ON-LINE INFORMATION STORAGE AND RETRIEVAL

Prepared for the

AGARD Symposium on
"Storage and Retrieval of Information"
For Session 3 - The Evolution of Current Methodology

18-21 June 1968
Munich, Germany

Noah S. Prywes
The Moore School of Electrical Engineering
University of Pennsylvania

ON-LINE INFORMATION STORAGE AND RETRIEVAL

Summary of Paper for the
Storage and Retrieval of Information Symposium
for Session 3: The Evolution of Current Methodology

SUMMARY

The paper is addressed to those concerned with improving effectiveness of small or large libraries, or those considering the establishment of a new collection in a certain subject area. The large staffing and cost frequently discourages the setting of satisfactory services. The paper submits for consideration an avenue of using automatic aids to achieve effectiveness within the bounds of economic practicality.

A number of methodologies have been developed. This paper summarizes the respective methodologies and the state of the art, with suggestions of the advantages of immediate application.

Another objective of the paper is to point to the potential of truly satisfactory services to users considerably beyond present capabilities. Through on-line communication with computers the rapid memorizing and recall can be extended to vast information, normally confined to the shelves of a library or the drawers of filing cabinets. A human will be able to use terminals to recall information from huge repositories in an effective and convenient manner.

The methods and procedures employed in information storage and retrieval for a century, such as indexing, classification or, more recently, content analysis, have proved of lasting value and serve as a foundation for the newer systems. However, to cope in a practical manner with the mass of data, it is essential that these traditional approaches be modified in order that these functions be performed automatically with only guidance provided from humans.

For instance, the indexing of documents should be entirely performed by the computer. This however connotes an open-ended index-word vocabulary which is first semi-automatically processed to form a thesaurus. The next step would be the completely automated processing of a classification system, which also provides a scheme for placing of documents on shelves, in micro-form or in the memory of the computer.

Based on these storage methods, the interactive man-computer storage offers the best potential for achieving high retrieval effectiveness to the point where information storage and retrieval systems become really useful as an extension of human memory and recall.

LIST OF FIGURES

ACKNOWLEDGMENT

The work discussed herein was supported by Contract NOnr 551(40) from the Information Systems Branch, Office of Naval Research and Rome Air Development Center.

ON-LINE INFORMATION STORAGE AND RETRIEVAL

Prepared for the

AGARD Symposium on
"Storage and Retrieval of Information"
For Session 3 - The Evolution of Current Methodology

18-21 June 1968
Munich, Germany

Noah S. Prywes
The Moore School of Electrical Engineering
University of Pennsylvania

1. DISCUSSION OF THE PROBLEM

This paper deals with the interrelated issues of pre-processing of documents and the effectiveness of retrieval of documents. Justifiably, the storage and retrieval problem is currently of great concern. The effectiveness in retrieving documents is highly dependent on the amount of labor and processing invested in the storage of the documents. Namely, the retrieval is greatly facilitated by storage processing products such as catalogues or storage allocation schemes. These are used in retrieval in referencing catalogues or in following a convenient placing scheme while browsing through shelves of documents. In effect, the problems of storage and retrieval are a single problem. This paper reviews briefly the components of a total storage and retrieval system while referencing relevant developments.

The storage process described in this paper includes all the functions which take place in libraries and information centers from acquisition to the placing of the documents in the repository. This process, which includes indexing, cataloging and vocabulary maintenance, demands a great deal of time and expertise. In any one of the large libraries or information centers there are thousands of monographs and serials that are waiting to be

catalogued and indexed. These often lay unused because of the dearth of competent cataloguers and indexers, especially those expert in particular subjects and languages. The increased amount of material which is being circulated soon may require substantial increase in staff. Staff with this competence is extremely scarce; low salaries discourage young people from library work. For these reasons the storage process tends to constitute a serious bottleneck.

On the retrieval side, evaluation tests indicate that libraries and information centers operate at a low, almost inacceptable retrieval effectiveness. The library user requiring specific information is overwhelmed with information, much of which is irrelevant.

The mechanizing of procedures in an information center or a library does not need any more justification than the notion of mechanizing any other industrial, commercial, or service function. The premise of this paper is that automatic storage processing and on-line retrieval are competitive in effectiveness with manual procedures. The automatic procedures are not especially complex and they can be readily applied (see Sect. 3.2).

