AD NUMBER

AD429837

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; 30 OCT 1963. Other requests shall be referred to Army Electronics Research and Development Laboratory, Fort Monmouth, NJ.

AUTHORITY

USAEC ltr dtd 25 Jan 1967

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

REPORT NO. 5

30 OCTOBER 1963

RESEARCH IN INFORMATION RETRIEVAL

Fifth Quarterly Report

1 July 1963 - 30 September 1963

Contract No. DA 36-039-SC-90787
File No. 1160-PM-62-93-93(6509)

Technical Report 5201-TR-0058

U.S. Army Electronics Research and Development Laboratory

Fort Monmouth, New Jersey

ITT Data and Information Systems Division

Report No. 5 31 September 1963

RESEARCH IN INFORMATION RETRIEVAL

Fifth Quarterly Report
1 July 1963 - 31 September 1963

An investigation
of the techniques and concepts of information retrieval

Contract No. DA 36-039-SC-90787
File No. 1160-PM-62-93-93(6509)
Signal Corps Technical Requirement
SCL-4218 12 January 1960

Technical Report 5201-TR-0069

Jacques Harlow, Principal Investigator

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

1. PURPOSE

1.1 SCOPE

This report discusses the work performed for the U. S. Army Signal Electronics and Development Laboratory under Contract No. DA-36-039-SC-90787 during the period from 1 July 1963 to 31 September 1963.

1.2 OBJECTIVES

The objective of this project is to investigate the techniques and concepts of information retrieval and to formulate and develop a general theory of information retrieval. The formalization of this theory is oriented to the automation of large-capacity information storage and retrieval systems. This theoretical framework will be the basis for the use of general purpose stored-program digital computer systems to perform the storage and retrieval functions.

1.3 PROJECT TASKS

The task structure is based upon the information retrieval model specified in the First Quarterly Report to USAELRDL, the framework elaborated for it in the Second Quarterly Report, and the description of tasks presented in the Third Quarterly Report. The task structure is intended as an organizational guide for continuing investigations. It is not intended to exclude constructive effort in task areas that may not have been foreseen, nor is it likely that all the tasks and subtasks specified will receive equally intensive treatment.

The goal of this project is a theory or a model of a fully automated information content storage and retrieval systems. The task

structure deals with four areas of procedural capability that must be developed if this goal is to be achieved:

(a) Input capabilities

(b) Query capabilities

(c) Processing capabilities

(d) Information retrieval system theory and integration (integrative capabilities)

The first three areas are roughly analogous to the D, E, and F transforms of the basic information retrieval model. The last area is a supra-ordinate category that indirectly involves the other three.

The major tasks and subtasks were described in the Third Quarterly Report; these descriptions will not be repeated in this report. There were no significant changes to the task structure during this period. Therefore, the following paragraphs comment briefly on the work performed on various subtasks during the reporting period.

1.3.1 Input Capability - A large part of the documented effort in the past quarter is in this area. All of the subtasks mentioned under input capability have been considered either explicitly or implicitly in the material of section 4.2 below. As work has proceeded, it has become increasingly clear that input capabilities provide the basic foundation for the functioning of any information retrieval system. It has already been noted that they dominate query capabilities insofar as questions can be no better articulated than the inherent structure of the input analysis (and subsequent input determined processing) allows. In the work of the past quarter we have developed the groundwork for

applying the analysis of input capabilities to the development of efficient integrated information retrieval systems. Thus it is expected that the work on economics of descriptor use and on evaluation of descriptor importance will ultimately lead to a general model of descriptor efficiency.

Work focusing more strictly on the problem of input capabilities includes both an analysis of automatic corrective indexing procedures and an outline of a plan for empirical investigation of our information theoretical approach to automatic classifications. The actual performance of the experimental work outlined in these sections (4.2.3 and 4.2.4 below) is, however, outside the scope of the present project.

1.3.2 Query Capabilities - During the past quarter an analysis was made of the ultimately desired query capability in an ideal information system oriented primarily to fact rather than document retrieval. A system of this kind would require inferential processing capabilities in order to deal with implicit as well as explicit factual content. Some of the problems in the design of such a system are considered and salient issues in the logic of questions and questioning are highlighted. These issues clarify the isolation of a query capability subtask since they are not readily dealt with in any other task category.

1.3.3 Processing Capabilities - No documentation has been produced in this area for the present quarterly report. A good deal of analysis is in progress here. Work in the area of associative techniques has not as yet resulted in a significant original contribution. The value of

the multi-list system in this regard has not yet been completely evaluated. Work on Markov processes continues but is not yet conclusive. It should be noted, however, that the issues raised under query capabilities are relevant to the design of sophisticated processing procedures.

1.3.4 Integrative Capabilities - The significance of the work under input capabilities for the efficient integration of an information retrieval system and for the development of a coherent theoretical model have already been alluded to. There is no specific documentation on this task. It should be emphasized, however, that the eventual problem of integrating the various theoretical or pragmatic aspects of a system are constantly borne in mind. The model presented in the First Quarterly Report is admittedly a simplified concept, but it serves to relate the independent studies being conducted. The interrelationships among these studies are being continuously discussed by staff personnel.

2. ABSTRACT

Documentation in two of the four areas of capability described in the project task structure has been produced in the last quarter. Under input capabilities a plan for the empirical evaluation on procedures for automatic assignment has been developed.

The economics of descriptor usage, importance of ranking of descriptors and automatic corrective procedures have been considered. Work in the latter areas is also considered a significant contribution to ultimate system integration. Query capabilities have been considered from the standpoint of a fact retrieval system and the problem of developing a logic of questioning is discussed. These considerations provide a framework for the requirements of further developments in processing capabilities.

3. PUBLICATIONS, REPORTS, AND CONFERENCES

3.1 REPORTS

The following reports were issued during the reporting period:

(a) RESEARCH IN INFORMATION RETRIEVAL: Fourth Quarterly Report, 1 April 1963 - 30 June 1963, Technical Report 5201-TR-0062, (Manuscript Version), 31 July 1963.

(b) MONTHLY LETTER REPORT NO. 9, 1 July 1963 - 31 July 1963, File No. 5201-TR-0059, 31 July 1963; Research in Information Retrieval, George Greenberg.

