Searching Natural Language Text by Computer

Machine indexing and text searching offer an approach to the basic problems of library automation.

Don R. Swanson

What is information retrieval?—"Hypothesis after hypothesis has been tried, and still we can't find out. Yet we must never lose hope; more and more complicated theories, fuller and fuller collections of data, richer rewards for researchers who make progress, more and more terrible punishments for those who fail—all this, pursued and accelerated to the very end of time, cannot, surely, fail to succeed."—C. S. LEWIS, *The Screwtape Letters.*

It is through the medium of libraries that knowledge is preserved for the future, yet posterity stands to benefit little unless such preserves are both accessible and digestible. Rather little is known in terms of scientific "observables" about the true usability of library-preserved information; the following brief description of the problem is intended to make evident some important aspects of this ignorance.

A scientist who nowadays imagines either that he is keeping up with his field or that he can later find in the library whatever may have escaped his notice when it was first written is a victim of what might be called the "fallacy of abundance." The fact that so much can be found on any subject creates an illusion that little remains hidden. Although library searches probably seem more often than not to be successful simply because a relatively satisfying amount of material is exhumed, such success may be illusory, since the requester cannot assess the quantity and value of relevant information which he fails to discover.

The sheer abundance of recorded knowledge and the growth rate thereof, furthermore, seem to foreshadow a crisis of inundation. It is estimated that there are now 100,000 journals and 600,000 scientists in the world, and that these quantities are doubling every 10 to 15 years. If this rate of increase is sustained until the tercentenary of this article, there will then be over seven times as many scientists as there are people, a circumstance which should at least provoke (if not stagger) the imagination. The implied *Warmetöd* of scientific information might be forestalled by engineering breakthroughs, but such breakthroughs may depend first on acquiring deeper understanding of the conceptual nature of the problem itself.

The question of formulating criteria for judging the relevance of any document to the motive, purpose, or intent which underlies a request for information is profound and lies at the heart of the matter. The hierology of research on "information retrieval" from libraries is well documented with reported "experiments" whose results generated little more than an aggregate of conflicting opinions on whether any given document, after recovery, was or was not really relevant to the intent of the requester, even though it fulfilled all specified "search criteria." The carrying out of such search experiments when "relevance" per se cannot be consistently assessed by human judgment would seem to represent overly vigorous pursuit of a solution before identification of the problem. Questions of determining "relevance" and "responsiveness," so far as practical purposes are concerned, receive dominant emphasis in the research reported here.

Consider for a moment a "thinkable" solution to the whole difficulty. Clearly, if the requester himself could read all existing literature (that is, every page of the entire library collection) and apply to each piece of information found therein expert intuitive criteria in order to ascertain its relevance to the requirement at hand, there would be, in principle, no problem of communicating recorded knowledge; the procedure suffers only from total impracticality. The historical answer to such impracticality is of course to organize, classify, index, and catalog the library and then search only a very small amount of catalog-type information. (Thus, it is an obvious though sometimes unappreciated fact that indexes and card catalogs for libraries need exist only because libraries are too large to be read through for each information requirement.) In short, some kind of condensed "representation" of the contents of the library is created. This representation, since it is condensed, is necessarily imperfect; the price paid for the convenience of having a practical and manageable search capability is the loss of at least some (and quite possibly most) of the accessible information in the library.

Now, it is a remarkable fact that the extent of information loss for existing libraries is not only unknown but apparently no serious efforts have ever been made to define such loss in measurable terms. Intuitively, one knows that the larger and more detailed the index and catalog, the less the information loss but the greater the difficulty of searching. Indeed, in the extreme, if each sentence of each article of the collection were cataloged, the "representation" of the library so created might approach high fidelity but would be almost as impractical as the library itself for conducting searches.

The matter of impracticality notwithstanding, to suppose that fully successful catalog search is a possibility implies that all human knowledge can be organized by means of some immutable and unambiguous classification or indexing scheme. Apart from the cardinal disadvantage of being nonexistent, such a hypothetical scheme has the even more unfortunate attribute of requiring that information be cataloged with precognition to insure recoverability in response to all conceivably relevant future requests. Both human knowledge and language change in the course of time, and the basic function of the library—to serve as a communication link be-

Dr. Swanson is manager of the synthetic intelligence department of the Ramo-Wooldridge Laboratories of Thompson Ramo Wooldridge, Inc., Canoga Park, Calif.

tween past and future—cannot, within the framework of present indexing and searching concepts, accommodate such change.

