N00014-70-C-0297

Functions of a Man-Machine Interactive Information Retrieval System

Abstract

An effective man-machine interactive retrieval system is not achieved by simply placing a terminal on each end of an existing machine retrieval system. An interactive system requires a sequence of steps in which man and machine alternately take action. It should also provide different levels of services to experienced and inexperienced searches, recognize the difference between a narrow and broad query, furnish clues as to the next direction to be searched, reorganize the data base dynamically as the searcher changes his viewpoint, provide a ranking of responses in the most likely sequence and offer the searcher the option of overriding the ranking when a particular term is of extreme significance.

An online interactive system meeting many of these needs has been developed and tested. The objective of the development of this system, BROWSER, was to investigate the effectiveness of a free-form query with a combinatorial search algorithm and the effectiveness of various techniques and components to facilitate online browsing.

Functions of a Man-Machine Interactive Information Retrieval System*

J. H. Williams, Jr.
Federal Systems Division
International Business Machines Corporation
Gaithersburg, Maryland

o Introduction

Many machine retrieval systems have reached a plateau of effectiveness. A fervent hope exists that interactive systems providing the searcher with feedback will increase effectiveness in those systems having reached a plateau (1). The hope could be justified not only by providing empirical results (2) but also by noting that in many other physical and human operations feedback, or closed loop, systems offer superior performance to open loop systems. Increased performance is achieved by correcting the input on subsequent passes through the system on the basis of errors detected on previous passes. The notion of subsequent passes through the system implies several searches interspersed

---

*Support for development of the BROWSER system was provided by the International Patent Operations Department of the International Business Machines Corporation. Support for feasibility testing on naval documents and preparation of this article was provided under Contract N00014-70-C-0297 by U. S. Navy, Office of Naval Research

with corrections furnished by the human. Thus, interactive systems must be designed to furnish error information to the user and must be able to accept a modification to the input and continue to cycle until convergence on the desired output is reached.

This article describes the functions of the BROWSER online text retrieval system. The objective of the BROWSER development was to test the effectiveness of a free-form query with a combinatorial search algorithm and the effectiveness of various techniques and components to facilitate online browsing for concept retrieval.

Online concept retrieval systems have not progressed as rapidly as online data retrieval systems due to the several stages of uncertainty inherent in the creation, indexing, storing, and retrieving of concept information. In addition, at each stage ambiguity can exist in both directions from concept to term and from term to concept. A concept can be represented by more than one term, and a term can represent more than one concept. Further uncertainty arises with concepts that must be represented with more than a single term or a phrase. Concepts represented by a sentence or several sentences introduce additional uncertainty by allowing numerous combinations of words within sentences to represent the same concept.

Recognition of the distinctions between data and concept retrieval indicates the primary reason why online concept retrieval systems have not progressed as rapidly and why Boolean query systems may not be able to provide effective search performance for many subject retrieval application areas. Concept retrieval contains uncertainity and combinatorial

properties, whereas data retrieval is limited to logical properties. Boolean query statements admit only two conditions, true or false, whereas an algorithm needed for concept retrieval must admit a range of certainty or probability. In particular, a man-machine concept retrieval system should provide the user some degree of certainty associated with its output and should allow the user to interact in such a way as to increase this certainty.

This article describes: (1) some techniques that have reduced uncertainty in retrieval systems, (2) features to be included in man-machine concept retrieval systems, (3) functions of the BROWSER online retrieval system developed on the basis of a combinatorial search, and (4) the diverse needs of an information center.

o Concept Retrieval

The well recognized complexity of concept retrieval has been underestimated by those attempting to apply data retrieval techniques to concept applications. The confusion appears to be due to the emphasis on retrieval criteria rather than on relevance criteria. In data retrieval systems, retrieval and relevance criteria are identical, because they are measured in the same units. For example, in a driver's license file, the file content searched is license numbers, the query is expressed as a license number, the retrieval algorithm matches the units of the license number, and relevancy is based on license number units. Thus, all decisions concern only one dimension. All operations by machine and man can be performed with the same file.

In concept retrieval systems, however, retrieval and relevance criteria are not identical; they are each measured by different units. Retrieval criteria are expressed in term units. Whereas relevance criteria are expressed in conceptual units. As terms are physical units they can be stored and matched by machine. However, conceptual units are expressed and measured only internally within the human brain. Thus, a man-machine retrieval system should have two units, each with a different type of performance measure. The machine's units should be measured in terms of the physical words on which it operates, and the man's units should be measured in terms of the mental concepts on which he operates. The design problem becomes one of deciding how to assign the various processes of retrieval to the appropriate partner.