(c) MONTHLY LETTER REPORT NO. 10, 1 August 1963 - 31 August 1963, File No. 5201-TR-0063, 31 August 1963; Research in Information Retrieval, George Greenberg.

3.2 CONFERENCES

On 2 August 1963 a conference was held between ITT DISD and USAELRDL in Paramus. The purpose of the meeting was to brief USAELRDL on progress made during the fourth quarter of the information retrieval project. Researchers presented aspects of their work during the quarter which were included in the fourth quarterly report. Plans for the fifth quarter and future activity were also discussed.

During the fifth quarter attendance at the simulation of cognitive processes seminar continued until 26 July 1963. This meeting has already been described in the last quarterly report.

4. FACTUAL DATA

4.1 ORGANIZATION

This section is organized according to the major areas of capability described in the project task structure. This structure is summarized in Section 1.3; a full description of the tasks was presented in the Third Quarterly Report.

4.2 INPUT CAPABILITIES

Work done under input capabilities this quarter includes a section on the economics of descriptor usage, an analysis of the problem of the ranking of descriptors in terms of their importance for information retrieval processes, and an examination of automatic procedures for corrective indexing. The latter contains a brief description of an approach to experimental verification of the validity of the corrective indexing plan making use of available library data only. The final selection is entirely oriented to experimentation and contains a description of a plan for the empirical evaluation of the information theoretical methods of automatic document classification developed under earlier work on input capabilities in this project. The first two sections of the following material on input capabilities are also relevant to the ultimate query capabilities of a system designed with these considerations in mind and especially to the ultimate integration of component capabilities into an efficient working system.

4.2.1 The Economics of Descriptor Usage - The problem of economics of descriptor usage may be stated as follows: Given the set of frequencies

distributed over the members of the power set (on the set of all documents) find the most efficient allocation of descriptors.

In the above statement of the problem the word "efficient" is not exactly specified. What the exact meaning of the problem is will become clear only when the concept of "efficiency" is elucidated.

For the purposes of this note let us assume that the concept of "efficiency" implies the existence of a general Retrieval Utility Function. We define this function (in previous reports on Non-Boolean Retrieval) as a set function over the four categories of sets:

(a) The set of all correctly retrieved documents.

(b) The set of all incorrectly retrieved documents.

(c) The set of all correctly unretrieved documents.

(d) The set of all incorrectly unretrieved documents.

Assuming that we know the form of such functions, we may now conceive our task to be the maximization of this function; i.e., we would like to allocate descriptors in a way that will maximize this function.

Reflecting upon the above formulation, one observes, however, that the problem is not yet completely specified. There is nothing in the statement of the problem which prevents us from assigning a descriptor to every number of power set and thus from obtaining the maximum accuracy of retrieval (Maximizing Utility Function). To make the problem meaningful, we must therefore introduce certain constraints upon the process of allocation of descriptors. This we may do either by introducing explicitly the constraining factors into the Retrieval Utility Function, or by

stating the constraint conditions as separate constraint equations. It is conceptually simpler to take the second alternative, and at least for the time being we shall follow it.

What kind of constraint conditions can we introduce?

(a) There exists a cost associated with the number of descriptors used. The optimum condition is reached when the positive increment in Retrieval Utility Function is exactly balanced by the negative increment in the cost function.

(b) There exists a cost of concatenation. Associated with each retrieval there is a cost which depends upon how many descriptors are used to specify the request.

The process of optimization is the same as under (a).

There are several major drawbacks associated with the constraints specified under (a) and (b); the two most important ones are:

(a) The nature of cost function is unknown. This is so because we have not deduced the existence of these restrictions from some more basic postulates, but rather imposed them arbitrarily on the grounds of empirical feasibility.

(b) The value of the constraints under (a) and (b) is not sufficiently relevant to the problem of descriptor allocation.

This latter statement should be interpreted in the following sense: The Retrieval Utility Function is sensitive to the boundary relations between descriptors, as for example in a decision to apply an available descriptor to a member of a power set which is both infrequently used and is also a subset of member sets which are infrequently used. (In this case, the allocation is obviously inefficient.) On the other hand, the constraints under (a) have to do only with the number of descriptors used but not with their allocation. The constraints under (b) are

allocation sensitive but the difficulty here is that any allocation solution depends upon the nature of the frequency distribution among the members of the power set. In other words, any solution obtained will be valid only for a particular distribution and offhand it is difficult to see how in general any conclusion could be drawn.

At this point two alternatives present themselves. The first one would consist of an attempt to put all possible frequency distributions into a small number of major categories, and then to attempt to get a solution for each. The second would consist of imposing an extra constraint condition which would in a sense "freeze" some of the degrees of freedom with which the allocation activities are carried out. Both of these alternatives are under investigation.

In the discussion above it has been assumed that descriptors are matched with documents correctly. In other words the problem was stated in terms of choosing the best set of descriptors. The question of a descriptor correctly characterizing a document was not at all implicated. Such an approach implies that the connection between a descriptor and a document is established on grounds independent of the descriptor usage. This view is consistent with what one may call a "semantical" approach which views the usage of a term as being determined by its meaning.

In the context of automatic procedures, however, the relation between meaning and usage may be at least partially reversed. The determination of the meaning of a descriptor may have to be inferred from the way it is being used.

It seems natural then in this context to view indexing as being essentially probabilistic. Once the strictly semantical point of view is abandoned, the descriptors apply to documents with a continuous range of probablities.

The next two sections will delve into the implications of the point of view expressed above.

4.2.2 Descriptor Ranking in Terms of Their Importance For Information Retrieval Processes

4.2.2.1 General Background - Let us imagine a large collection of documents classified in some fashion. Each document in this collection is labeled by one or more descriptors. It is intuitively evident that not all descriptors are of equal importance. The deletion of some would result in almost no harm to retrieval processes. The deletion of some others would be, however, very detrimental.