Even though a theoretically complete "solution" to the whole problem of communicating knowledge is thus almost unthinkable, progress in a practical sense constitutes an objective deserving of serious scientific attention.

Equivalence of Text Searching To Automatic Indexing

The purpose of this article is to present and discuss a fundamental approach to the investigation of automatic indexing and retrieval and to report the results of preliminary experiments on the searching of text. It is thus important first to make clear the relationship between text searching and automatic indexing.

Consider again the aforementioned suggestion that, ideally, the full text of the entire library ought to be read and searched for each request, to make certain that nothing relevant is overlooked. The task, for a human being is, of course, impractical to the point of absurdity. Such is not necessarily the case if a computing machine of sufficiently large capacity, high speed, and low cost were to perform the search. Let us suppose such a machine to be available, and for our present purposes let us neglect matters of speed and cost altogether, for the purpose of emphasizing the conceptual problems that are of paramount importance.

It may be seen, after some thought, that "direct retrieval" through searching the entire collection by computer to discover all information relevant to an inquiry is not necessarily possible in principle. That is to say, there is no known machine method for processing the full text of an unindexed document to determine whether that document is or is not relevant to some stated inquiry.

But now consider the full text of a library as the extreme case of a complete index or catalog—namely, a representation in which there is no loss of information. Clearly, if a computer technique for search and retrieval from the full text of a collection of documents cannot be developed, then it is unthinkable that matters could be improved by using the machine to operate on just part of the information (a "condensed representation")—that is, on an automatically produced index. This line of argument demonstrates persuasively that the development of techniques for automatic full-text search and retrieval is prerequisite to automatic indexing. It is equally clear that a technique for automatic indexing can be derived immediately from a text-searching technique, and thus that the two processes involve conceptually equivalent problems. For it is necessary only to so organize the machine procedures that those operations of text reduction or reorganization common to all searches are performed just once, and prior to searching, in order to create directly an automatic indexing procedure.

The practicality and economy of automatic indexing can then be investigated in the light of a known and controllable trade-off between information loss and brevity of representation. It is not suggested that present computers are necessarily economical for library text searching; this study has thus far been addressed to the conceptual problems alone. Questions of economics will be deferred until the process to be mechanized has been defined and understood.

A Model Information System

Let us take as a novel point of beginning the users of the library and their information requirements.

At first sight it would seem not altogether unreasonable to collect and study a large number of questions that have historically been asked of librarians and libraries. However, such questions necessarily reflect not the asker's "requirement" so much as his translation of that requirement into a form to which he thinks the library has a capability of responding. Now, perhaps one could determine the motive or purpose which prompted each request, but even then it is not unthinkable that these motives are to an unknown extent linked indissociably with the user's concept of the library's ability to serve his needs. The danger of beginning with the library query rather than the motivating requirement is apparent from the fact that the relevance of a library document to some request depends on the viewpoint and purposes of the requester.

Thus, in the experiments discussed here the "information requirement" of the requester is based on a set of questions constructed without reference to library experience. These questions are highly specific and, to the extent possible, incorporate within themselves the requester's "viewpoint" and "motive." Examples of such questions are: (i) What nuclear reactions are sensitive to the spin and parity of mesons and hence are useful in measuring those quantities? (ii) How does charge polarization within a nucleus affect the Coulomb scattering of charged particles by that nucleus? (iii) What are the "magic numbers" for nuclear shell structure?

A collection of 100 articles was chosen as an experimental "library," and each article in the entire collection was studied in the light of its possible relevance to each of 50 questions asked. Degrees of relevance were recognized, and a weighting factor was estimated to represent such a degree. The subject matter of the articles was restricted to nuclear physics, and relevance estimates were made by physicists with postdoctoral experience in the specialties represented by the articles. This procedure thus led, insofar as is conceivably possible, to prior determination of all relevant responses to any subsequent retrieval question. Confidence in the process is based upon the fact that a satisfactory degree of consistency among independently rendered judgments as to relevance resulted.

The restriction of the collection to homogeneous subject matter was intended to permit some degree of extrapolation to the behavior of larger collections. Computer search techniques must necessarily be based on the language content of the documents searched, and it is clear that the more homogeneous the language the more difficult the problem of discriminating relevant material from irrelevant for any given request. Thus homogeneous subject matter in a small collection would tend to present about the same level of retrieval difficulty as diverse material in a larger collection.