The automated storage processing discussed here includes the following steps. Citations and sometimes abstracts of incoming documents are first transcribed into machine readable form. Natural language processing of title and abstract results first in a concordance of stem words. The concordance may also provide information about the frequency of stem words. In a semi-automatic process, words may be omitted, added, or various relationships established between words to form an open-ended thesaurus. Then, based on this thesaurus, the incoming documents are automatically indexed.

Finally, an automatic process may be applied which generates a library classification system for the collection. Such a classification then represents a scheme for placing documents on shelves, in microforms, or in the computer, as appropriate.

The interactive man-computer retrieval process, which follows the storage process, offers the best potential for improving retrieval effectiveness to the point where information storage and retrieval systems become really useful. This interactive process has a number of aspects. An individual can communicate with a central computer through a remote terminal "on-line", i.e., where the terminal is continuously monitored by a central computer. The computer deletes, changes and analyzes the queries and retrieves information in "real-time", compatible with the normal working speed of a human. To fulfill these functions, the computer must have a storage capacity of billions of characters with fractional second access to any information.

In the interactive retrieval process the user may search thesauri, classification schedules, catalogues or the documents in a manner very similar to that employed traditionally in libraries; however, with far less effort and much greater speed. He may for instance reference documents by title, author, publisher, citation, subject, or browse through citations or abstracts of documents on a common subject, placed together in the memory of the computer.

Methods and procedures like those described in this paper, such as content analysis, concordance and thesaurus preparation and indexing, which require merely clerical procedures, have been proposed for centuries.(1) They have been opposed by those who believe that manual processing of the

document has a "quality" superior to algorithmic processing based on selection of words from the abstract or even from the title. The manual approach has a number of ancillary positions that are contested here. For instance, the manual approach also conveys the notion that the subject term vocabulary needs to be controlled, and that only highly competent persons in specific areas should exercise judgment in regard to adding terms; these positions are contradictory to the approach in this paper. The objective of the procedures described here is to do away with much of the vocabulary maintenance work currently prevalent, especially the notes and instructions directed to indexers and cataloguers which would not be required in an automated system.

2. STORAGE

2.1 The Input of Documents and Content of the Repositories

The repository includes a collection of document representations. Each document is an integral entity in the collection. It may be broken down as shown in Fig. 1. The examination and analysis of documents in the storage or retrieval processes is usually conducted in the order from top down of the information shown in Fig. 1. Generally as one proceeds downward in Fig. 1, greater depth and frequently greater volume of information is provided; however, less frequent access is required to the more voluminous parts.

The upper four boxes in Fig. 1 are said to contain association terms. These are words or terms such as title, author, subject, etc., which identify single or entire classes of documents. The association terms may convey information about various relationships among respective documents, such as having a subject heading or citation in common, or sequence of events indicated by dates of publications, etc.

The language analysis in the storage process could be based on: 1) entire text, 2) association terms and abstracts, or 3) association terms only. The cost of transcription into machine readable form decreases greatly as the amount transcribed is reduced; however, this also reduces retrieval effectiveness. There is an indication however that effectiveness of retrieval based on the subject of the document increases considerably (20-25%) when the transcription of the abstract is added to the transcription of the association terms.(2) Language analysis of full text does not seem to sufficiently improve the effectiveness of retrieval to warrant the considerably greater cost of transcription. Also, the content analysis of text requires more complex procedures, including syntactic or semantic analysis.

The document collection is only one part of the information in the repository as shown in Fig. 2. The other parts contain directories of the association terms, and stratification of these directories. The directories may be considered to be information about information.

The directories may be generated a posteriori from the documents themselves. Namely, the association terms shown in Fig. 1 may be extracted automatically from each document as it enters the collection. In this way the concordance of the terms for all the documents may be derived automatically. The aggregate of the various types of association terms then constitute an all-inclusive directory or concordance of the association terms. Further processing then establishes the higher level directories which contain assignment of terms to categories and a variety of relationships among the terms.

The generally prevalent approach to indexing and vocabulary maintenance is that of applying human judgment a priori. An example of this approach is the establishment of the Dewey Decimal Classification which has divided the

library collection into progressively more specific classes. Using this system, professional indexers in libraries assign subject headings (stated in terms of class numbers) from a controlled schedule to the documents as they enter the libraries. In time, such a classification system must be expanded and revised by the library community to recognize new areas not included in previous schedules. Re-examination and reclassifying of documents already in the collection is then necessary to assign the new subject headings to them.