As a result of the library's growth or changed usage, librarians might wish to append new descriptors to the documents or possibly though less urgently to delete some others. One tends to think of such processes as being primarily dependent upon the subject matter contained in the documents. Undoubtedly this manner of thinking was well adopted to manual systems and relatively small collections. The emergency of automated retrieval systems and a vast increase in sizes of document collection creates, however, problems of a different kind.

The unchecked proliferation of descriptors may have actually

diminished the usefulness of a library, either by lengthening the physical processes involved in retrieval, by confusing the taxonomical logic of the collection, by simply straying too far from the natural usage of terms or for a number of other reasons. In any case, and for whatever reason the librarian may wish to restrict the number of new descriptors which must be introduced in order to keep the retrieval processes near the peak of efficiency.

Under such conditions the choice and even the allocation of descriptors may be governed by the criteria of descriptor importance mentioned above. In addition, the criteria utilized in automatic indexing procedures may of necessity lean more towards utilization of statistical type of information about the collection than is the case when indexing is done manually. To put the same ideas differently and more strikingly, when indexing is performed automatically the governing criteria may pertain more to statistical distributions of descriptors among the documents than to explicit relation between the subject matter of a given document and a descriptor.

This may be an overstatement. Still the two adduced reasons provide enough incentive to initiate the investigation of the problem. The important questions which ought to be answered are:

(a) Which factors govern the criteria of relative importance of descriptors?

(b) How can these factors be expressed formally and converted into the quantitative measures?

(c) In what way can these criteria be used to govern automatic selection and allocation of descriptors to documents?

(d) What statistical data is required for the determination of the order of descriptor importance?

Of these questions (a) and (d) will be dealt with in this section. The third question has been considered in the section dealing with Automatic Corrective Indexing Procedures. The second one will be delayed into the future.

4.2.2.2 Factors Which Govern the Criteria of Relative Importance of Descriptors - We can start the investigation of this question on the intuitive level.

(a) Let us suppose that a certain descriptor is never mentioned in any of the retrieval requests. Obviously such a descriptor could be deleted from the collection without any ill-effects for the retrieval processes. Conversely, descriptors used with high frequencies have a high probability of being important. At the present stage, we can only speak of the higher probability of importance since the relation of various factors to each other has not yet been formalized. So far as frequency relations are concerned, a certain assymetrical situation exists. Below a certain frequency threshold the frequency considerations are overwhelming. If a descriptor is not used with a certain minimum frequency it cannot be ranked high. However, the high frequency descriptors are not necessarily of importance. For example, a high frequency descriptor may be aynonymous with another descriptor.

(b) Descriptors are usually employed jointly. The importance of a descriptor is influenced by "the company it keeps." A descriptor may have little relative discriminatory power vis a vis descriptors that co-occur in a representative retrieval request. For example, let us assume that a certain descriptor say D is used jointly with descriptors:

$A_1A_2A_3A_4$

$B_1B_2B_3B_4$

and

$C_1C_2C_3C_4$

Let us assume that the increment of the retrieval collection due to the deletion of D is in each of the cases from 490 documents to 500 documents. The average "actual discriminatory power" of the D-descriptor is low.

(c) The average number of descriptors used in retrieval calls containing a given descriptor is an important indicator of the order of importance. Other things being equal, one may expect that a descriptor which co-occurs with large numbers of other descriptors in retrieval requests is of lesser importance than those which co-occur with few.

The above considerations dealt with descriptors as used in retrieval calls. These have to do with the actual usage of descriptors. We wish to distinguish these considerations from those pertaining to the potential usage. The next set of factors will deal with factors not related too directly to actual usage. These factors are dependent only upon the

distribution of descriptors among documents and not with their occurrence in retrieval calls.

(d) The larger the size of a document set spanned by a descriptor, the greater will be its ranking on the importance scale.

(e) Corresponding to the "actual discriminatory power" of a descriptor there is the "potential discriminatory power." This is a measure of unique coverage which is due to the given descriptor. Suppose that a given descriptor is deleted, certain sets of documents which could be previously retrieved, can now only be retrieved as subsets of other retrievable sets. For example, let us imagine a descriptor which spans a set of documents in such a way that its intersection with every possible intersection of the sets spanned by other descriptors is not a proper subset of such intersection. Such descriptors have clearly discriminatory power of zero, since every set of documents which can be retrieved by using it can also be retrieved without its assistance.

(f) A set spanned by a descriptor may intersect sets spanned by closely related descriptors or by sets spanned by descriptors remote from one another. We may call such characteristics a measure of dispersion of a descriptor. Other things being equal, the more dispersed a descriptor is the less highly will it rank. This is so because with high dispersion in any particular retrieval call the higher proportion of retrieved documents may be expected to be only

marginally relevant to the request.

4.2.2.3 Statistical Data Required For the Determination of the Order of Descriptor Importance - Unfortunately not all the factors mentioned in the preceding section can be conveniently measured. For some the amount of bookkeeping required is too close to astronomical to be of practical consideration. Therefore, one must take recourse to convenient substitutes, which encapsule the essential information without too much leakage, and at the same time reduce the requisite amount of data handling and bookkeeping.

The important consideration that has to be kept in mind is that detailed accounts of intradescriptor relationships cannot be kept. For example with 10,000 descriptors there are $2^{10,000}$ possible combinations or descriptors and if even .01% of these are active (i.e., there are some documents which are indexed by them) the number of entries which would have to be kept is astronomical. We wish therefore to keep track of selective data on the basis of which the important intradescriptor relationships could be approximately reconstructed.

The most difficult problem will consist of trying to reconstruct the "dispersion" and the "discriminatory power" of the descriptor set. Tentatively, the following set of parameters is suggested as a basis.

(a) Total document span of individual descriptors.

(b) Frequency of recall of individual descriptors.

(c) The number of documents spanned by a given descriptor in company with either k descriptors where k is 1, 2, etc.

(d) The document span of an average descriptor contained in a set of k of them present with a given descriptor.

(e) The frequency of recall of an average descriptor contained in a set of k of them present with any given descriptor.

(f) The number giving an overlap measure of an average descriptor contained in a set of k descriptors present with a given descriptor.