In effect, a "model" information system was adopted which consists of a small collection of articles, a list of questions, and a matrix of numbers representing the estimated degree or "weight" of relevance of each article to each question, based on careful application of expert human judgment. Within the framework of this model any method of information searching can be tested for its ability to yield relevant documents and reject those that are irrelevant. The scope of the investigation at present is based solely on the hypothesized model. Experiments and conclusions apply to the model only; the relationship of the model to reality is a separate matter involving a multitude of complex issues that will be dealt with in the course of time as understanding of the behavior of the model grows, and as reasons emerge for introducing modifications.

Experience thus far has indicated that this first model is neither too elementary to be challenging nor too complex to be manageable. In contrast to real information systems, the "observables" have been identified and the imponderables have been expunged—at least to an extent sufficient for practical purposes.

The latter point deserves some amplification. Information retrieval is not likely ever to become a precisely defined process (in a "formal" or mathematical sense) so far as machine procedures are concerned, and at best one can hope to achieve only partial success. The futility of seeking rigor and precision should not cause one to abandon pursuit of maximal understanding of the nature of the problem. It is this basic understanding which is sought in the investigation of a simple model, so that sound approaches toward the development of economic, effective, and practical large-scale information systems can then be initiated.

Text-Searching Experiments

To construct the model system described in the preceding section and to carry out the search experiments, two separate teams of physicists were formed. Group 1, through direct study of each article in the collection, determined its degree of relevance to each question asked. Group 2 was assigned the task of transforming each question into a search instruction for the computer. No member of group 2 was permitted any knowledge of the contents of the search library other than that it was made up of 100 nuclear physics articles selected from the *Physical Review* over a 10-year period.

Group 2 consisted of seven persons: three nuclear physicists, two computer programmers, one mathematician, and one librarian (with training in physics). Each of these seven persons formulated search instructions for each of 50 questions, using the three methods described in subsequent paragraphs. Thus, about 1000 retrieval experiments were carried out.

To illustrate the character of the search instructions, consider that into which the second of the three example questions listed above might be transformed: Deliver all articles which contain either the phrase *charge polarization* or *charge distribution* and at the same time contain the word *scattering, scattered,* or *scatter.*

The computer program utilized has a capability for testing each article in the collection to determine whether or not the conditions of the search are met. (The full text of each article was keypunched and recorded on magnetic tape.) In essence, an article can be tested to determine the presence of any given word, any one or more of a group of words, or the simultaneous presence of one (or more) members of each of several groups, as illustrated in the foregoing example. Phrases may be used instead of words in any part of the specification. For each article which satisfies the required conditions, the computer supplies a printed output which indicates the location of each of the discovered words or phrases within the search list.

The library (on magnetic tape) searched by the computer was not in any way indexed, classified, or organized by human processing. The experiments did not include searches for authors or reference citations.

In order to compare the effectiveness of text searching by computer with more or less conventional methods of using libraries, the experimental collection of articles was cataloged by means of a subject-heading index designed especially for the field of nuclear physics. The physicists of group 2, assisted by technical librarians, transformed each of the test questions into appropriate index entries in order to search the card catalog.

The success of machine retrieval and conventional subject-heading retrieval was measured in terms of the "direct examination" process carried out by the members of group 1. These measures of success take into account both the fraction of relevant material retrieved and the amount of accompanying irrelevant material.

Thus, the evaluation procedure is intended to establish some measure of the effectiveness with which human beings can transform questions into computer search instructions and to permit comparison of full text search with a conventional subject-heading catalog search. Of greater importance are the objectives of determining the extent to which the effectiveness of full text search can be improved through use by the requester of certain "retrieval aids," such as a thesaurus, and of developing techniques for systematically improving these aids as a result of collecting experimental data. Machine search experiments were carried out both with and without retrieval aids, and the two cases were compared.

The three methods of retrieval employed are designated by the following symbols and definitions. (*C*) "Conventional retrieval" based on a subject-heading index; no machine procedures are involved. (*U*) Retrieval based on specifications of words and phrases in disjunctive and conjunctive combinations, as illustrated above. No retrieval aids are employed (*U* denotes unaided). (*A*) Search requests are formulated as described for *U* but with the thesaurus-like word and phrase group list and the index thereto as retrieval aids.