One reason that term retrieval systems have met with some success is that some concepts can be expressed with a single term or with adjacent terms. Performance, however, decreases as attempts are made to extend such systems to retrieve more complex concepts requiring expression by a sentence, or several sentences. Sentences provide so many degrees of freedom that "term adjacency" itself no longer identifies all relevant material. In fact, the number of degrees of freedom over several sentences causes the problem to be one of a combinatorial nature. Many combinations of terms can be used in several sentences to express the same concept.

Increasing need and desire for a concept retrieval system have stimulated developments along two alternative paths: (1) the research path of developing a method for representing concepts in a physical

form that can be stored in a machine and developing a comparison method to decide whether two concepts are identical or (2) the engineering path of constructing a device and testing it under environmental conditions. A successful concept system must be able to create a concept search file by extracting and representing concepts from source documents and recording them in such a way that they can be quantitatively compared. The system must then be able to extract and represent concepts from written or verbal query statements and provide a comparison algorithm upon which an identity decision can be made. An effective test for the adequacy of such an algorithm can be made without extensive computer processing. It could be made with a group of users, because the relevance decision is within the human province. Various descriptions of closely related but distinct concepts could be read and an agreement reached on which sets of words represent which concepts. Then the output of the concept extraction algorithm could be compared with the previous human decision. Cuadra's report (3) indicates the difficulties in obtaining human agreement on relevance. W. N. Locke (4) also confirms the difficulty by stating that "information is not something we read out of a document, but something that we read into it."

Observing the current difficulties pursuant to following the first path, we have chosen the second. We have developed and tested an online man-machine system for the retrieval of concept information. We have endeavored to employ feedback components that aid in reducing uncertainty, and to exploit the combinatorial behavior of words as they express concepts. A more complete system might exploit syntactic properties as well.

However, we have noted that verbal communications are often effective without completely correct syntax. Further, effective syntactic procedures at the present time only operate within the bounds of a sentence, while the need extends over a paragraph. Therefore, we developed a system utilizing only the combinatorial properties of words within a paragraph, viz. an abstract, to determine if effective results could be obtained without the additional refinements of syntax. Gifford and Baumanis (5) also found that choice of terms far outweighed arrangement of terms.

o Term Retrieval Techniques

Several techniques previously used in term retrieval systems should be included in concept retrieval systems when required. Recording the relative position of each word within a sentence and paragraph attempts to reduce the number of false hits; while augmenting index and/or query terms with thesaurus terms attempts to increase the number of true hits.

The word position technique has been employed successfully in many specific applications. These applications, however, should be more properly considered term retrieval rather than concept retrieval, since the purpose of the retrieval as well as the relevance criteria is in "term" units. When the task is recodification of statutes, the problem is to find all references to a particular term or phrase so that a new term or phrase may be substituted. When the task is retrieval of descriptors or keywords, the problem is to ensure that individual terms that may occur in several phrases are retrieved only within the desired phrase.

Successful use of this technique in limited applications does not provide sufficient evidence that it will meet with the same success on full text applications with complex concepts. In fact, when concepts can be expressed by various combinations of terms, the word position requirement overdamps the response so much that true hits are missed.

In the following example, portions of three abstracts containing the concepts "data management" and "information retrieval" are illustrated.

Query:

DATA MANAGEMENT and/or INFORMATION RETRIEVAL SYSTEM

Excerpts from relevant documents:

U. S. Army Technical INFORMATION Office...
Using the approach of concept storage and RETRIEVAL SYSTEM

AD-640101

U. S. Navy Library...DATA base for bibliographic searches and RETRIEVAL of the information...a universally accepted INFORMATION SYSTEM...

AD-640117

SYSTEM for handling and MANAGING Air Force DATA and INFORMATION...

AD-640811

Note that many of the query terms occur in each abstract, but not necessarily the same combination of terms nor the same sequence. If a term adjacency condition were specified on each pair of query terms, these abstracts would be missed.

The term "thesaurus" has been expanded into so many synonomous meanings that its various contents and functions need to be discussed separately. A thesaurus is a dictionary of synonyms. However, for the purpose of man-machine retrieval, there are several dictionaries of synonyms that can be stored and updated easily with minimum ambiguity, while others may not be amenable to economic computer processing. The easiest for both man and machine are those based on orthographic similarities and term lists having a one-to-one synonomous relationship. The most complex for both are those dictionaries showing hierarchical (narrower and broader term), precedence (use and used for), and miscellaneous (related term) relationships.