4.2.2.4 Summary and Conclusions - Statistical properties of descriptors have been proposed as a basis for ascertaining the ranking of descriptors in terms of their importance in the retrieval processes. Specifically the concepts of descriptor's "discriminatory power" and of its "dispersion" have been defined. Several sets of data about descriptors have been suggested as a feasible basis on which the estimation of ranking parameters could be based.

4.2.3 Automatic Corrective Indexing Procedures - The objectives of this section are:

(a) To set the broad outlines of Automatic Corrective Indexing Procedures.

(b) To indicate what connection there is between measures of descriptor ranking on descriptor importance scale and the procedures outlined under (a).

Let us imagine the following process: The library user is permitted to state his request in terms of any descriptor he chooses. Some of the descriptors he chooses to characterize his request are already contained in the existing retrieval vocabulary, some others are not. The new terms used in the request are then cross-tabulated with respect to the presently employed descriptors. One would wish if it were all feasible

to register the information regarding the joint usage of the new term with every possible combination of all descriptors. Such procedures, however, are obviously too cumbersome and the amount of data which it is necessary to store too bulky for practical consideration. Obviously then we shall have to distill the essential information so as to reduce the complexity of attendant data handling to manageable proportion.

The most important piece of information about a new term is the frequency of its distribution with respect to other descriptors. We also wish to have some information concerning combinations of descriptors mentioned jointly with the new term. Such information can be partially conveyed by registering in addition to its frequency, the average number of descriptors which appear in the requests containing the new term and a given descriptor.

Finally, it is important to keep track of the number of times (relatively to other new terms) a term is mentioned. We may conveniently think of a new term being represented by a complex vector

$$\vec{d}_j = p\left[(f_1,g_1), (f_2,g_2),\ldots,(f_i,g_i),\ldots,\right]$$

where

p = modulus of the vector = the normalized percentage of times the new term is mentioned.

f_i = the frequency of co-occurrence of the i^{th} descriptor with the new term,

g_i = the average number of descriptors co-occurring with a given term and the i^{th} descriptor.

There are two decisions which must be made

(a) Which of the new terms are to be selected as additional descriptors?

(b) To which documents should a new descriptor be appended?

So far as the first decision is concerned, the most obvious factor influencing it is the modulus p of the vector $\vec{d}_j$, for it is this number which indicates the frequency with which a term is mentioned in requests. Yet it is not the only factor, and the selection decision must be based on more comprehensive grounds.

First of all a term frequently mentioned in the retrieval requests and thus a candidate for a new descriptor may be highly synonymous with another term in use. The suspicion that this may be the case will be based on the inspection of the vector belonging to the term. If any of the f_i's occurring as vector components is close to 1, one may suspect that the term act as synonym to the $i^{\underline{th}}$ descriptor. This is only however at best a necessary but not sufficient condition. For it is possible that the $i^{\underline{th}}$ descriptor is used more broadly than the new term. It is not as yet entirely clear how synonymity could be distinguished from the broader usage case with certainty. The fact that the modulus of the descriptor is larger than the modulus of the new term does not enable us to infer that the descriptor is used more broadly. It may well be that some of the users will not think of the synonymous terms writing their requests. The final resolution of this matter may proceed essentially along two lines not necessarily exclusive. The first would submit this question to an empirical investigation hoping that there exists some critical ratio between the two moduli (i.e., the descriptor vector and the new-term vector) which will serve as a dividing line between the synonymity

case and the broader use case. The second approach will utilize the "internal evidence." Suppose that in addition to being highly correlated with a certain descriptor the new term in question has a similar profile to it. The similarity of profile is determined by comparing the two vectors (but not their members) upon one another as the criterion, namely the similarity between the (f_1, g_1) components. Explicit methods of determination are under study.

Secondarily, one would not like to base the selection decision entirely upon the evidence of new terms contained in the requests. At least partially one would like to utilize the implicit evidence contained in the manner present descriptors are being employed.

We now turn to the second problem: To which documents should the new description be appended? It is assumed that this problem is to be resolved by automatic means without recourse to human judgment.

The solution to this question lies along the lines of trying to match as closely as possible the descriptor profile of a new term with the description profile of a document. What such procedure means will be best understood by considering a population of requests containing the new term. On the basis of the statistical information contained in the term vector $\vec{d}$, it is possible to approximate resolution of this population into specific requests. That is to say, a set of specified requests is reconstituted each carrying a numerical weight in proportion to its frequency. Now, it seems reasonable to suppose that documents which contain a larger proportion of descriptors also present in requests

containing the new term are more likely to contain it than those with a small proportion. Thus it will be possible to assign the new term under consideration to documents with a certain probability value. The assignment of probabilities is a rather complex process depending upon many factors. We shall discuss them below qualitatively, leaving the exposition of the actual computational processes for further development.

(a) Any combination of descriptors has an a priori probability of co-occurring on any document with $1,2,\ldots,k$ extra descriptors.

(b) To any combination of descriptors one may correspond the average number of different descriptors co-occurring with it. Both quantities are computational.

These parameters express roughly expectancy of another descriptor appearing on a document on which a given combination of descriptor appears. Now, if we assume that the shift, occasioned by the introduction of the new term, in these expectancy numbers and probabilities is either nil or very small, then these parameters will serve as very useful guides in the process of assignation of new terms to documents.

The Automatic Corrective Indexing Procedure outlined and the theory underlying it are at present not grounded in empirical corroboration. At some stage empirical corroboration will become absolutely required not only to test the soundness of the fundamental assumptions but also to choose between competing alternative assumptions and to yield numerical values for the parameters since these can in no way be deduced theoretically.

The Automatic Corrective Indexing Procedure may be tested by applying it to existing libraries. This is how the experiments would be conducted.

Library catalogues or files would be scanned with the purpose of gathering required statistics. However, in gathering the statistics some of the present descriptors would be treated as non-existent. Likewise, statistical data concerning requests would be collected. The requests containing deleted descriptors would be either simply omitted or treated as ordinary requests with the deleted descriptors ignored. At some point after statistical material has been compiled the hitherto deleted descriptors will be treated as new terms. The procedure outlined in the preceding section will be followed and new descriptors selected and applied to documents. It will then be possible to compare the original allocation of descriptors to the one which resulted from the application of the Corrective Procedures.