Figure 2 illustrates the a priori and a posteriori approaches to generating the directories as opposites. (A variety of mixes of these two approaches are possible.)

Retrieval effectiveness tests indicate that a posteriori indexing performs as well as a priori indexing; and that the lack of term control in a posteriori indexing does not cause deterioration in performance.[2,3] This will be further discussed below in connection with evaluation of retrieval effectiveness.

2.2 Language Processing

The simplest language processing procedure is to analyze a text to recognize and generate stems of words encountered in the input material. This involves recognizing the suffixes of words. A suffix editing procedure for English is described by Stone, et al.[4] A similar procedure for French has been described by Gardin and his associates.[5] Similar procedures have been developed by numerous other investigators.[6,7] More sophisticated procedures including matching of stem words against a thesaurus and syntactic or semantic analysis of text may be employed in the automatic indexing and classification as discussed below.

Natural language processing and machine translation research are relevant, as many of the algorithms developed there are directly applicable to automatic indexing. However, the systems employing the more complex procedures are highly experimental and in many cases the research has not advanced beyond the theoretical considerations.

2.3 Concordance and Thesaurus Generation

Although completely automatic thesaurus generation procedures have been under development for some time, considerable experience has been accumulated with a semi-automatic approach.(8) Computer aids are provided, but human intellect is applied to the discrimination and grouping of words. The first step in this process is to use computer aids which accept the transcribed portions of the documents as an input and generate a concordance of stem words. This concordance includes title or abstract words in addition to the other association terms in Fig. 1. The computer aids also provide frequencies of occurrence for the words in the concordance.

The first step in deriving a thesaurus may be the elimination of the very high and very low frequency words.(8) Another step would be the indicating of "broader", "narrower" or "related" relationships between words. Especially important also is the recognition of synonyms. It is necessary to establish such relationships as the documents have been authored by many people at different times who use a variety of words to designate similar meanings. Categories may be constituted which contain various instances of word usage, each such word may be given in context. Another approach is to prepare separate thesauri for some specific subject areas, where appropriate relationships between words are established in the context of the subject areas.

The thesaurus generation process is similar to the vocabulary maintenance functions in conventional libraries. However, the on-line automated aids may provide suggestions in regard to words and categories which deserve the attention of the individual engaged in establishing relationships among words. For instance, frequencies of terms used in retrieval queries and index terms of relevant documents, which have been retrieved in response to these queries, may serve as a guide regarding association and relationships among terms. Various statistics about frequencies of co-occurrence of terms may be used to combine terms into phrases which will be used in their entirety as a single term in the indexing process. Finally, the automatic generation of a classification, described later, may provide further information about grouping and sub-groupings of terms and respective documents to form progressively more generic subject areas.

2.4 Automatic Indexing

A variety of automatic indexing approaches and systems have been described by Stevens.(9) The objective here is to review briefly the simplest procedures which have proven effective. In the most simple procedures, stem words derived from titles or abstracts are considered to be the index terms of the respective documents without reference to the thesaurus at all. This simple process has proven effective for retrieval in situations where a user is satisfied with retrieval of any one or few relevant documents. This method proves especially effective in an interactive mode of search where the user may guide the computer in search for relevant material.

Automatic indexing may however utilize far more sophisticated approaches. A perusal of the thesaurus for stem words derived from titles or abstracts may result in important indexing decisions. It would eliminate undesired terms, or assign documents to classes or categories. Still a more complex process may assign term phrases based on words co-occurrences or based on syntactic analysis.

2.5 Automatic Generation of a Classification System and Assignment of Location For Documents

The automatic generation of a classification system in fact groups citations of documents in cells in the memory of the computer, very much as the documents on a common subject are grouped on respective library shelves. The retrieval process then consists of a search of several shelf areas in a large library to find the documents relating to a subject on which information is demanded. A classification system, automatic or conventional, has then a dual purpose. It is a methodology for placing like documents together but it is also a retrieval methodology by which one may be guided to the group of "like" documents which deal with the area of his interest. Like conventional classifications, an automatic classification system may be used to put documents away, but only after the classification system itself is derived from the documents. Namely, it does not precede the documents, but follows them. The automatic classification process is a follow-up on the automatic subject indexing. It attempts to put together in a cell documents which have most index terms in common.