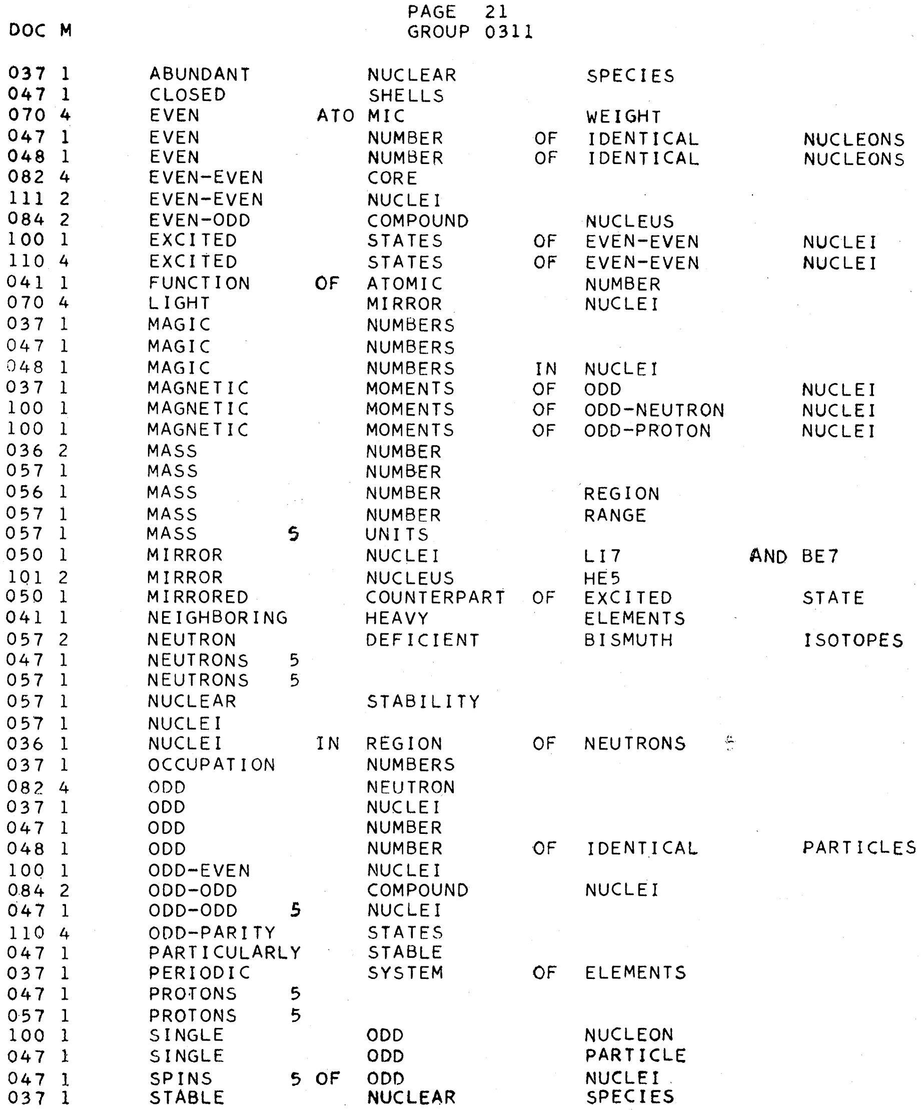
Each individual completed formulating search instructions for condition *U* before undertaking *A*, since otherwise *U* might be invalidated by recollection of the retrieval aids employed in *A*.

Use of a Thesaurus

The process of formulating a computer search instruction from the posed question consists of conjecturing lists of possible words or phrases which the requester believes responsive material ought to contain and nonresponsive material ought not to contain. He must be alert to different ways of expressing the same concepts and to the frequency or commonness of the words or phrases which he chooses. Potentially the retrieval aid of most use to the requester should be a collection of thesaurus-like groupings of words and phrases which tend to mean about the same thing for most foreseeable purposes of information retrieval and which can be used to stimulate his memory in the process of constructing "alternate ways of expressing the same concept."

Such a thesaurus for nuclear physics words and phrases was compiled for the purpose of these experiments. The use of this thesaurus can best be illustrated by an example. Suppose that in order to formulate a search instruction from the question "What are the magic numbers for nuclear shell structure?" the requester first conjectures that, although a search for the phrase *magic numbers* would no doubt be productive, there might be relevant articles which do not contain that phrase and, indeed, perhaps articles that were written before the phrase was invented. Thus, in order to be reminded of words or phrases in addition to *magic numbers* which might be clues to the presence of information relevant to the question, he consults the thesaurus by means of an index. He looks up the word *magic* and is referred to thesaurus group 0311 consisting of the list of words and phrases illustrated in Fig. 1. Among other things, he is reminded of the appropriateness of the concept of stability (*stable nuclear species, nuclear stability, particularly stable*) and somewhat less directly of the implication that an unusually high abundance connotes unusually high stability. On the basis of these reminders he designs an appropriate search instruction.