4.2.4 Empirical Evaluation of Information Theoretical Methods of Automatic Document Classification

4.2.4.1 Purpose - The purpose of this section is to present plans for an experimental evaluation of previously proposed techniques for classifying documents automatically using information theoretical methods.

4.2.4.2 Introduction - In previous reports certain information theoretical methods of document classification were presented. These methods made use of word occurrence and word frequency information as clues to the classification of a document. The methods were based entirely on a theoretical analysis of the document classification problem; no experimental evidence as to the effectiveness or practicality of these methods was introduced.

In reviewing the literature on automatic document classification, two articles were found which were of special interest. They were "Automatic Indexing" by M. E. Maron [6], and "Automatic Document Classification" by H. Borko and M. Bernick [2]. These researchers used statistical techniques to find the correlation between word occurrence in a document and document categorization. Maron used Bayesian techniques, while Borko and Bernick used factor analysis techniques. These methods were applied to the same set of data, and the results of each method were compared in reference [2].

The information theoretical methods of document classification fall into the same theoretical framework as the above two methods. They represent another way of statistically analyzing the data. Then a test of the information theoretical methods which should be of great interest would use the same set of data that Maron and Borko and Bernick used, permitting the comparison of all three methods. In the following sections, Maron's experiments and Borko's and Bernick's experiments are summarized, the new experiments are described, and the expected results are indicated.

4.2.4.3 The Data - Both references [2] and [6] used abstracts of computer literature published in the IRE Transactions on Electronic Computers [5]. There were 405 abstracts, which were divided into two groups. Group 1 consisted of 260 abstracts published in the March and June 1959 issues of the Transactions; Group 2 contained the remaining 145 abstracts. Group 1 was known as the Experimental Group, Group 2 as the Validation Group. The documents of the Experimental Group were analyzed

and the clue words and classification system were derived on the basis of this analysis. The Validation Group was then used to test the effectiveness of the statistical procedures; this group of documents had been put aside while these procedures had been developed with the Experimental Group.

4.2.4.4 Maron's Experiment - Maron discarded the classification system that had been published with the abstracts and chose a finer categorization of 32 categories, which he felt better reflected the nature of the abstracts. He next chose "clue words," or words whose occurrence could predict the categorization of a document. First he eliminated all high frequency words like "and," "the," etc., together with words that were very common like "computer," "machine," "system," etc. Then extremely low frequency words were eliminated because they would be inefficient. The remainder was listed showing the number of times each one appeared in a document in a particular category, and from this list, the words that seemed to peak in particular categories were chosen. No automatic techniques were used ~~for this selection; the 90 clue words~~ were chosen from this list by inspection. Then, using the probabilities associated with each of the clue words in a Bayesian prediction formula, he predicted document category using first, the experimental group and later the validation group. The prediction formula gave the probability of a document falling into a particular category given that certain clue words had appeared in it. The category with the highest probability was chosen. Then these results were compared with the results obtained by human indexers.

4.2.4.5 Borko's and Bernick's Experiment - Borko and Bernick used the same 90 clue words that Maron used in their experiment. Using the statistics from the Experimental Group of these 90 clue words, a 90 by 90 correlation matrix was set up, measuring the correlation of each clue word with the other clue words. This matrix was then factor analyzed [1]. A set of 21 orthogonal factors was obtained which was felt to be meaningful and adequate when interpreted as a set of classification categories.

A prediction formula was chosen, a function of the sum of the products of the normalized factor loadings and the index term. Using this prediction formula, the category with the highest score was chosen. This scheme was used for both Experimental and Validation Groups, and the results were compared with the results obtained by human classification of the documents into the 21 factor derived categories.

Borko and Bernick also proposed other experiments on the same data base to shed light on certain questions that arose during the original experiment. One experiment would use the 21 categories already derived, but select clue words in the manner suggested by Maron, and would then use Maron's prediction equation. The second experiment would factor derive a new classification system, based not on Maron's 90 clue words but on clue words derived on a frequency basis. This classification system would then be used with both prediction schemes.

4.2.4.6 The Proposed Experiment - A brief description of the information theoretical classification methods and the proposed experiments is contained in reference [7]. The information theoretical methods assume

a given classification system. In this case, three classification systems would be used:

(a) Maron's 32 categories.

(b) Borko's and Bernick's 21 categories used in their experiment.

(c) The categories derived by factor analysis on a frequency basis in Borko's and Bernick's proposed experiment.

Based on these three classification systems, three sets of 90 clue words would be derived by applying the two information theoretical measures of reference [7] to word occurrence information for all words appearing in the abstracts. These words would be compared with the sets obtained by the other researchers. It is expected that the set of words obtained by information theoretical methods would be fairly similar to the sets of words obtained by Maron's methods. The selected words would then be used with Maron's prediction equation, as well as certain empirical prediction equations.

The results, using the human classification already performed in the previous research as the criteria of correctness, will then be compared with the automatic classification results of the previous research. It is expected that the results would be close to those obtained by Maron's methods, because of the basic similarity in method, although an improvement is expected because of the more methodical selection of effective clue words some of which may have been overlooked by Maron's method. In addition it may be possible to use a more effective prediction formula than the one Maron used.

After the results of these initial experiments are analyzed, word frequency statistics might be used to determine clue words, following the methods outlined in reference [7]. In addition, the experiments should be repeated, using not just 90 clue words, but using those words which have information theoretical measures beyond a certain cutoff point. Both these last experiments should lead to improved classification results, but the exact path they take must be determined on the basis of analysis of results from the first set of experiments.

4.2.4.7 Summary - A test of information theoretical methods of document classification has been proposed using the same data used by other researchers in the field. The new results will be compared with those obtained by Maron and Borko and Bernick. It is estimated that the results will be similar to those obtained by Maron because of the basic similarity of method; however, an improvement is expected because of:

(a) A more methodical procedure for selecting clue words.

(b) A possibly improved prediction equation.

Further improvement may be possible using more detailed statistical data in the information theoretical approach.

4.3 QUERY CAPABILITIES

4.3.1 User Orientation - The users of an information system are often conceived as a univocal mass that knows precisely what type of information it wants from the system. The problem of system design is then reduced to the simple expedient of devising means of access to the general body

of stored information for this class of users.