The scope of this paper does not permit a description of the process for automatically creating a classification system. Various methodologies have been used for this process. These consist of employing statistical techniques, [10,11,12] computing "distances" between documents, [12,13] and employing co-occurrence of index terms. [14,15] The latter approach is simplest in terms of the complexity of the process and amount of processing required. A collection composed of 4,000 documents with a vocabulary of 6,000 index terms has been processed to date. [15] Experiments are continuing at the University of Pennsylvania with collections of tens of thousands of documents.

It is important to note here that automatic classification may be used not only to complement a coordinate-indexing retrieval scheme, but it also constitutes an alternative to coordinate-indexing. If used in a coordinate indexing system, automatic classification methodology provides a storage arrangement and a directory which greatly speeds up the search and retrieval. [16] As an alternative to coordinate indexing, automatic classification and the arrangement of documents in cells allows the user to direct the computer in its search toward the area of interest. This is further described below in connection with interactive retrieval techniques.

3. RETRIEVAL

3.1 Retrieval by Association Terms

A basic property of an on-line retrieval system is a man-computer language which includes in its vocabulary all the association terms. (See Fig. 1) A simple search may be initiated by the user communicating to the system a description of desired information. To "describe" a single or a class of documents, it is necessary to supply in a query the association terms

as well as the relationships among the terms. The procedure consists of specifying the association terms of the desired documents and the requisite logical or arithmetic relationships among the terms or among other information elements within the document. It is important that a user at the terminal should be capable of expressing a query in terms most convenient for him. For that reason ample choice must be given to him to search by various types of association terms such as author, publication, title words, accession numbers, references, etc. In addition, he should be able to reference the various directories, such as the thesaurus or the automatic classification, to aid him in selection of terms. Similarly, he should be able to specify for instance, a generic term to include all the narrower terms which correspond to it. Finally, he should be able to examine the citations which are being retrieved by the system and respond by indicating their relevance to his subject of interest. In these interactions with the computer, the display formats of the computer responses are important to the facility with which a system may be used. These formats are arranged to minimize the user's labor in selecting terms or documents.

On-line retrieval systems may be divided into two classes. The systems which aid user formulation of queries and retrieve respective documents are referred to here as key word systems. The second type of systems provide automatic reformulation of the query based on indications from the user of satisfaction or dissatisfaction with the retrieved material. These may be called automatic reformulating systems. In fact, in this manner the user guides and directs the search of the computer system.

3.1.1 Key-word Retrieval Systems

An outstanding example of the key-word system is the BOLD system at System Development Corporation, developed by Borko.(17) BOLD utilizes on-line displays which both assist the user in acquiring a mastery of the system itself and in performing guided searches. No language analysis technique is used in BOLD and the indexing is entirely manual. The MULTILIST system at the University of Pennsylvania is another example of a key word retrieval capability based on list processing which facilitates split-second retrieval from large document collections. The MULTILIST system includes both manually indexed (artificial intelligence) and automatically indexed (Physics) collections.(18)

BOLD and MULTILIST are representative of typical current systems. With these systems retrieval is easier but the basic content of the query is not altered except at the insistance of the user. Namely, while formulation of the query is assisted by the system, there is no attempt at reformulation based on the results of previous searches.

3.1.2 Interactive Query Reformulating Systems

The procedure in retrieval with a reformulating system may be as follows. A user may desire to search the collection to obtain a bibliography on a certain subject. He would then submit a query to the system consisting of word-stem terms. These terms may be found in directories (Fig. 2). The system then will use the automatic library classification which has been generated to find the cell(s) which correspond to the largest number of terms in the query. (Alternately, weights may be associated with the terms and cells are selected which have documents indexed with the maximum total weight

of the terms.) The user may then consider a number of citations from the respective cell or cells, and he may indicate acceptance or rejection of certain citations as relevant or irrelevant respectively. The terms corresponding to the accepted or rejected documents will then be examined by the computer and the initial query may be reformulated. It will include additional terms derived from acceptable documents or it will omit some of the initial terms that are in the rejected documents. Based on the newly reformulated query, a search is repeated, new cells are found and their content is displayed to the user. This process may continue with the input from the user being primarily the approval or disapproval of retrieved material.