A term list that groups all inflected forms of a term under its root, can be easily created and maintained by a computer because it is based solely on character matching. A list of the root forms then represents the only search arguments for the file. A corresponding list of all inflected forms is not required because the computer compares roots to the fully inflected input; when a match is found, document number is stored in the search file under the appropriate root. The same root-match technique is used again when query terms are input. This performs a useful function of increasing true hits, because the searcher need not be concerned with the precise inflected forms employed by various authors

in their documents. Any inflected form input as a query term will match any form contained in the source document through the common root. A root word list also decreases storage requirements, as the dictionary length is reduced by one-third, and reduces search time because the term list is shorter.

A dictionary of abbreviations also increases true hits without the adverse effect of increasing false hits. An abbreviation list can be easily programmed, as the only requirement is table of address pointers referring to the term the abbreviation stands for.

In applications containing one-for-one synonyms, a related term list will also increase true hits without increasing false hits, provided all users of the system agree on the pairwise equivalency of the terms. Incorporation of this type of dictionary presents no problem to programmers as, again, the only programming technique involved is one of establishing address pointers to the corresponding terms. The critical responsibility lies with the user group responsible for ensuring that the terms are indeed in one-to-one correspondence with each other.

A dictionary of acronyms may appear to be similar to a dictionary of abbreviations. However, due to the proliferation of acronyms, there are now many letter combinations that stand for many different organizations and titles. Thus, the one-to-one correspondence has been lost. Only in applications containing limited subject material, where the acronyms that will be encountered are known, can the effect of false hits be limited. Otherwise, the problems of an acronym dictionary are the same as for the thesaurus, discussed next.

A thesaurus including relationships more complex than simple one-to-one substitutable terms contains the inherent potential of increasing the false hits. This disadvantage frequently outweighs the desired effect of increasing true hits. Searching a thesaurus for the term desired to represent a concept is fraught with the same problems as that of searching the data base. Thus, inclusion of a thesaurus within a system is more than an aid; it actually extends the overall search process into two distinct successive searches: first, a search of the thesaurus for the correct terms and, second, a search of the data base, using those terms, for the correct document. The same retrieval measures of relevance and recall should be applied at the end of the thesaurus search and at the end of the data base search. This two-stage procedure then allows one the opportunity of isolating errors due to the thesaurus, and due to the searcher's ability to express his mental concept of the problem in words.

If the thesaurus is used during indexing, the advantage of expanding an index term set is that the searcher's mental effort is reduced, as he does not have to recall related terms. The consequent disadvantage is that each document index set has been automatically expanded to such an extent that fine resolution between related subconcepts cannot be achieved. Another critical disadvantage is that terms are automatically added to an index set without regard to the content and concepts of the documents. A requisite of automatic index set expansion is that the user group agree on the term equivalences established by the thesaurus.

When the thesaurus function is employed at query time, the searcher has the advantage of maintaining conceptual control over terms (by adding only those terms appropriate to his concept and needs) at the expense of sacrificing time rather than performance. The total time required for the searcher to augment his query and review his responses may be less than the time of reviewing the larger set of responses retrieved from an automatically expanded query system.

The need for requiring a thesaurus within a system should be expressed in terms of the increase in true hits expected versus the expected increase in false hits. The placement of the thesaurus—within the machine system or exterior to it—as well as whether it is employed at index time, search time, or both, depends upon tradeoffs associated with the particular application.

Having established that a thesaurus will increase retrieval performance for a particular application, the decision of whether to include it within the machine system is one of cost, not performance. A well constructed thesaurus, such as the one published by the Engineers Joint Council (6) may be used both by indexer and searcher in book form exterior to the machine system. The cost of maintaining it within the machine and providing a search program to search it and display appropriate entries must be compared with the expected reduction of the human effort involved in looking up desired terms in a single well organized volume.

o Purpose of the Human in a Man-Machine System

Online systems are employed for various reasons. The overriding reason in concept retrieval systems is to allow the searcher to reduce the uncertainty in his final output by providing him means to iteratively converge on the desired response. An interactive system is not achieved by simply adding an input terminal onto a batch searching system. An interactive system requires a sequence of actions. The man and machine alternately take actions based upon information received during the preceding step. In particular, a closed loop feedback system furnishes error information indicating the type of correction to be applied to the input on the subsequent pass through the search system. The error information should indicate either that the current query input is too narrow, and insufficient true hits have resulted, or that the current query input is too broad and too many false hits have resulted. The error indication should not only concern magnitude, but should also give a hint as to the next direction in which to proceed. Thus, a designer must establish the types of information that will enable a searcher to take another action to cause convergence and at which phases in the overall search and retrieval process interactive points are needed. An interactive system has the potential of improving overall retrieval performance because it restores the searching portion of the overall search and retrieval process to the human. Searching is the human decision-making process of finding or discovering something through careful examination, whereas, retrieval is the mechanical process of bringing back identified information.

