# *A NEW METHOD OF RECORDING AND SEARCHING INFORMATION*

H. P. LUHN*

This method applies to the procedures required to record a legend concerning a document and to enable an inquirer to locate this document by means of the legend, if it is related to a specified subject.

The conventional methods of indexing and classifying attempt to evaluate the relative importance of a plurality of aspects contained in a document and makes the most important one the key for locating the document within an orderly scale of a certain dimension. Subordinated aspects are covered by way of reference in appropriate other locations of the scale.

One of the disadvantages of the conventional system is that the standard of value on which the indexer bases his decision may change and, what suddenly is considered an aspect of major significance, may not have been included in the classification or index at the time, even though it was contained in a document.

Another drawback is that it becomes difficult for an inquirer to reverse the process of classification or indexing and pose his query in a form matching to a reasonable degree the values of a potential reference.

The new method uses the principle of characterizing a topic by a set of identifying elements or criteria. These elements may be of any dimension and as many may be recorded as is desirable. Also, they are not weighted and no significance need be implied by the order in which they are given.

One of the main functions of the new method is that of producing a response to an inquiry in all cases, even if the reference appears to be remote, it being the understanding that it is the closest available.

The elements enumerated by recorders to identify a topic will necessarily vary as no two recorders will view a topic in identical fashion. Similarly, no two inquirers, when referring to the same subject will state their query in identical fashion. It is therefore important that a system recognizes that these variations arise and that they cannot be controlled. It must then become the function of the system to overcome these variations to a reasonable degree.

When identifying a topic by a set of criteria or identifying terms, the more terms are stated the more specifically the topic is delineated. Each term in turn may be a concept which in itself may vary as to specificity. If we consider a concept as being a field in a multi-dimensional array, we may then visualize a topic as being located in that space which is common to all the concept fields stated. It may further be visualized that related topics are located more or less adjacent to each other depending on the degree of similarity and that this is so because they agree in some of the identifying terms and therefore share some of the concept fields. Figure 1 is a diagrammatic illustration.

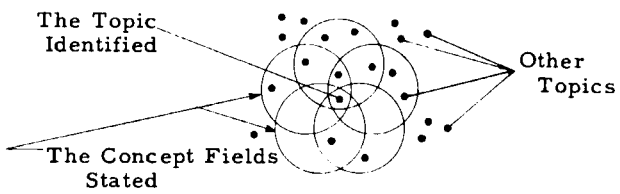



FIGURE 1.

In order to understand the nature of the arrangement, let us assume a vocabulary of 100 concepts and let us identify a topic by five conceptual terms. By using all possible combinations of five terms, a total of 75 million patterns of criteria result, each of these patterns having a fixed location within the system. If then a topic is identified by five terms of the vocabulary, it is thereby assigned to a definite one of these fixed locations.

While assuming that there is an ideal and true location where a topic belongs, it is un-

*International Business Machines Corporation, Engineering Laboratory, Poughkeepsie, New York.

were to do the same job. There will result a deviation from the true location proportionate to the degree of disagreement of either. For instance one recorder may diverge to the extent of matching only 3 of the 5 criteria while the other matches 4. The resultant displacement is shown in diagram Figure 2.

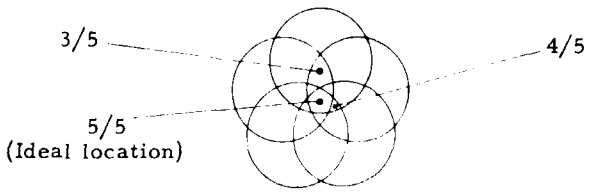



FIGURE 2.

Such disagreements will be the more pronounced the more specific the conceptual terms are and it is a further function of the new method to minimize variations by broadening the concept used in the terms and by using as large a number of broadened criteria as possible even to the extent of redundancy. This approach is based on the realization that an inquirer could not match excessive specificity when stating his query and that his position is similar to that of the recorder.

The process of broadening the concept involves the compilation of a dictionary wherein key terms of desired broadness may be found to replace unduly specific terms, the latter being treated as synonyms of a higher order than ordinarily considered. Translating criteria into these key terms is a process of normalization which will eliminate many disagreements in the choice of specific terms amongst recorders, amongst inquirers, and amongst the two groups, by merging the terms at issue into a single key term. However the dictionary does not classify or index but maintains the idea of terms being fields and applies the identification principle to the terms in the manner it is applied to the topics, even though to a lesser degree. A specific term may appear under the heading of several key terms and if according to its application an overlapping of concepts exists then the term is represented by the several key terms involved, as shown diagrammatically in Figure 3 for 'b'.

