CHAPTER 1

INTRODUCTION

1. The Document Retrieval Problem

The attention devoted to document retrieval systems in recent years is based in some form or other on the assumption that such systems can satisfy needs exhibited by a class of users or can satisfy needs likely to be exhibited in the near future.[x] Although there is general agreement that situations requiring reference to some body of accumulated knowledge do exist in modern society, there is no such consensus as to the most effective means of satisfying such needs under the varying sets of circumstances in which they arise.[2,3,4,5,6] One of the alternatives under consideration is the application of automatic information processing equipment to the mechanization of reference providing systems.[7] This report characterizes the basic functions required in such systems, and develops optimization techniques applicable to a certain class of implementations of these functions. In addition the basis for the evaluation of retrieval system performance is examined and some novel evaluation criteria are introduced.

In general terms the document retrieval problem can be introduced with the following assumptions: a body of recorded knowledge

[x] The references cited in the introduction are only illustrative and are by no means exhaustive. The Proceedings of the 1958 International Conference on Scientific Information[1] contains a number of papers dealing with document retrieval and allied problems.

exists in the form of a collection of documents (where a document connotates any segment of natural language text); a population of users exists with reason to believe that the collection may contain information pertinent to its needs. The problem, therefore, is in determining if in fact there are relevant documents (i.e. documents with information content useful to users) in the collection and in obtaining those which may be found. It will be assumed that the determination of the existence of relevant documents implies their identification and that some unspecified means is available for obtaining tokens of such documents once identified. In this context it should be noted that the document retrieval problem is considered distinct from the data-providing or fact retrieval problem.[8] The information content of a document in the former is considered an atomic element of the system; and as such, a document or a set of documents (or unique referents there to) is provided in response to user demands. But in the data-providing or fact retrieval problem, specific items of information, e.g. facts, messages, statements, answers to questions, etc., are extracted from source material and provided in response to users' queries. Automatic data-providing systems raise a class of problems such as the mechanization of deductive and inductive inference which are not considered here.

2. A Functional Model

Any document retrieval system, automatic or manual, can be functionally characterized by three basic elements:

a) the representation of the information content of source documents, i.e. the indexing function;

b) the representation of the information needs of the users of the system, i.e. the search request formulation function;

c) the matching operation between search request representations and source document representations, i.e. the search or retrieval function.

In addition to this functional characterization, other elements of document retrieval system organization are important in an operational framework. Such characteristics as storage organization, input-output facilities, document acquisition policy, economic factors and others may be critical in an operational sense[9], but for the purposes of this report these will be considered primarily as secondary factors. In this sense, then, the main purpose here is to consider the logical and methodological aspects of the mechanization of document retrieval systems, and in so doing to ignore many of the operational factors which may be important in other contexts.

The true information content of a document or segment of natural language text might be defined as existing only in the mind of its author. The representation of this content in recorded form via the natural language can be considered as an attempt at communication. That in fact such communication is successful on the average might in part be measured in terms of human progress. In any case, the information content of a document is a theoretically tenuous concept.

In document retrieval, one does not necessarily desire to extract and represent the "content" of a document, but rather to characterize that content in a manner which can consistently lead to the recovery of its primary representation, namely the document (the representation of the natural language).

The first functional aspect of any document retrieval system, therefore, involves the means for representing or characterizing the information content of source documents. Traditionally, this is the process of subject indexing. Useful referents to documents in a retrieval system may be indicative of attributes other than information content. In particular, referents such as the author's name, publication date, journal or publisher identification, cited references, etc., can be useful in several contexts.[10] For the current purposes, however, those referents not directly indicative of information content will be ignored with the understanding that their practical usefulness to the retrieval process as a whole must be considered in special circumstances. Chapter 2 of this study considers the role of indexing in document retrieval systems. The index function is discussed in terms of its goals, as well as in terms of the linguistic aspects of its mechanization, and of the possibilities of optimization of automatic indexing techniques.

The second functional aspect of the retrieval system, that is the search request formulation, is primarily a user function. In the broad sense it is also a system function in that a retrieval system includes the user. In a narrower sense, however, when the system is designed to react with the user so as to insure that a given search request becomes an effective representation of the user's information needs in terms of the system's capabilities, the request formulation process must be considered a critical system function.[11,12] Chapter 3 develops this premise and considers various means for optimization of user search requests in terms of system parameters and the objectives of the retrieval process as a whole.

The third functional aspect of document retrieval relates to the nature of the matching criteria used to select source documents in response to a user's input query.[13,14] In Chapter 4 the influence of the structure of the information representations on the matching criteria is developed. Major emphasis is placed in this chapter on the relation of the matching function to document classification and searching, and in this context an automatic classification algorithm is developed. This algorithm is specifically designed to increase search efficiency and is shown to be applicable to a certain class of matching functions.

Chapter 5 considers several aspects of the general problem of the evaluation of document retrieval systems, particularly as they relate to the functional model. In addition to examining the statistical basis for evaluation parameters, some novel measures are derived which have several advantages over those in current use.

Some of the salient features of the SMART automatic document retrieval system are presented in an appendix. The SMART system is used both as a concrete model, and as a simulation device for the experiments reported in this thesis.

3. A Specific Model - The SMART System

The experimental results to be presented here in connection with the optimization and evaluation algorithms were obtained by simulation, assuming a specific model for a mechanized document retrieval system. This model is based on the SMART retrieval system developed at Harvard University under the direction of Professor Gerard Salton.[15] The primary features of those elements of the SMART system of interest here are briefly outlined, so that it may become possible to refer to them in succeeding chapters. A more thorough outline of the SMART system is given in Appendix A.