This approach has been experimented with in the SMART Project and the results have been evaluated to determine the effectiveness of this powerful strategy.(2) Experiments with this approach have been also conducted by Edwards.(8)

3.2 Evaluation of Retrieval Effectiveness

As has been amply illustrated, there are a great variety of thesaurus generation and automatic indexing strategies as well as of retrieval strategies. It is also quite apparent that the selection of a strategy is very critical to the cost and retrieval effectiveness of the system. An evaluation methodology has been developed to determine retrieval effectiveness of systems.(19) As has been already indicated, increased costs and labor in storage processing may result in improvement of retrieval effectiveness. However, the amount of cost measure as related to the improvement in retrieval effectiveness is very important. Also, for various retrieval applications different degrees of effectiveness in retrieval are required.

Although tests of retrieval effectiveness have often been seriously challenged on a variety of grounds, two measures of retrieval effectiveness appear to receive wide acceptability.(20) One of these measures - the recall ratio - is the ratio of the number of relevant documents retrieved to the total number of documents in a collection which are relevant to a search. The other measure, the precision ratio, is the ratio of relevant number of documents retrieved to the total number of documents retrieved in a search.

For a sequence of queries interactively executed in the search, a plot can be made of precision vs. recall.(21) It is important to point out here that only the conjunction of these two measures is meaningful as an indication of effectiveness of retrieval strategies. The most ideal conditions would be those corresponding to unity recall and unity precision. For instance, perfect recall can always be achieved by retrieving an entire collection; the precision however would be then extremely low. On the other hand, if the number of retrieved documents is very small, the precision might be unity, but the recall would be very low. This illustrates that the combination of recall and precision must be considered in the evaluation. A strategy is considered to be more effective if its plot of precision vs. recall is described by a curve closer to the ideal point of precision = 1 and recall = 1. Examination of literature(21) indicates that in this respect, that joint recall precision retrieval effectiveness improves as well chosen, more sophisticated language processing techniques are applied, or as the retrieval process is carried out on-line, interactively employing greater choice of association terms.

4. CONCLUSION

The cost and staffing that are demanded of a library that desires to offer effective retrieval services are currently very large. Many smaller libraries try to use cataloging and indexing material generated in large information centers but even utilization of such resources requires considerable staff and cost. These smaller libraries may be the real beneficiaries from a total on-line storage and retrieval facility as described in this paper. The state of the art indicates that such a system is feasible and economical to develop at this time.

REFERENCES

1. W. C. B. Sayers, A Manual of Classification For Librarians and Bibliographers, 2nd Edition, Grafton and Co., 1944.

2. G. Salton, Scientific Report No. ISR-11 and No. ISR-12, Information Storage and Retrieval, Dept. of Computer Science, Cornell University, June 1966 and June 1967 respectively.

3. C. W. Cleverdon, et al., Factors Determining the Performance of Indexing Systems, Vo.. 1, Design, Part 1, Text. ASLIB Cranfield Research Project, 1966. Also, C. W. Cleverdon, Report on Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, ASLIB Cranfield Project, October 1962.

4. Philip S. Stone, et al., The General Inquirer: A Computer Approach to Content Analysis, The MIT Press, 1966.

5. J. C. Gardin, Syntol, Vol. II, Rutgers State University, 1965.

6. G. Salton, Content Analysis, Paper given at Symposium on Content Analysis, University of Pennsylvania, Nov. 1967.

7. N. Sager, "A Syntactic Analyzer for Natural Language," Report on the String Analysis Programs, Dept. of Linguistics, University of Pennsylvania, March 1966, pp. 1-41.

8. J. S. Edwards, Adaptive Man-Machine Interaction in Information Retrieval, Ph.D. Dissertation, The Moore School of Electrical Engineering, University of Pennsylvania, December 1967.

9. Mary Elizabeth Stevens, Automatic Indexing: A State of the Art Report, National Bureau of Standards Monograph 91, 1965.

10. H. Borko, "Research in Automatic Generation of Classification Systems," Proceedings Spring Joint Computer Conference, 1964, pp. 529-535.

11. J. H. Williams, Jr., "A Discriminate Method for Automatically Classifying Documents," Proceedings Fall Joint Computer Conference, 1963.

12. Frank B. Baker, "Information Retrieval Based Upon Latent Class Analysis," Journal of the ACM, October 1962, Vol. 9, No. 4, pp. 512-521.

13. R. M. Needham, "Automatic Classification in Linguistics," Rand Corporation Report, December 1966, AD 644 961.

14. N. S. Prywes, "Browsing in an Automated Library Through Remote Consoles," Computer Augmentation of Human Reasoning, M.A. Sass and W.D. Wilkinson, Editors, Proc. of a Seminar, June 1964, Spartan, 1965, pp. 105-130.