The thesaurus and other retrieval aids were compiled with the aid of text from a sample library that did not contain any of the search library documents, although, as in the latter, the subject matter was restricted to nuclear physics. It was hoped, of course, that the sample library would be useful as an aid to constructing the language relationships that form an essential part of the retrieval aids to be tested by using the search library. Had the search library itself been used to compile thesaurus-like word groups, then the particular relationships existing within the search library, and perhaps highly peculiar to that library, would become so readily apparent to the searcher that almost complete success in retrieval would be a foregone conclusion; the experiments in this circumstance would have had little scientific validity. This observation is particularly important in view of the small size of the search library.



```
                              PAGE  21
DOC M                         GROUP 0311

037 1     ABUNDANT         NUCLEAR          SPECIES
047 1     CLOSED           SHELLS
070 4     EVEN         ATO MIC              WEIGHT
047 1     EVEN             NUMBER       OF  IDENTICAL        NUCLEONS
048 1     EVEN             NUMBER       OF  IDENTICAL        NUCLEONS
082 4     EVEN-EVEN        CORE
111 2     EVEN-EVEN        NUCLEI
084 2     EVEN-ODD         COMPOUND         NUCLEUS
100 1     EXCITED          STATES       OF  EVEN-EVEN        NUCLEI
110 4     EXCITED          STATES       OF  EVEN-EVEN        NUCLEI
041 1     FUNCTION     OF  ATOMIC           NUMBER
070 4     LIGHT            MIRROR           NUCLEI
037 1     MAGIC            NUMBERS
047 1     MAGIC            NUMBERS
048 1     MAGIC            NUMBERS      IN  NUCLEI
037 1     MAGNETIC         MOMENTS      OF  ODD              NUCLEI
100 1     MAGNETIC         MOMENTS      OF  ODD-NEUTRON      NUCLEI
100 1     MAGNETIC         MOMENTS      OF  ODD-PROTON       NUCLEI
036 2     MASS             NUMBER
057 1     MASS             NUMBER
056 1     MASS             NUMBER           REGION
057 1     MASS             NUMBER           RANGE
057 1     MASS       5     UNITS
050 1     MIRROR           NUCLEI           LI7          AND BE7
101 2     MIRROR           NUCLEUS          HE5
050 1     MIRRORED         COUNTERPART  OF  EXCITED          STATE
041 1     NEIGHBORING      HEAVY            ELEMENTS
057 2     NEUTRON          DEFICIENT        BISMUTH          ISOTOPES
047 1     NEUTRONS   5
057 1     NEUTRONS   5
057 1     NUCLEAR          STABILITY
057 1     NUCLEI
036 1     NUCLEI       IN  REGION       OF  NEUTRONS  5
037 1     OCCUPATION       NUMBERS
082 4     ODD              NEUTRON
037 1     ODD              NUCLEI
047 1     ODD              NUMBER
048 1     ODD              NUMBER       OF  IDENTICAL        PARTICLES
100 1     ODD-EVEN         NUCLEI
084 2     ODD-ODD          COMPOUND         NUCLEI
047 1     ODD-ODD    5     NUCLEI
110 4     ODD-PARITY       STATES
047 1     PARTICULARLY     STABLE
037 1     PERIODIC         SYSTEM       OF  ELEMENTS
047 1     PROTONS    5
057 1     PROTONS    5
100 1     SINGLE           ODD              NUCLEON
047 1     SINGLE           ODD              PARTICLE
047 1     SPINS      5 OF  ODD              NUCLEI
037 1     STABLE           NUCLEAR          SPECIES
```

Fig. 1. A thesaurus group.

