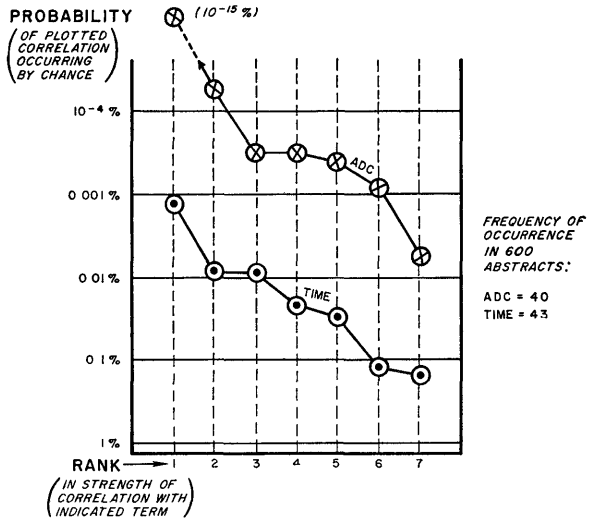


Fig. 3—Probability of highest observed correlations as a test for subject words.

other hand is a good solid subject word, standing for "Air Defense Command." Our alphabetized index enables us easily to select the words which correlate most highly with these two topics, and we can calculate and compare probabilities.

One can reason that a good subject word should have certain other words which co-exist with it in the same documents with a frequency much greater than expectable from chance distribution, but that nonsubject words, which are likely to be used by anybody writing about any subject, should not have high correlations.

To test out this notion I calculated the probabilities for the highest observed correlations occurring by chance, both for "ADC" and for "time," given that all words in the library are randomly assigned to documents. Of course, the words are not randomly assigned, so I got some very low probabilities. As Fig. 3 shows, the ADC correlations are very much more improbable than those for "time."

Now these correlations will become weaker for any word, subject word or otherwise, as they are present in fewer documents. This means that in order to apply any sort of correlation test to select subject words, one has to make allowance for the frequency of the word. Unfortunately, the correlation test will fail altogether when a word is present in only 3 or 4 documents. One has to have a large enough sample. Also author biases in use of common words could conceivably cause many nonsubject words to pass a correlation test. But, fortunately, as collections increase in size these effects should become less important, and it is in very large collections where this sort of methodology will be needed.

## Conclusion

As a final comment: libraries and other collections of written information can be thought of as realms of nature, subject to scientific observation. Science brings valid simplicity to that which is apparently complicated, and it is hard to find anything more complicated than masses of ideas recorded on paper. And so I make explicit an idea which I hope has been implicit in this presentation—that general-purpose computers of today give us the opportunity to apply scientific method to uncover the principles of the nature and use of information in order that we may put to better use the vastly more powerful computers of tomorrow.

# A Theory of Information Retrieval

CLINTON M. WALKER†

THE mathematical formula which best describes my conclusions from reading the literature on information retrieval (IR) is the following:

4 U OK 4 $ = ET

For you, better for dollar equality.

This states that a more economical approach for organizations interested in information retrieval might be collectively to support as an information retrieval center some nonprofit organization, such as SRI or SDC. Such an organization could be a center for receiving and dissemination of up-to-the-minute retrieval literature of organizations concerned; could advise on the practicability of certain undertakings; and could perform experiments in the field of IR.

Aside from this one equation, formulation should proceed from basic principles. Perhaps the most basic of all principles is that meaning, rather than information alone, needs to be retrieved. Just how does one produce or obtain meaning? Take the example of a small child. All a child knows at first is himself. He gets acquainted with his hands and feet, and then with his near associates by relating them to himself. He gradu-

† Hughes Aircraft Co., Culver City, Calif.

ally learns to classify things in terms of roundness, which things he might call a ball; in terms of use, such as food. New things learned are related to things already known. For example, at an early age, any man might be classified as "daddy." Throughout his life, meaning is obtained by relating what is familiar to that which is unfamiliar.