15. D. Lefkovitz and T. Angell, "Experiments in Automatic Classification," Report No. 85-104-6, Computer Command and Control Company, 31 December 1966.

16. N. S. Prywes, "Structure and Organization for Very Large Data Bases," to be presented at Critical Factors in Data Management/1968, Problems and Solutions, Symposium, University of California, Los Angeles, California, March 20-22, 1968.

17. H. Borko, "Design of Information Systems and Services," Annual Review of Information Science and Technology, American Documentation Institute, Vol. 2, John Wiley and Sons, New York 1967.

18. A collection of Physics Articles prepared by a project at the Massachusetts Institute of Technology under the direction of M. Kessler. The experiments conducted with this collection are a subject of a Master's thesis, "Automatic Introduction of Information Into a Remote Access System: A Physics Library Catalog," by P. Gabrini at the Moore School of Electrical Engineering, Rept. No. 67-09, University of Pennsylvania, December 1966.

19. C. P. Bourne, Evaluation of Indexing Systems in Annual Review of Information Science and Technology, C.A. Cuadra, Ed., Wiley, 1966.

20. D. R. Swanson, "The Evidence Underlying the Cranfield Results," Library Quarterly, Vol. 35, 1965, pp. 1-20. Also, "On Indexing Depth and Retrieval Effectiveness," Proc. of the 2nd Congress on Information System Sciences, Joseph Spiegel and Donald Walker, Eds., Spartan and McMillan, 1965.

21. G. Salton and M. E. Lesk, "Computer Evaluation of Indexing and Text Processing," Journal of the ACM, January 1968, Vol. 15, No. 1, pp. 8-36.