In fact, however, the users are neither univocal nor certain; if they were, the problem of information retrieval would be vastly simplified. Any intermediary for gaining access to stored information would be superfluous, since the users by definition, have a priori knowledge about the nature of the information they seek. The difficulty is that users approach any information system--even a library card catalogue--because their questions are vague and ill formed. Furthermore, each user wishes to fulfill a different need.

In confronting a new system, any user is wary at first; the mechanism of the system stands as a barrier (and possibly a threat) between his questions and whatever answers may be available. The first criterion for gaining the user's confidence, then, is simplicity; the mechanics of the system should be readily grasped after a few moments of study. The second criterion is that the user quickly gain confidence that the system can indeed produce reasonable responses to reasonably well formed queries.

This second factor poses the greatest difficulty. If a user has confidence in the system, he is willing to enter a tacit dialogue. A simple question, however ill formed, produces sufficient information to lead to another, more cogent question. The dialogue continues from question to answer to question until the user eventually frames precisely the right question to gain access to the information he originally sought. This process with the familiar card catalogue is heuristic; the same

process should occur with an automated system, but the interposition of a machine may easily restrain the facility of the dialogue.

An information system deals with the functional elements of information in such a way that a sequence of operations upon these elements or upon concatenations of these elements produces the requested information. What is desired is information explicitly or implicitly contained in the data received by the system. Thus, ultimately, logical implications, generalizations, correlations, and even logical appraisals of the original data (credulity measures and ordering relations) may be the results of these operations.

The requirements for performing operations upon the information parallel, at least in part, those for storing information. These operations should be defined so that information can be recombined into forms that are not explicitly formed in the original information. Such processing operations should be specified in relation to the storage operations. The retrieval processes may then gather relevant material from the stored data so that it may be operated upon and used to answer questions. Some of these operations are based upon statistical analyses of the data. Other operations are functions performed upon the question in order to improve the formulation of a query. In this way the inherent difficulties in establishing a dialogue between the user and the system may be reduced, if not entirely eliminated.

Additional operations on information may be necessary. The system may be expected to derive logical relationships existing among data

contained in its memory. In addition to logical inferences (deductions), the system may be expected to perform inferential processes (inductions). Such inductive inferences differ from deductive inferences in two important respects: the relationships derived are not necessarily valid; and not all the rules of inductive reasoning are explicitly formalized.

Implied relationship is a generic term for all relationships not explicitly contained in a system. Such relationships are derived by means of inferential processes, i.e., inductions and statistical correlations. The term implied relationship includes relationships derived on the basis of inductive, or non-rigorous, inferential processes. Such relationships are by their nature not as well defined as relationships obtained deductively. The system must, therefore, be designed with the capacity to estimate the degree of credibility of such derived relations and the degree of relevance to other information. On the basis of such estimates the system may accept or reject the derived conclusions.

Since the set of implied relationships is not well defined, such a system will arbitrarily limit the range of derivable relationships. It cannot be expected that the system will attempt to derive all the implied relationships that lie within a specified range without being requested to do so, either directly or indirectly, in terms of a question. On the other hand, some of the implied relationships might be so important to the functioning of the system that they ought to be derived even without any initiating query. An information system would, therefore, be more powerful if it possessed a set of decision algorithms for determining at

which point it must stop its inferential activities.

It is necessary to state the criteria employed to select the relationships the system will derive. While the set of explicit relationships stored in the memory of a system may be well defined, the corresponding set of implicit relationships may not be. The derived implicit relationships depend not only upon the set of explicit relationships, but also the nature of the formal or informal inferential methods as well as upon other factors--e.g., the richness of association--less amenable to precise description. Because of these factors it may be questioned whether the notion of the set of all implicit relationships derivable from the information is meaningful. From a practical viewpoint, some limitations upon the range of implicit relationships must be imposed.

The criteria for the limitations that are to be imposed upon a system's ability to derive implicit relationships ought to include:

(a) Only implicit relationships possessing potential utility to the users of the system should be derived.

(b) The system should not try to derive implicit relationships of so complex a nature that the attempt is likely to end in failure.

(c) The limitations should be flexible enough to leave room for learning.

The system may be able to increase the range of derivable implicit relationships as it obtains more input information or elicits more information about a question from the user; again the importance of a dialogue is apparent. The criterion for the selection of derivable relationships, which includes all three of these characteristics is: The system is only concerned with those implied relationships that can be derived in response

to a definite procedure specified by the user. This principle may be considered as the organizing principle of the system.

There are several points that will clarify the meaning of this principle. In addition, the adoption of this principle has certain implications for the learning processes that will take place in an information system. The phrase, "....in response to a definite procedure specified by the user," does not mean that the user is obliged to supply the directives that could be directly translated into programs--that is, a sequence of action resulting in an output consisting of the appropriate implicit relationships. Neither does it mean that such a specification need be supplied to the system initially.

The principle simply states that the user knows how to go about solving the problem embodied in a query addressed to the system; he knows how to solve the problem in terms of human mental processes. Moreover, the principle does not require the user to state the procedure formally. The concept of knowing how to go about solving problems implies no more than that the user know enough about his own procedures to answer questions about his approach to the problem.

4.3.2 A Concept of Questioning - In order to optimize the retrieval ability of a system, the user should question the system within the framework of a theory of questioning. The development of a concept of questioning has occasioned considerable scientific interest within the last decade. In part, such an interest is related to problems of retrieving information, for even a cursory examination of questioning indicates that

its plays an important role in the retrieval of information. Every pragmatically important question has a correct answer associated with it. Such a correct answer is a statement that provides a person with information--knowledge that he did not possess at the time that he asked the question. The statement may be true or false and still fulfill this criterion. Given a framework of this kind, the concept of questions requires a development along two parallel lines: the semeiology and the methodology of questions.

The semeiology of questions pertains to the form and nature of queries. Questions are a type of linguistic structure. Composed as they are of signs--letters and words--questions have meaning. Such meaning may be even more complex than the meaning of declarative statements, since questions may also be logical functions of such meanings.