Results

Data on retrieval results are presented in terms of a scoring algorithm which depends, as mentioned above, on the percentage of relevant material retrieved and on the amount of irrelevant material also retrieved. The relative weight given to these two factors is arbitrary and represents an intuitive estimate of the relationship between "reward" for successful retrieval and "punishment" for irrelevant retrieval. The penalty for irrelevant retrieval we take proportional to the "cost" (in terms of time spent) of reading the irrelevant document. The retrieval score, then, is given by $(R - pI)$, where R is the sum of the relevance weights of the retrieved documents divided by the total sum of the relevance weights (for the given question) of all documents in the library; I is the effective amount of irrelevant material (and is given by $N - LR$, where N represents the total number of documents retrieved and L represents the total number of relevant documents in the library); and p is the irrelevance penalty and may take on arbitrarily assigned values. The scoring formula can be derived by taking a linear combination of the fraction of all relevant information retrieved and the cost of reading all retrieved material and then normalizing so that the score for "perfect" retrieval (all material which is relevant plus none that is irrelevant) is 1.

Results have been compounded into a single score; such a compound score facilitates comparisons for different individuals and comparison of different methods of retrieval. The sensitivity of the various scores (averaged over 50 questions) to the factor p is of especial interest, and a graph showing this relationship is presented in Fig. 2. The $p = 0$ intercept shows the percentage of relevant retrieval, while the slope is directly proportional to the amount of irrelevant retrieval.

The first conspicuous implication of the results is that the proportion of relevant information retrieved under any circumstances (intercept for $p = 0$) is rather low. Though wide variation of success within the 50 questions was encountered, for no single individual did the average amount of relevant material (that is, the summed weight factors) retrieved (taken over 50 questions) exceed 42 percent of that which was judged by the members of group 1 to be present in the library. It is clear that in a library consisting of only 100 documents a fairly heavy penalty should be assigned to irrelevant retrieval. Intuitively, it seems reasonable to assume that p ought to lie in the range 0.05 to 0.15; unless otherwise specified, comparison of data will be based on that assumed range.

A second implication of the data in Fig. 2 is the apparent superiority of machine-retrieval techniques over conventional retrieval within the framework, of course, of our model. Conventional retrieval was carried out under the favorable conditions of a highly detailed and specific subject-heading list, tailored to the sample library, which in turn was selected from the same journal (the *Physical Review*) as the search library (but which contained no documents in common with the latter). The assignment of subject headings to the documents was liberal, with an average of about five headings per (1000-word) document; the number ranged up to 14. However, the assignment was not based upon knowledge of the search questions themselves, nor was it intended to exhaustively identify all conceivable subject relevancies.

A still greater margin of success of text search over subject retrieval was provided by what may be called the "source" documents. Each of the 50 questions was inspired by some particular document within the model library; that "source" document was therefore known to be highly responsive (weight 10) to its corresponding question. Because of this direct relationship, all results for retrieval of source documents were separately tabulated and not included in the graphs of Fig. 2. These results, in terms of "percent of relevant material retrieved averaged over all requesters and all questions" can be summarized in the following manner: C [conventional (subject heading)], 38 percent; U [unaided machine text search], 68 percent; A [thesaurus-aided machine text search], 86 percent.

It is a reasonable surmise that the language of the "source" document may have influenced the language of the question, and hence the language of the search request, thus increasing the prospects for successful text search. Generally, though, it was found that all documents with high relevance weights (whether source documents or not) were proportionately more effectively retrieved by text search; such correlation between high relevance and successful retrieval was less pronounced for the conventional search. Details of these results are given in the reports cited in (*1*).

It is expected that the relative superiority of machine text searching to conventional retrieval will become greater with subsequent experimentation as retrieval aids for text searching are improved, whereas no clear procedure is in evidence which will guarantee improvement of the conventional system, other than purposeful indexing to exhaustion or perhaps the provision of a team of physicists as consultants to the librarian-indexer. For either circumstance it is hardly appropriate to use the designation "conventional."

Although much detailed study still must be carried out to determine, for each question asked, just why relevant information was missed, it seems likely that for the "conventional" case the problem is largely that of a mismatch in viewpoint between indexing and questioning. Subject headings relevant to the questions asked were largely nonexistent.

It is intended to test, in future experiments, the effectiveness of a subject-heading list developed directly from a set of "sample" retrieval questions, separate from but similar to the test questions. It is possible that tailoring an index to user plus library rather than to librarian plus library would result in better "conventional" retrieval, though again the term *conventional* seems to lose its appropriateness.

In general it was found that a thesaurus-like collection of word and phrase groups is clearly of potential use as a source of suggestions, reminders, and stimuli in transforming questions to search instructions. The effectiveness of the physics thesaurus compiled for the first series of experiments was marginal but measurable; of greater significance is its apparent improvability, as indicated by the analysis of the results.