A. Property Vector Indexing

Index images of source documents in the experimental system are assumed to be property vector representations of document content. For present purposes it is sufficient to assume that the index image of a reference document is an n-dimensional vector in a property space in which the weight or magnitude of a given component (or attribute) reflects the degree to which that attribute characterizes the content of the source text. Specifically, the index images experimentally used were constructed by a thesaurus transformation of the input text. An attribute of the resulting index space corresponds to a thesaurus category (group of semantically related natural language terms), and attribute weight is derived from the frequency of occurrence of the category terms in the input text.

B. Request Processing

Search requests in the experimental system are introduced into the computer in the natural language with no format restrictions. The index language representations of the search requests are identical in structure with those of the reference documents and are derived by applying thesame transformation rules to the request text. It will be assumed, in general, that a search request is to be interpreted as a description of a single topic area, i.e. a request describing topics "A" and "B" will be assumed to be satisfied by documents dealing with A and B, or with A in relation to B, etc. A user interested in documents either about A or about B is, by assumption, required to submit two search requests. The implications of this assumption are discussed in more detail in Chapter 3.

C. Angular Distance Matching

A retrieval operation in the model system is performed by matching the index vector representation of the search request with the index vector representations of all reference documents. The range of the matching function is assumed to introduce at least a partial order on the reference collection. Since the length or absolute magnitude of an index vector under the assumed index transformation is a first order function of the length (number of words) of the text which it represents, rather than of the content thereof, a matching function is desired which is independent of the vector magnitudes involved. Under these circumstances it is natural to assume that the information carried by the index vector is contained in its angular position (i.e. its orientation in the property space). The matching function assumed, therefore, is the angular distance or a monotonic function of this distance between the search request vector and the source document vector, wherein decreasing distance is assumed to indicate increasing probability of relevance.

D. Terminology

In dealing with the foregoing model,[†] the following definitions are required:

1) Let $\mathcal{D} = \{D_1, D_2, \ldots, D_m\}$ represent the set of source documents in the natural language comprising the reference collection.

2) Let $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_k\}$ represent a set of sample search requests in the natural language comprising a test set of retrieval queries.

3) Let T represent the index transformation from the natural language to the index language. The index image of a document $D_i \in \mathcal{D}$ is $d_i = T(D_i)$, and the index image of a search request $Q_i \in \mathcal{Q}$ is $q_i = T(Q_i)$. Further let $D = \{\bar{d}_1, \bar{d}_2,$

[†] Where alternative models are considered, e.g. index images represented by sets rather than vectors, the required notation will be introduced following the framework defined here.

$\ldots,\bar{d}_m\}$ be the set of source document images and $Q=\{\bar{q}_1,\bar{q}_2,\ldots,\bar{q}_k\}$ be the set of query index images.

4) Let V represent the index language, an n-dimensional vector space defined as the range of the transformation T. An index image $\bar{d}$ may, therefore, be written:

$$\bar{d}=d_1\bar{v}_1 + d_2\bar{v}_2 + \ldots + d_n\bar{v}_n,$$

where the d_i, i = 1,n are the scalar weights assigned to the orthogonal unit vectors $\bar{v}_i$, i = 1,n which constitute the basis of the index language vector space.

5) Let M be the search request - source document matching function from which the retrieved ordering is derived; M is then a function from the couple $(\bar{q},\bar{d})$ $\bar{q} \in Q$, $\bar{d} \in D$ to the real line.

REFERENCES

1. Proceedings of the International Conference on Scientific Information, 2 Vols., Washington, D.C., 1958.
2. de Grolier, E., "Problems in Scientific Communication", IBM Journal of Research and Development, Vol. 2, No. 4, October 1958
3. Bar-Hillel, Y., "The Mechanization of Literature Searching", Symposium on the Mechanization of Thought Processes, National Physical Laboratories, Her Majesty's Stationary Office, 1959.
4. Mooers, C.N., "The Next Twenty Years in Information Retrieval, Proceedings of the Western Joint Computer Conference, March 1959
5. Kent, A., Textbook on Mechanized Information Retrieval, Interscience Publishers, New York, 1963
6. Oettinger, A.G., "A Bull's -eye View of Management Engineering Information Systems", Proceedings of the ACM National Meeting, 1964
7. Salton, G., "Progress in Automatic Information Retrieval", IEEE Spectrum, Vol. 8, No. 1, Jan. 1965
8. Simmons, R.F., "Answering English Questions by Computer: A Survey", Communications of the ACM, Vol. 8, No. 1, Jan. 1965
9. Becker, J. and Hayes, R. M., Information Storage and Retrieval: Tools, Elements, Theories, Wiley, New York, 1963
10. Garfield, E., "Citation Indexes for Science - A New Dimension in Documentation Through Association of Ideas", Science, Vol. 122, No. 3159, July 1955.

11. Rocchio, J. J. and Salton, G., "Information Search Optimization and Iterative Retrieval Techniques", AFIPS Conference Proceedings, Vol. 27, Part 1, Spartan Books, Washington, D.C. 1965

12. Curtice, R. M. and Rosenberg, V., "Optimizing Retrieval Results with Man - Machine Interaction", Center for the Information Sciences, Lehigh University, Bethlehem, Pa. 1965

13. Stiles, H. E., "The Association Factor in Information Retrieval", Journal of the ACM, Vol. 8 No. 2, April 1961

14. Le Schack, A. R., "A Note on Measures of Similarity", Report ISR-7 National Science Foundation, Harvard Computation Lab., Cambridge, Mass., June 1964

15. Salton, G. and Lesk, M. E., "The SMART Automatic Document Retrieval System - An Illustration", Communications of the ACM, Vol. 8, No. 6, June 1965