There are two possible ways to investigate the meaning of a question. A question may be correlated with a class of statements, any one of which is a correct answer to the question. In this sense, the question defines the scope of possible answers; it is neither responsive nor meaningful to answer the question "What time is it now?" with the statement "The Parthenon is located in Athens, Greece." On the other hand, there are questions that do not define the kind of statement that is a correct answer. Consider the question "How many horns does a unicorn have?" "There are no such things as unicorns," is as correct an answer as, "A unicorn has one horn." In other words, a question may pragmatically admit unclarity about the boundaries of a subject. Only procedurally correct

questions request information within a framework of concepts and statements accepted as true by both the questioner and the informer.

The realization that a question is related to a given state of knowledge requires further exploration. It is clear that a question is meaningful only if the questioner refers to a set of interrelated concepts either explicitly or implicitly. When a questioner asks "What time is it?" he knows that the answer is a set of numbers that have a certain order--for example, "later than." But it remains a problem whether some concept must be assumed explicitly or implicitly for any question to be meaningful. It may be that in order for a question to be meaningful, some restriction of its scope must be present.

The meaning of a complex term is not only determined by its relationship to non-linguistic factors, but also by its logical relationship to other terms. The meaning of questions is in part specified by their logical or syntactical relationship to other questions. What is required, then, is a formal logic of questions. Such a logic would rigorously formulate:

(a) The syntax of a formal language into which questions in natural language are translatable.

(b) The rules of deduction for such a language.

(c) The theorems concerning logical relations formulatable in such a system.

It seems that the language in which the logic is formulated may be constructed out of declarative sentences by the use of an undefined logical operator [3,4]. Logical functions analogous to deduction can then be defined. In any system the correlation between questions and

permissible answers must be formally modeled by mapping a question on a set of sentences. Semantically, at least, the range of variables should also be specified for answers that are specifiable for standard types of questions.

In addition to logical deducibility that would be studied by such a calculus, there is another dimension of logical analysis. This area pertains to the relative complexity of questions. It may be, for example, that in a certain context a Why question is translatable into a finite set of How questions. In this context, Why questions are more complex than How questions. But there are many types of questions. In addition, there are disjunctive and conjunctive questions as well as general and particular questions. This brief discussion indicates that a logical theory is necessary to consider problems of this kind systematically.

Once a formal analysis of questions has been developed, it will provide insight into the methodology of questions. If the questions that imply other questions are known or are reducible to other questions, then it is easier to develop strategies for sequencing questions so as to obtain maximum information for a minimum set of questions. It is advantageous for any information processing system to allow this condition to be fulfilled.

Besides purely logical and formal considerations, there is a problem of methodology--the strategy or heuristic of interrogation. This problem centers on the problem of efficiency and purposefulness in interrogation. The main objective is to relate the formal characteristics of questioning

to intentions that the questioner may have. From the nature of the problem it is evident that, unlike the inquiry into formal properties of questions, this discussion is mainly concerned with sequences of questions.

There are two types of goals that can be associated with the procedure of interrogation. The first is the desire to obtain more factual information. A simple example of this type of interrogation is: "How many people reside in Rome?" The second goal is to obtain a better understanding of a certain area of inquiry. This objective may be related to the interrogator's perception of gaps in the flow of information or to his lack of understanding of the information. Efficient and intelligent questioning depends upon the precision with which the interrogator can pinpoint the kind of information he wants as well as upon his ability to formulate the appropriate sequences of questions.

The objective of this concept of questioning is to establish procedures for an interrogator to discern the intention of his interrogations. The concept is not psychologically oriented. The problem is not to correlate subjective states of mind with the objectives of the questioning process. The concept seeks to associate the properties of sets of information with the rational formulation of interrogative intentions. These intentions are then fulfilled if the sequence of questions is appropriate for its purpose.

The ordering and the retrieval of information depend upon initially specified rules for information handling. These rules may not be the only rules for data handling necessary for the proper and efficient

operation of an information system. The system must be able to acquire new rules and modify old rules as it continues to process information. The acquisition of rules may be divided into two categories.

One category includes processes based upon success-failure criteria. In processes of this kind an information system attempts to improve its performance without an interchange of complex questions with the user. If the criteria for adequate performance are not satisfied, the system seeks to improve its performance solely on the basis of its store of data and its own experience.

The second category includes processes based upon a system's attempt to elicit information pertinent to the formation of adequate processing rules from the user. Such processes are more complex than those in the first category. In addition to being able to use its own experience, the system is able to question human beings and to use human guidance. In this way the essential dialogue between a system may head to the necessary well formed questions that will elicit the required information for the user.

The implication of this discussion is that the user-system dialogue will necessarily span a range of questions over a period of time, however short the time. But this implied constraint need not follow. A simple question may be simply answered; yet in a simple question the necessary clues to the relevant information are almost apparent. Consider a slightly more difficult instance. If the system contains N categories of information, then N! question combinations are possible. The information may also be stored so that a relation (A,B,C,...) holds. The query may be framed

(C,B,A). A simple response would state: "If your request could also be (A,B,C), then your answer is..." This approach appears too easy, but it is not uncommon. And if these functions were automated, the demon of interrogation could be greatly simplfied.

4.4 REFERENCES

(1) Borko, H., "The Construction of an Empirically Based Mathematical Derived Classification System," Proceedings of the Western Joint Computer Conference, May 1962.

(2) Borko, H., and Bernick, M., "Automatic Document Classification," Journal of the Association for Computing Machinery, Vol. 10, No. 2, April 1963, pp. 151-162.

(3) Harrah, D., "The Logic Questions and Answers," Philosophy of Science.

(4) Harrah, D., "The Logic of Questions," Proceedings of Congress of Philosophy.

(5) Institute of Radio Engineers, "Abstracts of Current Computer Literature," IRE Transactions, EC-8, Nos. 1, 2, and 3, 1959.

(6) Maron, M. E., "Automatic Indexing: An Experimental Inquiry," Journal of the Association for Computing Machinery, Vol. 8, No. 3, July 1961, pp. 404-417.