The design of man-machine retrieval systems must be based on a clear definition of those processes best performed by men and those best performed by machine. Man is less skilled at highly repetitive tasks, but performs well in his adaptive capability, for example, changing the characteristics of a problem statement as he reviews intermediate results. In addition, his learning capability increases his performance with time. Machines, excellent in highly repetitive tasks, but as yet poor in adaption and learning, should be used in complementary functions rather than competitive ones. Man thrives on freedom and abhors restraint; machines require order and control. In some systems, man is requested to force his thinking into the simple patterns required for orderly machine processing, with a consequent degradation in his performance. Human motivation and reaction are so closely tied to performance that man-machine systems must be designed to enhance man's performance while allowing the machine to serve him. A tendency has been developing, however, toward letting the man serve the machine, as shown in the proliferation of Boolean statement systems employed for concept retrieval.

A display terminal provides the searcher with the means of displaying information upon which he can make conceptual decisions. In this manner, complex arrangements of terms and text in its original form can be presented to the searcher for relevancy decisions; thus, the machine does not attempt conceptual decisions. Therefore, interactive systems should contain two levels of information one in the

form of terms upon which the machine performs character matches and a second in the form of combinations of terms and original text upon which the searcher makes relevancy decisions. Thus, the searcher need not repeat the machine decision on the same index terms, but can operate on a higher level of abstraction. Uncertainty is reduced by adding information. The function of the machine, operating on the term level, then becomes one of eliminating that portion of the data base containing none of the query terms and presenting to the searcher those documents containing at least some of the query terms. He can then apply meaning to terms and combinations of terms (in phrases and sentences) to make the final relevancy decision.

The underlying assumption is that a concept can be adequately expressed, and the target document located, if some of the query terms occur in the abstract. The assumption allows for even more stringent conditions than the classic synonym problem, as it does not require that the target document contain all of the query terms. The target document will be located whether the missing query terms are represented by synonomous terms, broader terms, narrower terms, or a few terms. The query expressing the desired concept is usually one or two sentences in length, with the choice of terms left to the searcher. Thus, with the selection of his terms, the searcher can intentionally make the query broad or narrow.

A browsing system should provide the searcher with the following capabilities, so that he can maintain control of the search process and dynamically adapt the system to his requirements:

1. Provide different levels of services to inexperienced and experienced searchers, particularly in areas of inter-disciplinary searching.
2. Recognize the difference between a narrow and a broad query with respect to the data base.
3. Furnish clues to relevant responses containing incomplete information.
4. Furnish clues to the next direction or subject area to be examined.
5. Organize and reorganize the data base dynamically as the searcher changes his point of view at the terminal.
6. Aid the searcher in his recall of synonomous terms.
7. Provide convenient means for adding, deleting, or modifying search terms.
8. Rank the responses numerically in the sequence most likely to fulfill the intent of the user.
9. Offer the searcher the opportunity to override a numerical ranking and emphasize the importance of a particular term.
10. Selectively print pertinent documents identified by the searcher.
11. Provide the option of performing Boolean or free-form queries.

o The BROWSER System

IBM has developed and tested an online interactive concept retrieval system, BROWSER (7). To "Browse" is to look over books, especially in order to select one to read. By implication then, the output of a browsing system will contain more documents than the precise answer set. In fact, if the computer were able to output the correct answer set, browsing would be unnecessary. The BROWSER system initially responds with a larger answer set as a result of the unrestricted query form and the searching algorithm. Through interactive cycles, it allows the searcher to converge to his desired output. The unrestricted query form allows the searcher to input one or more sentences describing the desired concept, a string of terms with or without correct syntax, a string of terms with or without function words, a phrase, or a single term.

Queries may be input from several sources. They may be input online or keypunched and loaded with the program. Standing queries may be stored and searched against periodic acquisitions. Because the input is free form, an existing abstract may be used in its original form as a query. Thus, the searcher is not obligated to paraphrase an author's own words nor determine which descriptors might have been assigned to the document. Any abstract in the file may be used as a query by setting the query number equal to the abstract number. Thus, the request, "find abstracts like this one," is met without need of interpretation. This enables specialists to communicate at the detailed level of their jargon without resorting to more general levels of all-encompassing terms. This facility

is particularly useful when browsing. When pertinent abstracts are found to indicate another direction that must be perused, the abstract just viewed may be immediately called from storage and employed as the next query.