We might christen this process the "Mew-Mew" theory of meaning. Each of us, as a "me" looks at something else as a "you," which we interpret in terms of the "me" or what is known, but which we might also take back into the relative "you" for objective evaluation.

This process of classification is an operational way to produce meaning. A language, in effect, classifies nouns; descriptions of "which," "what kind of," and "how many" have meaning when related to other nouns. What the nouns do—and how, when, and where they do it—has meaning when related to what other nouns might be doing.

So, an operational language is one in which classification takes place in familiar areas or domains. In these domains, dictionaries can be constructed of key nouns; definitions can include relationships to other key nouns within the area. For retrieval purposes, reference to a key noun could have a built-in potential reference to other key nouns, thus providing a built-in meaning potential. The prospects are exciting. But, before we develop the idea further, let us lay down some basic postulates.

We can set up a number series as a set $R$ of objects called nouns, with the relationships defined by three operations denoted by $\Sigma$, $\Pi$, and $\int$. Concomitant with this set is another set $M$ whose members can be derived from certain operations on the set $R$. The symbol $\rightarrow$ means "results in a relationship of," or "implies that"; the symbol "$+$" means "and"; the symbol "$-$" means "not"; "( )" are used in the usual enclosure sense. An IR specific operation is one denoted by the symbols $\Sigma$, $\Pi$, or $\int$. An IR nonspecific operation is any other operation in real or complex variable theory. We will assume that IR nonspecific operations will follow the manipulative rules of real and complex numbers for IR specific operations. For example, the operations are additively commutative.

$$(A\,J\,B) + (C\,J\,D) = (C\int D) + (A\int B)$$
$$(A\Sigma B) + (C\Sigma D) = (C\Sigma D) + (A\Sigma B)$$
$$(A\Pi B) + (C\Pi D) = (C\Pi D) + (A\Pi B).$$

The operations within the parentheses are IR specific; those between the parentheses are IR nonspecific. Additional postulates are required for defining the operation processes in an uncompleted operation. The following postulates and definitions are offered for consideration.

### Definition

The domain of $A$ consists of all subcategories and subsequent subcategories under $A$.

$A\int B$ states that a word, $A$, which is classified in category $B$ is put in a relationship such that $A$ is in a hierarchy less than that of $B$, and that the domain of $A$ includes not more than the domain of $B$. That is, $A$ is part of $B$.

### Postulate 1

$$A\int B\int C \rightarrow A\int C$$

states that a member of a subcategory is also a member of a category; for example, shoelace is a subcategory of shoe, which is a subcategory of clothing, which implies that shoelace is also a subcategory of clothing.

### Postulate 2

$$A\int B \rightarrow B\int A$$

means that a category cannot be a member of a subcategory unless it is the only member. True, in ordinary language, sight can be thought of as a subcategory of sensing and perhaps sensing at the same time can be thought of as a subcategory of sight; but, for the convenience of constructing an unambiguous dictionary, we can exclude this possibility until such time as we find it absolutely required. Thus, we shall construct a dictionary in a domain with rigid hierarchical relationships among nouns. If later, we want to relax this requirement, we may find some interesting experiments available in the realm of "machine thought processes."

### Definition

$A\Sigma B$ means that $A$ is synonymous with $B$.

### Postulate 3

$$(A\int C) + (A\Sigma B) \rightarrow B\int C$$

means that synonyms within an area are necessarily members of the same category.

### Definition

$A\pi B$ means that $A$ and $B$ are related by means of the characteristics of some domain. We shall call these words "relatives."

### Postulate 4

$$(A\Pi B) + (A\int C) \rightarrow B\int C$$

means that relatives are subcategories of the same category.

### Definition

$M$ is a set of elements of meaning derived by categorizing two or more elements of $R$ at the same time.