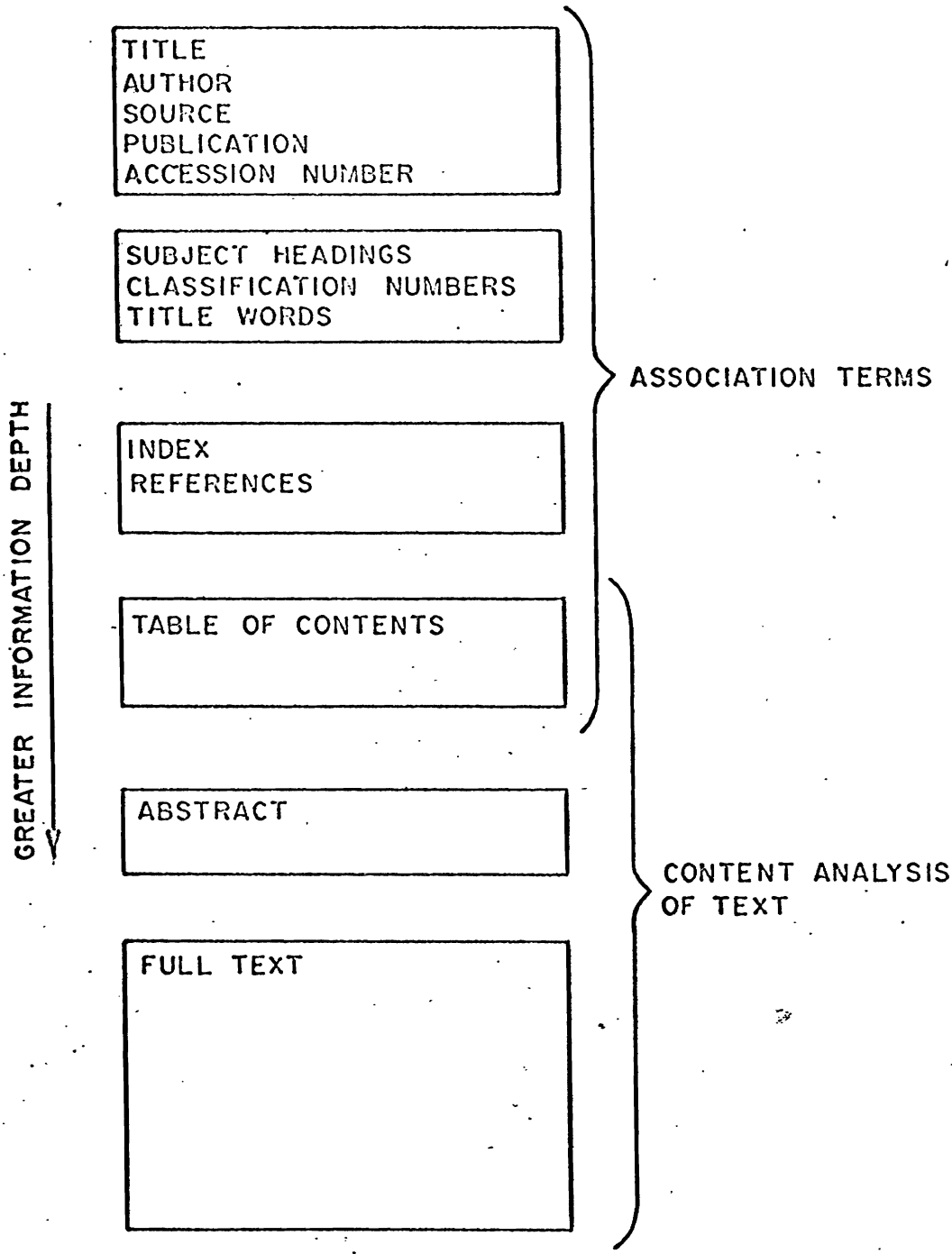

FIGURE I HIERARCHICAL BREAKDOWN OF A DOCUMENT

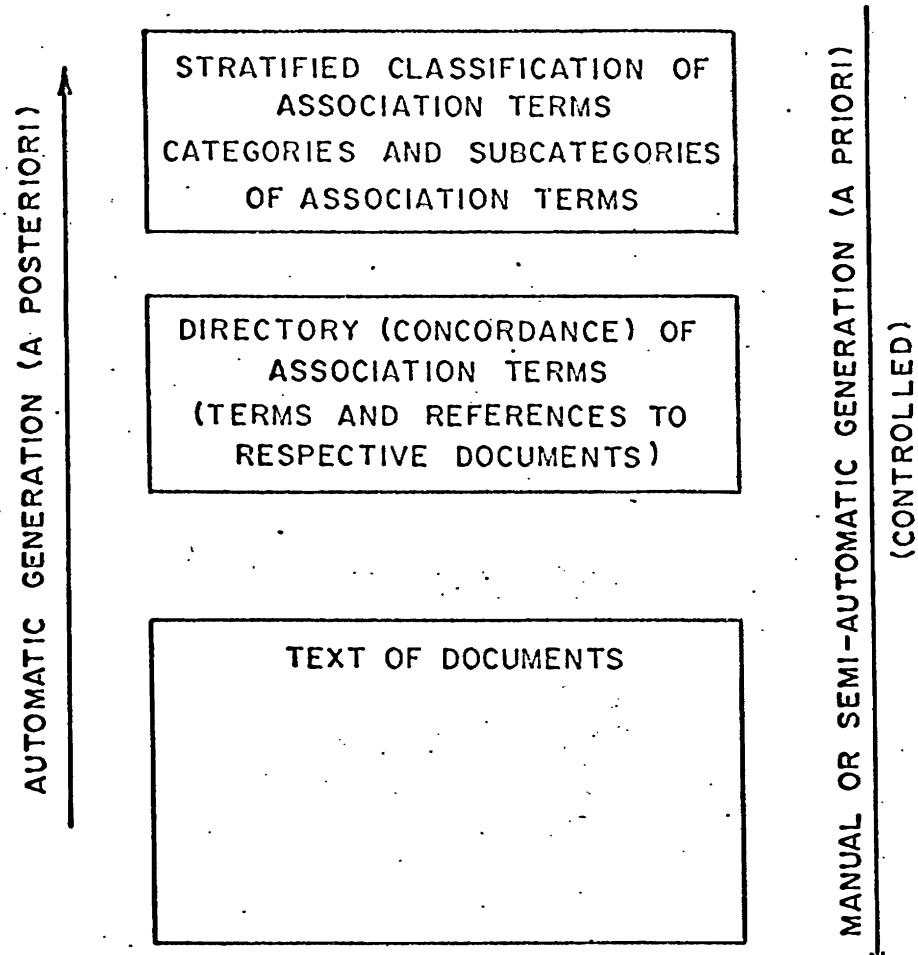

FIGURE 2 HIERARCHICAL BREAKDOWN OF INFORMATION IN A REPOSITORY