has a probability of .998 of belonging to class 1, i.e., it is concerned with computer automated instruction.

Table 3 shows the relationships between the response pattern, expected frequencies, and ordering ratios of the 1000 documents analyzed, in terms of their latent class structure.

Response Pattern	Expected Frequency		Total Fitted	Ordering Ratios	
	Class 1	Class 2		Class 1	Class 2
++++	158.76	.24	159.00	.998	.002
+++0	105.84	2.16	108.00	.980	.020
++0+	68.04	.96	69.00	.986	.014
+0++	68.04	2.16	70.20	.969	.031
0+++	17.64	.56	18.20	.969	.031
++00	45.36	8.78	54.14	.838	.162
+0+0	45.36	19.44	64.80	.700	.300
0++0	11.76	5.04	16.80	.700	.300
+00+	29.16	8.60	37.76	.772	.218
0+0+	7.56	2.39	9.95	.768	.232
00++	7.56	5.04	12.60	.600	.400
+000	19.44	77.76	97.20	.200	.800
0+00	5.04	20.16	25.20	.200	.800
00+0	5.04	45.36	50.40	.100	.900
000+	3.32	20.16	23.48	.142	.858
0000	2.16	181.20	183.36	.012	.988

Table 3. Expected Frequency of Response and the Ordering Ratios Based Upon the Estimated Latent Structure[1]

The table readily reveals the applicability of latent class analysis for information retrieval. This application is still in the theoretical and experimental stages. It has yet to be tested with empirical data from actual files.

AUTOMATED DOCUMENT CLASSIFICATION

The preceding discussions of factor analysis, clump theory, and latent class analysis all dealt with methods for devising empirically based classification categories. These and other researchers have been investigating mathematical methods for deriving classification categories because of their belief that empirical classification systems will provide a more efficient means for the classification and the retrieval of information than the traditional methods of document classification. This belief has been subjected to scientific tests and evaluations.

In a study by Borko and Bernick,[3] an attempt was made to test the hypothesis that a classification system derived by factor analysis provides the best possible basis for automatic document classification and would result in more accurate automatic classification of documents than would be possible using more traditional classification categories. Maron,[7] in pursuing his interests in automatic indexing and classification, worked with 405 abstracts of computer literature which had been published in the *IRE Transactions on Electronic Computers,* Volume EC-8. In essence, Maron proposed a set of 32 subject categories which he felt were logically descriptive of the computer abstracts. Then he selected 90 clue words in such a manner that they would be good predictors of his 32 categories. The 405 documents were divided into two groups—260 abstracts made up the experimental group and the remaining 145 comprised the validation group. Maron classified all 405 documents into the 32 categories. Working with the documents of the experimental group only, he computed the value of the terms in the Bayesian prediction equations. He then used this formula to automatically classify the documents into their categories. Automatic document classification was correct in 84.5% of the cases in the experimental group and in 51.8% of the cases in the validation group.

Borko and Bernick decided to test the hypothesis that a higher percentage of correct classifications could be made using the same set of documents if a factor-analytically derived classification system were used instead of Maron's logically derived categories. However, their results were approximately the same as those obtained by Maron. Because of the nature of the experimental design used, it was impossible to determine whether the difficulty lay in the mathematically derived classification system or whether the factor score method used to predict correct document classification was not as effective as the Bayesian prediction equation.

Another series of experiments were designed and executed.[4] It was concluded from this series that, while there was no significant difference between the predictive efficiency of Bayesian and factor score methods, automatic document classification is enhanced by the use of a factor-analytically derived classification schedule. Approximately 55% of the documents were automatically and correctly classified. While this 55% current automatic classification of the documents is statistically very significant, it will have little practical significance until greater accuracy can be demonstrated.

Up to this point the criterion for correct classification has been the human classifier, but this is not necessarily the best criterion. We know that humans are not perfectly reliable, and therefore, it is not possible to predict human classification with perfect accuracy. The ultimate criterion of the usefulness of any indexing and classification system is whether it retrieves relevant information in response to a search request. Automatic document classification procedures should be evaluated on how efficiently they retrieve information and not on how well they can match the imperfect human classifier. This is a much more difficult problem, but research is already under way to evaluate the retrieval effectiveness of automatic document classification. As work progresses on the evaluation and improvement of techniques for automatic document indexing and classification, it can be anticipated that the bottleneck which now exists between the collection and the processing of documents will be eliminated and automated storage and retrieval systems will become possible.

BIBLIOGRAPHY

1. BAKER, F. B. Information Retrieval Based Upon Latent Class Analysis. *Journal of the Association of Computing Machinery,* Vol. 9, No. 4, Oct. 1962, 512–521.
2. BORKO, H. The Construction of an Empirically Based Mathematically Derived Classification System. *Proceedings of the*

Spring Joint Computer Conference, San Francisco, May 1–3, 1962, Vol. 21, 279–289. (Also available as SDC document SP–585.)

3. BORKO, H., and BERNICK, M. Automatic Document Classification. *Journal of the Association of Computing Machinery,* Vol. 10, No. 2, April 1963, 151–162. (Also available as SDC document TM–771.)

4. BORKO, H., and BERNICK, M. Automatic Document Classification: Part II—Additional Experiments. *Journal of the Association of Computing Machinery,* Vol II, No. 2, April 1964. (Also available as TM–771/001/00.)

5. BRITISH STANDARDS INSTITUTE. *Guide to the Universal Decimal Classification (UDC)*. British Standards House, London, 1963.

6. HARMAN, H. H. *Modern Factor Analysis.* University of Chicago Press, Chicago, 1960.

7. MARON, M. E. Automatic Indexing: An Experimental Inquiry. *Journal of the Association of Computing Machinery,* Vol. 8, No. 3, July 1961, 407–417.

8. NEEDHAM, R. M. The Theory of Clumps, II. M. L. 139, Cambridge Language Research Unit, Cambridge, England, March 1961.

9. OLNEY, J. C. FEAT, An Inventory Program for Information Retrieval. SDC document FN–4018, July 1960.

10. PARKER-RHODES, A. F. Contributions to the Theory of Clumps, M. L. 138, Cambridge Language Research Unit, Cambridge, England, March 1961.

11. SHERA, J. H. and EGAN, M. E. *The Classified Catalog: Basic Principles and Practices,* American Library Association, Chicago, 1960.

12. TANIMOTO, T. T. An Elementary Mathematical Theory of Classification and Prediction, IBM, New York, 1958.