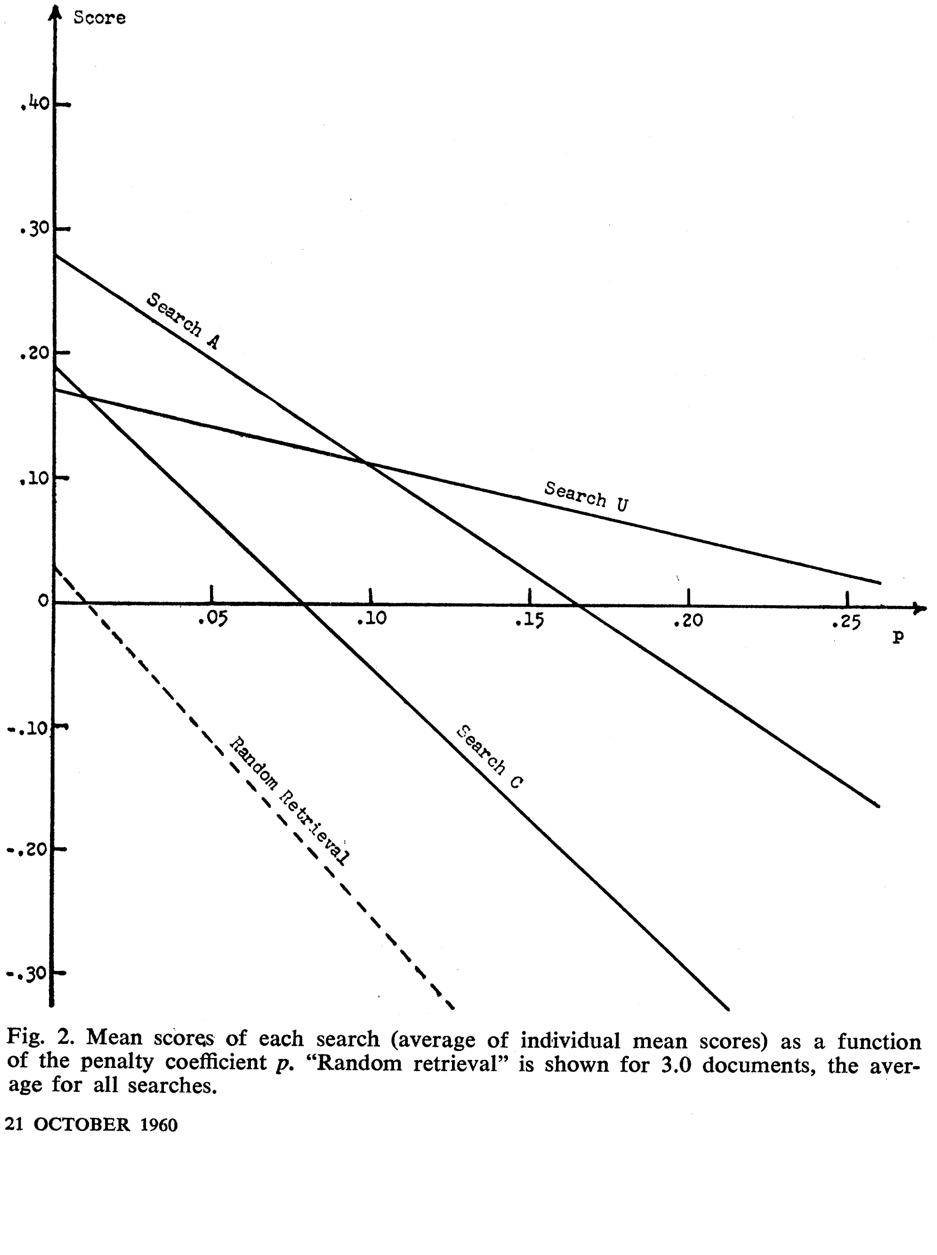


Fig. 2. Mean scores of each search (average of individual mean scores) as a function of the penalty coefficient p. "Random retrieval" is shown for 3.0 documents, the average for all searches.

A study is in progress to determine for each question asked the specific reason why relevant information was missed and why irrelevant information was retrieved. The results of partial analysis may be summarized as follows:

1) As was to have been expected for the first experimental cycle, a good many of the failures were due to obvious causes for which there are simple remedies. Inadequate recording of symbols and equations (particularly of subscripts and superscripts), confusion of capital, lower case, and other "engineering" problems accounted for failure to retrieve a significant amount of relevant material. Pure "machine synonyms" (such as π as an abbreviation for π-meson) not foreseen by the requester were encountered quite frequently.