### Postulate 5

$$(B\Sigma C) + A\int(B + C) \rightarrow (A\int B) + M = (A\int C) + M$$

means if $B$ and $C$ are synonymous, and $A$ is a common subdivision of both of them, then the classification of $A$

into $B$ and $A$ into $C$ simultaneously adds meaning to one of them, but it would be redundant to use both classifications.

POSTULATE 6

$$(A \amalg B) + (A + B) \int C = (A \int C) + (B \int C) + M$$

means that when relatives $A$ and $B$ are, together, classified as members of $C$, they contain an element of meaning which is not present when they are separately so classified.

POSTULATE 7

$$(B \amalg C) + A \int (B + C) \rightarrow A \int B + A \int C + M$$

means that to categorize a subdivision of two relatives, $B$ and $C$, is to add meaning to both of them.

We have, by these postulated operations, created a language of classification—an operational linguistics which should be compatible with operational mathematics. Hopefully, a classification of nouns accessible by data processing equipment can relieve the information seeker of the trouble of searching the entire haystack for his needle of information and thread of meaning. Purposely sacrificed is the richness of redundant normal language in favor of the more important feature of exactness. Not only do we attempt to be more exact, but also to minimize ambiguity, to allow easy translation of concepts, to assure objective criteria of meaning, and to provide a basis of agreement in discrimination.

Postulate 1 tells us which words can be classified in a given domain. Postulate 2 prevents common words from being counted in esoteric categories, unless they are subsumed under those categories. Postulate 3 permits the counting of synonymous words in the same frequency tally. Postulate 4 permits the discovery of alternative paths for continued search. Postulate 5 permits singling out of the representative path among equivalent paths to be followed. Postulate 6 shows that two words are more significant if the context does classify them together. Postulate 7, finally, shows that paths which originally diverge become significant upon reconvergence. In all those postulates which have symbol $M$ as an added element, significance is increased since $M$ represents meaning and meaning is of prime importance in transference.

In any system of information retrieval, there are factors of cost, speed, and power. These three criteria can be used to determine, under a given circumstance, which of several alternatives is to be preferred. In many instances, the major purpose of the retrieval system is to perform a rough scanning job for a literature searcher, to determine for him whether a particular document is worth further reading. Often the author can furnish, in addition to his name and topic, a list of his main ideas and purposes. He might even estimate a degree of correlation between the concepts embodied in his document and a list of key nouns in its general area.

To apply the power criterion, the machine or human doing the segregation of valuable from useless information can be simulated by a filter separating relevant message from total signal. Assuming homoscedasticity and linearity in the specified direction, a function

$$F = \Sigma(y - bx)^2$$

can be constructed, the parameter $b$ minimized by least squares, and a Pearson correlation coefficient, $r$, obtained between simulated relevant message and simulated total message. The parameter, $b$, would represent the error term between, for example, time and amplitude. An autocorrelation can also be performed minimizing the error between message power and an amplitude-attenuation factor representing noise.

With power evaluated, we can set whatever boundaries we desire as to speed and cost and make our choice by linear programming.

An example of a low-cost, high-speed retrieval system with fair retrieval power is one based on the key nouns with which an author titles his document. Other key nouns are likely to be found in the same sentences as the title key nouns; therefore, the searcher, machine or human, can reduce the volume of the document to a desired degree of abstractness by selecting the frequency and location of the sentences containing these title nouns which he wishes to extract. A simple experiment was performed by the author, using as an abstract the first sentence containing a title noun in each major subdivision. Questions pertaining to the documents concerned were asked participants in the experiment, some of whom had read the author's abstract; some, the abstract of key words; and some, the entire document. Results of the scoring were roughly comparable for the three categories, for equal reading time. The experiment itself is not so important except as an illustration of the power of the use of key words. Properly categorized, the use of key nouns could become an effective means of speedy, powerful, and, in large volume, relatively inexpensive information retrieval.