Query terms which match terms in the search dictionary become search terms. Typically, the number of search terms ranges from 4 to 20 (current program maximum is 50). Terms in the search dictionary must have occurred in at least one document in the data base. The searching algorithm has the effect of computing all combinations of the n search terms r at a time. The theoretical number of combinations is extremely high; however, the number of combinations actually occurring in documents is reasonably low. Rather than compute all theoretical combinations, the system records only those combinations actually occurring in documents in the data base. For example, a free-form query on the subject of "Automatic Document Classification" contained 20 search terms. No document in the data base contained all 20 terms. In fact, one very relevant document contained only 7 of the 20 search terms. Each of 30 relevant documents contained a different set of search terms, with only one exception. Three documents describing the same system contained an identical set of search terms. Authors, however, do not always employ the same terms when describing their own work. The set of 30 relevant documents contained three pairs of documents written by three authors. In each pair the author had used a different set of search terms to describe his own work. Thus, if a query were constrained to an exact

match on a fixed set of query terms, an author could not even retrieve all of his own papers.

This missing term problem is overcome by introducing redundancy into the query, that is, including more search terms than the anticipated minimum. The algorithm then, by forming all combinations of search terms, retrieves documents containing combinations of search terms not anticipated by the searcher. Thus, the searcher need not know the subject area to the detailed level of knowledge of the author.

Boolean search systems provide effective retrieval to those who know the search terms being employed in their field and to those who have searched the area previously. The effectiveness decreases rapidly, however, as the unfamiliarity with a subject area increases (8). A system must offer varying levels of search services to personnel having varying levels of understanding of the subject field. In fact, many researchers have indicated that their greatest need for mechanized retrieval exists when they must leave their own areas and search in new areas or perform interdisciplinary searches. The BROWSER system answers this need by permitting searches with any terms the searcher may know and providing him with a display of the text of documents located. The searcher, having input very general or broad terms, reads and learns more about the area at the terminal. By observing the specific terms used by the authors in the field, he can decide which terms to add to his query. He can continue this cycle until he can express his problem in the terms of the new field and then request output of documents relevant to his query.

In addition to extending one's capabilities into new areas, a browsing system must also be able to aid the searcher in his own area by providing him with information about the data base. Too frequently searches are performed without knowledge of the scope and depth of subject matter in the computer's file. The terms "narrow" and "broad," when referring to queries, are becoming increasingly ambiguous. A query that appears narrow to one person appears broad to another. To a novice in the field, a narrow query broadens as his knowledge of the field increases. A query that is broad in one data base becomes narrow in a data base containing few documents on that topic.

The BROWSER system quantifies the notions of narrow and broad with respect to the documents of the search file. Weights are computed for each search term when the data base is loaded, and are kept current at each update cycle. The weight assigned to each term is inversely proportional to the number of documents indexed by that term: terms occurring in many documents have low weights and terms occurring in a few documents have high weights. After a query has been entered, the weights for each term are displayed. Examination of these weights indicates whether the query is "narrow" with respect to that data base. Searches consisting of terms existing in only a few documents are narrow with respect to the data base. When the display indicates search terms existing in many documents, the searcher can narrow the query (i.e., the potential output) by deleting the broad terms prior to executing the search.

The importance of weighting a term on the basis of its document frequency in a particular data base, rather than on general usage, cannot be overemphasized. The value of weighting terms lies in the functional utility of the weight for the machine search, not in an implication that a correspondence exists between frequency and relevance. The same term will thus have a different value in different data bases. The term "computer" in an IBM data base will have a low value, indicating its broad distribution throughout many documents. In a linguistic data base, however, it will have a higher value, proving useful as a narrowing term with respect to the entire field of linguistics. A searcher approaching a strange data base need not be concerned that his query will result in an output of hundreds of documents. A display will inform him of the characteristics of the file prior to the execution of the search. He may then limit the output to a few documents and view them on the screen to observe more detailed terms. On the next query he may add the detailed terms and delete the term that had a high frequency with respect to this data base.

The output of the machine search phase is ranked numerically. A sum is computed for each abstract based on the number of query terms occurring in the abstract and their corresponding weights. Abstracts containing detailed terms occurring in a small portion of the data base will be ranked higher than those only containing general terms occurring in a large portion of the data base. The approach is analogous to information-theoretic concepts stating that low-probability items contain

more information than high-probability items. The machine phase yields abstracts having the largest amount of information in terms of the search file vocabulary with respect to the combination of query terms input.

There have been many attempts to classify documents into categories and form clusters of index terms and documents during the index phase. As stated by Sharp (9), "The failure of these approaches is due to the virtual impossibility of providing all the points of view that searchers need to access relevant material." He elaborates on the magnitude of combinations necessary to represent all points of view. The BROWSER approach does not precluster documents nor determine a minimum number of standard combinations possible. It effectively generates a new document cluster with respect to each query. Thus, each query defines the center of a new cluster, and all abstracts are measured from that center. During the browsing phase, the searcher may change the shape and direction of the cluster by adding and deleting terms from the query.