(7) Trachtenberg, Alfred, "Automatic Document Classification Using Information Theoretical Methods," Proceedings of the American Documentation Institute 26th Annual Meeting, Chicago.

5. CONCLUSIONS

All areas of capability have been extended by analytical studies of the aspects of the information retrieval problem that required fuller definition and articulation. Input capabilities have been specifically analyzed in terms of the economics of descriptor usage, the ranking of descriptor importance and the development of automatic corrective procedures. The work on automatic classification based on information theoretical evaluation of clue words has been brought to a level of specificity that makes it possible to describe a plan for empirical validation of this work. Actual experimentation along these lines is presently regarded as being outside the framework of this project.

Query capabilities have now been explicitly analyzed as required by a sophisticated fact retrieval system. Processing and integrative capabilities are under development but results have not yet been sufficiently conclusive for inclusion in this report. Even the contributions that have been documented are still essentially in the analytical and research stage. The only exception here is the work on empirical evaluation of automatic classification. Actual performance of this experiment requires expansion of the scope of this project or separate project implementation.

6. PLANS FOR NEXT QUARTER

Activities during the next quarter will proceed with the over-all goal of developing a theory of information retrieval for use as a tool in the design of information retrieval systems. This work will proceed within the specific task framework described in the Third Quarterly Report. Part of the emphasis will continue to be analytical with the primary purpose of developing methods to evaluate the relationship among significant system parameters. The major orientation will, however, shift toward the integration of the material produced thus far and toward the development of integrative capabilities in information system design.

Under input capability work on clue word selection is essentially complete except for possible expansion of the scope of the project to include the empirical work described in this report. Such a decision can only be arrived at in collaboration with USAELDRL. Further development is planned for the concepts of corrective procedures based on the statistics of descriptor usage. Quantitative measures of the discrimination and dispersion of a descriptor will be developed for inclusion in the general model of efficient information system design. Rational predictive formulas for going from clue words (selected by information theoretical methods already described) to categories may also be developed.

Under query capabilities no extension of the reported work on automatic extracting is planned. Further extensions on the logic of questioning is being considered. The formulation of a general theory of descriptor languages based upon frequency and accessibility will have important implications for improving query capabilities.

Under processing capabilities the evaluation of multi-list and Markov process techniques will be continued. If there is no conclusive breakthrough in these areas, work on them will be terminated in the next quarter. The requirements raised in the present section on query capabilities may also be considered from the viewpoint of processing.

Under integrative capabilities it is planned to attempt more rigorous formulations of a system model establishing the relationship between frequency and indexing. Further development in the area of general theoretical considerations may also be expected.

7. IDENTIFICATION OF PERSONNEL

7.1 PERSONNEL ASSIGNMENTS

The following personnel were assigned to the project during the period covered by this report:

Name	Title	Man-Hours
Jacques Harlow	Manager	30
Quentin A. Darmstadt	Research Specialist	50
George Greenberg	Senior Specialist	120
Maralyn Lindenlaub	Senior Program Analyst	400
Alfred Trachtenberg	Senior Program Analyst	400
Alexander Szejman	Senior Specialist	430

7.2 BACKGROUND OF PERSONNEL

The backgrounds of personnel assigned to the project were described in previous reports.

DISTRIBUTION LIST

Receipient	Copies
OASD (R&E) Rm 3E1065 Attention: Technical Library The Pentagon Washington 25, D. C.	1
Chief of Research and Development OCS, Department of the Army Washington 25, D. C.	1
Commanding General U. S. Army Materiel Command Attention: R&D Directorate, Res Div, Elect Br. Washington 25, D. C.	1
Commanding General U. S. Army Electronics Command Attention: AMSEL-AD Fort Monmouth, New Jersey	3
Commander, Armed Services Technical Information Agency Attention: TIPDR Arlington Hall Station Arlington 12, Virginia (Reports)	10
Commanding General USA Combat Developments Command Attention: CDCMR-E Fort Belvoir, Virginia	1
Commanding Officer USA Communication and Electronics Combat Development Agency Fort Huachuca, Arizona	1
Commanding General U. S. Army Electronics Research and Development Activity Attention: Technical Library Fort Huachuca, Arizona	1

Receipient	Copies
Chief, U. S. Army Security Agency Arlington Hall Station Arlington 12, Virginia	2
Deputy President U. S. Army Security Agency Board Arlington Hall Station Arlington 12, Virginia	1
Director, U. S. Naval Research Laboratory Attention: Code 2027 Washington 25, D. C.	1
Commanding Officer and Director U. S. Navy Electronics Laboratory San Diego 52, California	1
Aeronautical Systems Division Attention: ASAPRL Wright-Patterson Air Force Base, Ohio	1
Air Force Cambridge Research Laboratories Attention: CRZC L. G. Hanscom Field Bedford, Massachusetts	1
Air Force Cambridge Research Laboratories Attention: CRXL-R L. G. Hanscom Field Bedford, Massachusetts	1
Headquarters, Electronic Systems Division Attention: ESAT L. G. Hanscom Field Bedford, Massachusetts	1
Rome Air Development Center Attention: RAALD Griffiss Air Force Base, New York	1

Receipient	Copies
AFSC Scientific/Technical Liaison Office U. S. Naval Air Development Center Johnsville, Pennsylvania	1
Commanding Officer U. S. Army Electronics Materiel Support Agency Attention: SELMS-ADJ Fort Monmouth, New Jersey	1
Director, Fort Monmouth, Office USA Communication and Electronics Combat Development Agency Fort Monmouth, New Jersey	1
Corps of Engineers Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Marine Corps Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
AFSC Scientific/Technical Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Logistics Division Fort Monmouth, New Jersey	9
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Director of Research/Engineering Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Technical Documents Center Fort Monmouth, New Jersey	2

Receipient	Copies
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: SELRA-NPE Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Technical Information Division Fort Monmouth, New Jersey	3
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Exploratory Research Dr. Reilly Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Engineering Sciences Department Mr. Hennessy Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Exploratory Research Jack Benson Fort Monmouth, New Jersey	2
Headquarters, Aeronautical Systems Division Air Force Systems Command, USAF Attention: ASRCEM-1 (Mr. Thompson) Wright-Patterson Air Force Base, Ohio	1