2) A somewhat surprising reason for retrieving certain "irrelevant" information was the fact that such information turned out not to be irrelevant after all. Relevance implications had been overlooked by the members of group 1 in spite of independent assignment of weighting factors by two persons and a resolution of all variances through discussions with a third. The number of variances exceeding 2 weight points (on a scale of 10) was less than 20 percent of the total, for independent assignment, and most of this variance was attributable to obvious oversight. It is inferred that the exposure by machine search of the occurrence of certain words, phrases, and combinations of words and phrases can suggest to the reader relevance relationships that were not apparent beforehand, even with reasonably diligent study. Although this phenomenon was not very frequent, it is clear that the group 1 diligence level should be increased for subsequent experiments, since presumably a small amount of nonretrieved material erroneously received a relevance score of zero.

3) In many cases irrelevant material was retrieved through the conjoint occurrence of words which were not related to one another in the manner intended by the requester. This phenomenon is, of course, well known in any process of coordinating key words. To some degree, however, whenever irrelevant retrieval occurs because of lack of relationship specification, it should be possible to remedy the situation by requiring that certain of the specified terms occur within the same sentence or at least in relatively close proximity to each other. In effect, the substitution of proximity for syntax is suggested; the potential value of this suggestion is being investigated. Preliminary results indicate that, on an indiscriminate or purely statistical basis, proximity is not helpful. All retrieved documents were partitioned into two groups, the relevant and the irrelevant. A search word-proximity parameter was defined, measured, and found to be not significantly different for the two groups.

4) The most difficult problem encountered is that of indirect implication or even pure analogy. On a number of occasions an article which would unequivocally be judged by a physicist to be relevant to a given question would contain no words which, in any reasonably direct way, would be at all likely to be designated relevant on the basis of the question alone. It turned out, however, that conventional indexing and retrieval systems in these latter cases were no more successful in recognizing the relevant associations involved.

5) Irrelevant material was retrieved due to unforeseen contexts of single words. Expansion of the thesaurus phrase index would probably remedy the situation to a large degree by providing sample contexts for frequently used combinations.

6) A rather large amount of missed relevant information would have been retrieved if certain fairly common near-synonyms of the specified search words had been listed as disjunctive alternates. It is obvious that augmentation of the thesaurus would remedy the defect to a degree, but it cannot yet be surmised how far one can go toward building into a thesaurus all conceivable associations of retrieval interest.

Summary

A fundamental approach to automatic indexing and retrieval of library-stored information through investigating machine search of natural language text is described above, and the results of preliminary experimental studies based on that approach (*1*) are presented. A limited and manageable model of a library (together with search questions) constituted the object of the investigation; the effectiveness with which responsive information could be recovered was measured. The small scale of the model permitted direct examination of the entire collection as a basis for establishing practical measures of "relevance" or "responsiveness" to questions.

In terms of these measures, the effectiveness of all information search techniques tested on the model was found to be rather low. Text search by computer was, however, significantly better than a conventional, nonmechanized subject-index method. Thus, even though machines may never enjoy more than partial success in library indexing, a small suspicion might justifiably be entertained that people are even less promising.

Note

[1] This work was carried out under the sponsorship of the Council on Library Resources, Washington, D.C. I acknowledge with sincere appreciation the assistance of L. C. Ray, who planned and supervised the computer programming efforts; of J. L. Kuhns, who made invaluable contributions in mathematical analysis and in interpretation of the results and also suggested the C. S. Lewis quotation; of J. M. Fontana, who assisted in the programming; and of D. V. Black, who developed and applied the subject-heading index. I extend thanks also to the members of group 1 (R. W. Wright, A. Risley, J. M. Brabant, L. M. Frantz, A. L. Morse, W. Blocker, L. Higgins, W. Hershberger, and H. Trueblood) and to members of group 2 (D. L. Drukey, R. Norton, K. Watson, J. Mersel, D. Wenger, J. L. Kuhns, and D. Black). I am also indebted to Professor N. Chomsky and Professor P. L. Garvin for helpful and stimulating discussions. For further details and additional data on the experiments discussed see "Word Correlation and Automatic Indexing," *Ramo-Wooldridge Reports; Progr. Rept. No. 1, C82-9U9* (1959); *Progr. Rept. No. 2, C82-0U1* (1959); "An Experiment in Automatic Text Searching," *Phase I Final Rept. C82, 0U4* (1960).