DOCUMENTATION
INCORPORATED

1556 CONNECTICUT AVENUE, N. W.
WASHINGTON 9, D. C.
COLUMBIA 5-4577

THE ACTUAL AND POTENTIAL ASSOCIATION OF IDEAS IN INFORMATION SYSTEMS

TECHNICAL REPORT NO. 3

PREPARED UNDER

CONTRACT Nonr-1305(00)

FOR THE OFFICE OF NAVAL RESEARCH

JANUARY
1954

DOCUMENTATION
INCORPORATED

8621 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

TECHNICAL REPORT NO. 3

THE ACTUAL AND POTENTIAL ASSOCIATION OF IDEAS IN INFORMATION SYSTEMS

In Technical Report No. 1 we considered the problem of the relation of the number of possible associations of ideas in an information system to the actual number of associations.

In an information system of n terms, the number of possible associations (using only one mode of association, namely, logical conjunction) is 2^n-1; but we recognize intuitively that there is a much smaller number of actual associations. Suppose in a system of information using 5000 terms to analyze a collection of documents, no individual document required more than 10 terms for a complete analysis of its contents. It would follow that any association of 11 terms would be an empty function, i.e., there would be no information in the system corresponding to any association of 11 terms. This does not give us the number of actual associations, but it sets an upper limit to the number of possible functions which although still very large is much smaller than 2^n-1.

The number of possible associations of 10 terms in a system of 5000 terms is, of course

$$\frac{5000 \times 4999 \times 4998 \cdots 4991}{10!}$$

The magnitude of this number is approximately $\frac{5^{10} \times 10^{30}}{2 \times 10^{6}}$. Since the number of maximum sized associations in a system of information may be less but can never be more than the number of documents in the system; the system would have to cover $\frac{5^{10} \times 10^{30}}{2 \times 10^{6}}$ documents in order to exhibit this number of actual associations of 10 terms. If not all possible associations of 10 terms are actual, then not all possible associations of 9, 8, 7, etc. terms will be actual.

Hence, in terms of the conditions we have established so far we can conclude:

1. There will be no associations longer than 10 terms; and

2. Not all associations of 10 or less terms will be actual.

We can establish an equivalence between the number of maximum sized associations and the number of documents in the system by assuming that each document in the system is analyzed into an association of 10 terms and that each association of 10 terms is unique; that is to say, two items might be analyzed into 9 identical terms, but the tenth term will serve to distinguish the analysis of any document from all the others. Thus, if there are 50,000 documents in the system, there will be 50,000 different associations of 10 terms. We can also calculate the maximum and minimum number of different associations of 1 term, 2 terms, 3 terms, etc. in such a system.

The number of associations of 1 term, 2 terms, 3 terms, 4 terms, etc. in a set of 10 terms is again $2^{n}-1$ or 1023. This

means that a single document represented by an association of 10 terms represents 1023 different sets of associations. If each set of associations were unique, the number of associations in the system would then be 50,000 x 1023. But, we know from our conditions that this isn't true, since, if there are only 5,000 terms in the system, some of the associations in one set will be equivalent to associations in other sets. Consider, for example, the following sets of associations:

A B C D E F G H I J

A B C D E F G H I K

Each set represents $2^{10}-1$ or 1023 associations, but 2^9-1 associations will be common to the two sets. Hence, the number of different associations in these two sets is equal to $(2^{10}-1)+\lfloor 2^{10}-2^9 \rfloor$.

The condition under which we will have the minimum number of associations is, of course, the condition of the maximum degree of common terms in each combination of ten terms. If we varied our initial conditions and supposed that there were 50,009 terms in our system used to analyze 50,000 documents, then the formula for the minimum number of associations would be

$$1023 + \lfloor 49,999 \times (2^{10}-2^9) \rfloor$$

This formula is for the condition in which all 50,000 documents have 9 terms in common and differ by only one term and requires, as we have said, 50,009 terms. With only 5000 terms we can have only 4991 documents having 9 terms in common; for these

4991 documents the number of associations will be

$$1023 + \lfloor 4990 \times (2^{10} - 2^9) \rfloor$$

But it is possible in a system of 5000 terms to have $\frac{5000}{9}$ absolutely unique combinations of 9 terms. Each of these unique sets of 9 terms can be varied 4991 times by the addition of an additional term. To provide for our 50,000 documents with a minimum number of associations, we need utilize only 10 sets, each of which includes 4991 combinations differing by one term. Our formula then is

$$10 \times \{1023 + \lfloor 4990 \times (2^{10} - 2^9) \rfloor\}$$

This number is roughly 25,000,000, which is the minimum number of associations in a system of 50,000 documents and 5,000 terms, when each document is uniquely indexed and all documents are indexed by 10 terms. We need not attempt to calculate the maximum, since we know that it will be less than 50,000,000 or $\lfloor 50,000 \times (2^{10} - 1) \rfloor$ and is of the same order of magnitude as the minimum.

The magnitude of the minimum is significant because it indicates that mechanization cannot take the form of recording and searching for each one of the 25,000,000 associations as individual entities on a punched card or even a magnetic drum.

It will be recalled that we began this investigation with the realization that if we mechanized the manipulation of terms rather than documents, the number of elements to be handled by our machine is appreciably reduced. Now we see that our terms

...rate an alarming number of associations, and the problem ... one of including all associations virtually or potentially ... actually. In fact, what should be mechanical in the ... dictionary is the presentation of any association in ... through the manipulation of a much smaller number of ..., rather than the actual storage and manipulation of all ... associations. The implications and elaboration of this con... will be presented in subsequent reports.