The manner which an inquirer approaches the process of searching for desired information becomes one similar to that performed by the recorder. He first states hiw query in as many and as specific terms as he desires. Then with the aid of the special dictionary he normalizes the conceptual terms of identification to arrive at a statement adjusted to the requirements of the system.

The actual process of searching involves the comparing of his statements with all the statements contained in the collection of records prepared by the recorder. This task, being beyond human capability, may be performed automatically by a scanning machine which is capable of not only matching similar portions of information but of doing this in accordance with any conceivable pattern of conditions.

As indicated earlier, the intended purpose of a search is to produce a response to a query. Because it is not usually known how specific a response can be expected, the initial query is stated rather broadly thereby extending the field to include less related material. The extent of responses obtained on this bases is a valuable indication of the amount of attention devoted to the subject area in the past. The material abtained would then be subjected to increasingly more specific searches in order to get the closest match possible. Also, material uncovered by this approach may lead to the discovery of unsuspected, but pertinent other related information.

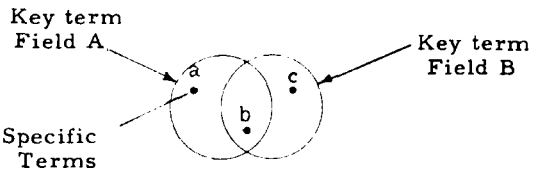
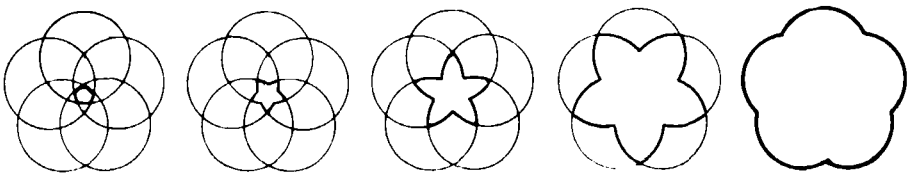



FIGURE 3.

In particular, the scheme of broadening the field of response consists of asking that a fixed fraction of the given terms be met by the records. This procedure is quite different from that used when broadening a generic search by dropping subclasses. The effect is illustrated by the following diagrams, Figure 4, showing progressively broader fields formed by 5 terms.

Using the proportions of the example previously given and assuming an evenly distributed population of topics, the relative probability of response is expressed by the factors

listed below each fraction. While applied to an idealized situation, the results are nevertheless indicative of the advantages the method of identification has over other methods of indexing information.



| 5/5 | 4/5 | 3/5 | 2/5 | 1/5 |
|---|---|---|---|---|
| probability factor of response: | | | | |
| 1 | 96 | 4,656 | 152,096 | 3,764,376 |

FIGURE 4.

# *THE USE OF THE UNIVAC FAC-TRONIC SYSTEM IN THE LIBRARY REFERENCE FIELD*

HERBERT F. MITCHELL, JR.*

The tremendous increase in the volume of technical literature of all kinds and fields is presenting the librarian with an almost impossible reference task. The sheer volume of these documents is creating a filing problem of the first magnitude. When this volume is combined with the fact that many documents cut across classification lines, the problem of providing reference bibliographies is made that much more difficult.

Several persons concerned with the furnishing of reference material have approached those of us engaged in the manufacture and utilization of digital computers to see if these machines might be of assistance to the librarian. Such an occasion arose a little over a year ago when the Centralized Air Document Office in Dayton, Ohio, approached Remington Rand to ascertain the suitability of our equipment for this work. A study was made to see how the UNIVAC Fac-Tronic System might be applied to the task of obtaining all possible documents from a large file which could answer a specific query submitted to this office. The model studied invisioned a library of 1,000,000 documents. Each document was identified by an eight-digit shelf number. A master reference file was to be compiled, each item of which would consist of the shelf number followed by a series of coded approaches. Each such approach would represent some pertinent feature of the document, such as: author, data, contract number, and descriptors of the subject or subjects treated by the document. It was anticipated that each document would have an average of fifteen approaches with a maximum of thirty.

In order to obtain a list of all documents which might possibly answer a given query, the computer would be supplied with the appropriate coded approaches included in the query. It would then search through the entire master file and select all document items which contain the approaches given in the query.

For such a system as the above to be work-

*Director, UNIVAC Applications Department Remington Rand, Inc.