For those situations where the searcher knows the required search terms and is unconcerned with synonyms, he may override the numerical ranking and cause the ranking to be a function of only those terms he deems significant, regardless of their frequency in the file.

Exhaustive searching was the overriding criterion in the design of the BROWSER system, as it was originally developed and tested for a patent-searching application. Therefore, the system was biased in favor of a high recall ratio. To compensate for the accompanying high relevance ratio, the system output includes an index of the search

responses to enable a searcher to scan the ranked output for assurance that all relevant documents were found. The index, similar to a Key Word In Context (KWIC) index, consists of a one-line summary for each abstract located. Each summary line shows which search terms occurred in that abstract. As each page contains 50 summary lines, which would normally contain only the full text of two abstracts, the information compaction ratio is 25:1.

Weighting terms based on document frequency overcame one of the problems confronting dictionary-preparation committees. The problem of: which terms should be placed on the common word exclusion list? After the 200 function words have been agreed upon for exclusion, the next set of high-frequency terms usually does not obtain unanimous approval. For most queries some of the terms will not be used. However, for state-o the-art surveys and novices entering the field, these terms provide valuable queries. Such terms can be included in the search dictionary, with the decision of whether to employ them in a query deferred to the searcher. By weighting terms in inverse proportion to their frequency in the file, the high-frequency terms will have relatively low weights. Their contribution to the total score of an abstract will be insignificant. Moreover, a searcher may exercise a decision to exclude them from his search by setting a threshold value. Any term having a weight lower than threshold will not be used by the search program. For state-of-the-art searches, the threshold may be set to zero and thus allow all terms to be employed in the search.

Aside from its information-selection properties, weighting value based on document frequency within a file also provides an aid to the physical problems of a storage and retrieval system. Search time is generally a function of data base size, but, more specifically, it is a function of the number of documents indexed by a query term. Providing a searcher with the number of documents indexed by each of his query terms enables him to affect the search time. The natural inclination to reduce time and cost will prevail as experience grows with browsing systems that permit control over the process. In addition, no degradation in retrieval performance results, as many searches have been performed with frequency terms suppressed (such as, information retrieval, programming, and electronic circuits).

Economical use of disk packs, particularly in online systems, also could result from the use of a weighting value based on document frequency. In large data base systems inverted search files may extend over several disk packs. Because retrieval of terms is on a random basis, the file sequence of the terms may be decided by the system designer. The order may be alphabetical, frequency, chronological, or by subject area. The advantage of a frequency order is that the high-frequency terms may be stored on a separate disk pack. In one application, the inverted file for the 25 highest frequency terms required as much space for document number entries as the next 2,000 search terms. Thus, the high-frequency term pack, not required by most searches, can be mounted on an as needed basis, rather than being online continuously.

In the BROWSER System indexing is performed, by computer, on the basis of words occurring in the abstract text. Accompanying descriptors may also be included in the search file. The terms in the search file are in root form. Thus, one form of a thesaurus, that of grouping inflected forms of a term together with its root, has been implemented within the current version of BROWSER. Other forms of thesauri can be implemented for specific applications based on need and agreement of a user thesaurus group on desired equivalencies. Another source of synonyms, the documents themselves, when displayed to searcher may be of more value than a collection of synonyms out of context. Due to the clustering algorithm, the initial search finds documents within the desired area. By scanning these documents on a screen, other terms used by authors in the field and within the context of the search are easily observed and added to the query at the searcher's discretion, rather than by a program.

o Diverse Needs of an Information Center

Many information centers have been using mechanized retrieval systems of one type or another. Most operating systems are based on keyword and descriptor or subject heading indexing and Boolean statement retrieval techniques. To determine the value of supplementing these techniques with free-form queries on the full abstract text, a feasibility exercise was performed on a data base furnished by the Navy Automated Research and Development Information System (NARDIS).

The exercise consisted of loading 1600 abstracts into the BROWSER system, updating the dictionary with naval terms, and performing searches representing diverse needs. The dictionary of electronic and engineering terms used in previous IBM internal tests was used on the first automatic indexing run. Of these dictionary terms, 2700 root words also occurred in NARDIS documents. Thus, the program and the existing dictionary saved considerable human effort as the basic dictionary did not have to be recompiled by Navy indexers. During the dictionary update cycle, 600 terms were added to the dictionary. The selection and review of terms to be added required only two days of effort. Some of the terms added were naval acronyms, such as AAW, ASR, ASW, ATP, IFF, NARDIS, NASL, NAVFAC, NAVMATINST, and so forth.

The set of queries created by NARDIS represented three explicit levels of concept information to be retrieved from a text file and thus proved useful in identifying the different retrieval techniques required to meet the diverse needs of an information center. The three levels of information were term, phrase, and sentence. In addition, the set also contained an example of subject category retrieval, numeric term retrieval, acronym term retrieval, and synonym enhancement.

Term retrieval consists of a single term search in which the user assumes responsibility for the ambiguity of the term; that is, any document containing the term is considered relevant. Terms can be words, acronyms, or alphanumeric character strings. Three queries were searched representing the various term types, respectively: "sonar," "3M," and "5000 angstroms."

Phrase retrieval consists of more than one term. A query containing the phrase "heavy list helicopters" represented this case.

Sentence (or sentence fragment) retrieval consists of a sentence or two or more phrases representing a concept. The concept itself is invariant, but various combinations of terms can be used to express the same concept. A concept query may appear to be similar to a phrase query because the terms used to express the concept are adjacent; however, the concept may also be expressed by a different phrase. The following four queries representing concepts were searched:

1. Prevention of acute respiratory disease in recruits.
2. Studies of fluid flow or viscous elastic flow or hemodynamics of blood.
3. Military applications of infrared detectors, detection, scanning, scanners, optical systems, and viewing devices.
4. Find reports concerning chemoelectric energy conversion, including batteries, fuel cells, energy storage and conversion (electrochemical only), galvanic cells, electrodes (electrochemical) and electrochemistry. Do not include subjects of corrosion, solar cells, power supplied (electronic). Avoid electrolytic cells and electrolysis if concerned with manufacturing chemicals (Cl_2, Mg, caustic, etc.), but include these subjects if concerned with reversible energy conversion of power generation phenomena.

The search for the concept "3M," which stands for management, material, and maintenance, indicated the value of the combinatorial

algorithm and the learning of new terms from displayed abstracts that further define the concept. No abstracts contained the acronym "3M," nor did any abstract contain all three terms. However, several abstracts containing combinations of two of the three terms were retrieved and deemed relevant. In fact, during the browsing phase of viewing the abstracts, additional terms such as "equipment" and "reliability" were observed and subsequently added to the query.

The search for studies of fluid flow relating to blood flow demonstrated the system's ability to relate documents from two previously unrelated disciplines. The researcher did not expect to find documents containing blood flow and the dynamics of fluid flow, as the purpose of his research was to relate them. He wished to find documents on fluid flow so that he himself could utilize previous research in another field to analytically attack the problem of blood flow.

The search concerning chemoelectric-energy conversion was handled more easily in this system than in a Boolean query statement system. The query was entered directly in sentence form rather than tediously attempting to state the concept in a series of "and," "or," and "not" logical statements. Emphasis in the BROWSER system is on stating the problem in sufficient detail to positively reduce the number of false hits containing general terms not on the specific topic. Therefore, the two negative sentences in this query were not input. Because the first sentence described the topic in sufficient detail, the number of false drops was below an acceptable minimum.

The diverse needs of an information center must be thoroughly explored with reference to the specific form and components of questions asked by the center's users prior to the selection of a particular machine retrieval system or technique. Retrieval systems processing a variety of questions cannot perform effectively on a small set of techniques. For simple term retrieval, the matching algorithm may need to include word-position information. For complex concept retrieval expressed in sentences or sentence fragments, a combinatorial algorithm will be required until analytic solutions of content analysis are fully developed.

o Summary

An automatic indexing and text retrieval system allowing free-form query and providing online browsing capabilities, through the IBM 2260 Display Terminal, has been developed and tested. The prototype system has operated on data bases of 25,000 German language patent abstracts, 9,000 English language patent abstracts, 8,000 Defense Documentation Center abstracts, and 1,600 Navy abstracts. The objective of the BROWSER System was to test the effectiveness of a free-form query statement with a combinatorial algorithm and the effectiveness of various techniques and components to facilitate online browsing. The techniques developed are valuable additions to existing retrieval techniques (such as Boolean statement queries and word position restrictions), particularly in retrieving abstracts containing incomplete information and retrieving from various viewpoints not anticipated by typical indexing procedures.

The essence of the retrieval problem is that some concepts are referred to by more than one term, and some terms refer to more than one concept. Thus, the multiple meaning problem causes both false hits and missing true hits. When a searcher uses a term having more than one meaning he will receive false drops. When a searcher attempts to find a concept that can be expressed by more than one term, he fails to retrieve it when he uses one term and the author used the other term.

This situation corresponds to Zipf's (10) hypothesis that ideal communication occurs when the vocabulary is in balance. The vocabulary is in balance when the effort expended by the author in selecting unambiguous terms and the effort expended by the reader in interpreting the terms is at a minimum. Deviations from the ideal occur in both directions: (1) if the author decreases his effort by selecting frequently used terms having multiple meanings, he increases the readers effort of selecting the proper meaning for each term; (2) if the author increases his effort by expending more time to recall more different words each having a unique meaning, he decreases the readers effort of interpretation because each term is now unambiguous.

In a retrieval system the document's words or index terms have been stored prior to search time. The responsibility for obtaining the vocabulary balance then lies solely with the searcher. Thus, interactive systems need to provide tools to the searcher for him to communicate with the author through the text terms or index terms. Information retrieval is similar to other human endeavors in which performance is a function

of effort expended. The performance of current retrieval systems may be indicative of Zipf's conclusion as expressed in the title of his book: Human Behavior and the Principle of Least Effort.

Improved retrieval performance will not necessarily result from interactive systems that simply provide a terminal on each end of an existing machine retrieval system. The entire search process must be reanalyzed so that appropriate functions can be properly assigned to man and machine. Each function offered to the man through a terminal must elicit responses that will encourage him to converge on the desired response. Indexing is a mental process of recognition whereas searching is a process of recall. Because recall is much more difficult, aids are needed for increasing the searchers performance. The best use of a display terminal in an interactive system is one that facilitates the transition from recall to recognition.

References

1. Swets, J. A., Effectiveness of Information Retrieval Methods, American Documentation, 20:72-89 (1969).

2. Salton, G., A Comparison Between Manual and Automatic Indexing Methods, American Documentation, 20:61-71 (1969).

3. Cuadra, Carlos A., R. V. Katter, et al, Experimental Studies of Relevance Judgments: Final Report Vol. I Project Summary, TM 3520/001/00, System Development Corporation, Santa Monica, 1967.

4. Locke, W. N., Computer Costs for Large Libraries, Datamation, 16:69-74 (1970).

5. Gifford, C. and G. J. Baumanis, On Understanding User Choices: Textual Correlates of Relevance Judgments, American Documentation, 20:21-26 (1969).

6. Engineers Joint Council, Thesaurus of Engineering and Scientific Terms, New York, 1967.

7. Williams, J. H., BROWSER, An Automatic Indexing On-Line Text Retrieval System, IBM Federal Systems Division, Gaithersburg, Maryland (AD693143), September 1969.

8. Verhoeff, J., Goffman, W., and Belzer, Jr., Inefficiency of the Use of Boolean Functions for Information Retrieval Systems, Communications of the Association for Computing Machinery, 4:557-599 (1961).

References (Continued)

9. Sharp, J. R., Some Fundamentals of Information Retrieval, London House and Maxwell, New York 1965, pp. 68-98.

10. Zipf, George K., Human Behavior and the Principle of Least Effort, Addison-Wesley Press, Cambridge, 1949.

ACKNOWLEDGMENTS

The author wishes to express appreciation to co-designer, Matthew P. Perriens; to the programming staff of Robert T. Fausey, James Lira, and Eugenia N. Gregory; and for the support provided by the sponsors.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Federal Systems Division International Business Machines Corporation Gaithersburg, Maryland 20760	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE
Functions of a Man-Machine Interactive Information Retrieval System

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Annual Progress Report in the form of a periodical article.

5. AUTHOR(S) *(First name, middle initial, last name)*
John H. Williams, Jr.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
October 1970	32	10

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-70-C-0297 b. PROJECT NO. c. d.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT
Article has been submitted for publication to American Documentation.
[redacted].

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Information Systems Branch Office of Naval Research Dept. of Navy, Arlington, Va. 22217

13. ABSTRACT

An effective man-machine interactive retrieval system is not achieved by simply placing a terminal on each end of an existing machine retrieval system. An interactive system requires a sequence of steps in which man and machine alternately take action. It should also provide different levels of services to experienced and inexperienced searches, recognize the difference between a narrow and broad query, furnish clues as to the next direction to be searched, reorganize the data base dynamically as the searcher changes his viewpoint, provide a ranking of responses in the most likely sequence and offer the searcher the option of overriding the ranking when a particular term is of extreme significance.

An online interactive system meeting many of these needs has been developed and tested. The objective of the development of this system, BROWSER, was to investigate the effectiveness of a free-form query with a combinatorial search algorithm and the effectiveness of various techniques and components to facilitate online browsing.

DD FORM 1 NOV 68 1473

Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT

Security Classification