Title

A PRELIMINARY REPORT ON AN ON-LINE INDEXING- EDITING SYSTEM FOR DOE/TECHNICAL INFORMATION CENTER

Authors

Cerny, Barbara A.
Lawrence, J. Dennis.

Publication Date

1981-11-01

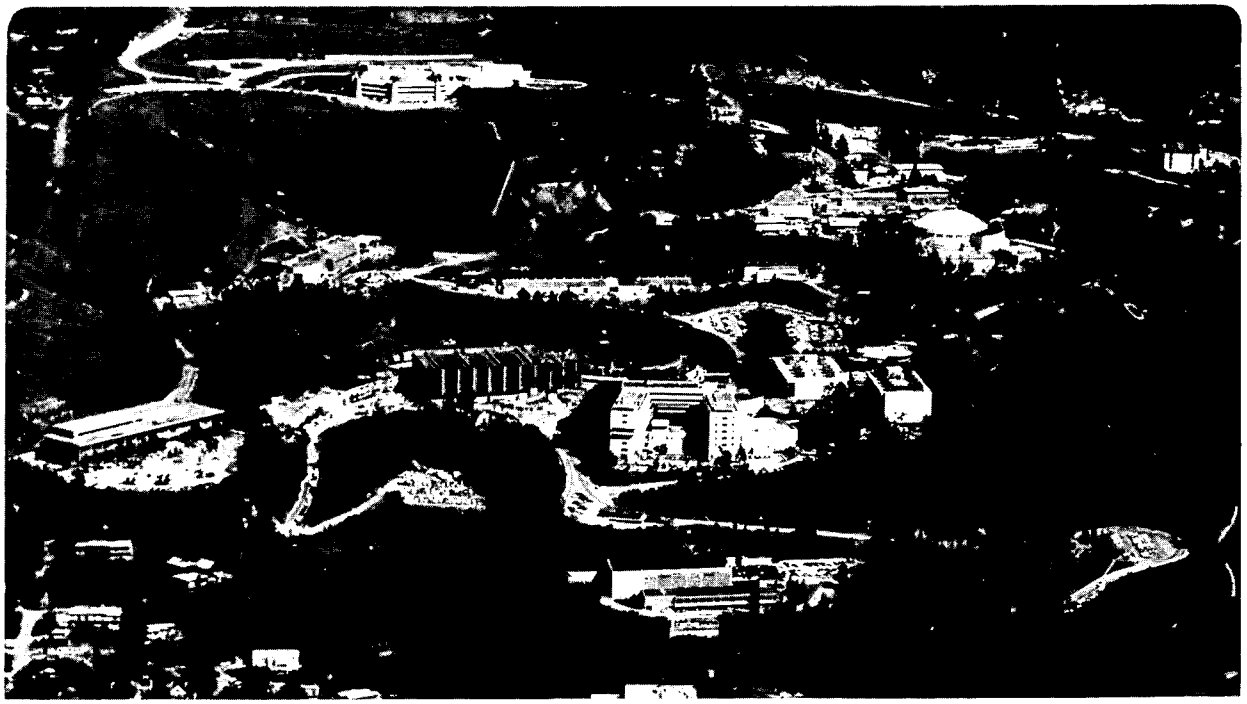
Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Engineering & Technical
Services Division

A PRELIMINARY REPORT ON AN ON-LINE INDEXING-EDITING SYSTEM FOR DOE/TECHNICAL INFORMATION CENTER

Barbara A. Cerny and J. Dennis Lawrence

November 1981



Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

DISCLAIMER

LBID - 446

A PRELIMINARY REPORT ON AN ON-LINE INDEXING-EDITING SYSTEM FOR DOE/TECHNICAL INFORMATION CENTER

by

Barbara A. Cerny

and

J. Dennis Lawrence

Information Methodology Research Project
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

November 6, 1981

This work was supported by the U.S. Department of Energy under Contract W-7405-ENG-48

INTRODUCTION

This report will discuss the motivation for and design of a system to provide computer assisted aids to document indexing for the DOE/Technical Information Center (TIC) Energy Information Data Base (EDB) [1]. This computer searchable data base covers all scientific and technical energy areas. It includes more than 1,000,000 citations and 15,000 items a month are added to it.

Each citation consists of bibliographic data elements to identify the item (such as title, author, author affiliation journal name, publication date, etc.), while the essence of the item is given by the abstract. These abstracts may be written by the author and edited, if necessary, by TIC, or may be prepared completely by TIC if an author generated abstract is unavailable. Additionally, the documents are indexed by specialists to reflect the subject content using descriptive terms from the EDB thesaurus [2].

Large numbers of documents are currently processed daily with a combination of manual indexing and batch computer processing (Fig. 1). Approximately 65% of the citations come in on magnetic tape from the American Institute of Physics, Engineering Index, INIS, and other sources, already indexed to varying degrees. The remaining 35% are printed documents. This report addresses the problem of reducing the backlog in this processing and offers as a solution a computer system designed to take advantage of state-of-the-art interactive programming techniques which allow an indexer to interact directly with a computer terminal. The set of programs that we are prototyping, while drawing on automatic indexing techniques, relies strongly on human decision-making, by including the indexer as a fundamental element of the system. Since the Energy Data Base is a very large, dynamic collection of citations, the implementation of completely automatic indexing for it

would have involved work of the magnitude and scope of Klingbiel [3] for the Department of Defense. And even in this successful, functioning system, approximately 25% of the documents must still be reviewed manually. Hence a pragmatic approach to increasing processing efficiency seemed to be to side-step the manual indexing vs automatic indexing debate and initiallly emulate the manual mode of indexer behavior. The decision making aspects of the indexers task would be enhanced while many of the clerical tasks would be minimized or eliminated. A future goal is to build "intelligence" into this user interface. This "intelligence" will be our point of contact with previous work on machine aided indexing (MAI) since it will be based on statistical and linguistic analyses of document text. We will then go beyond MAI to artificial intelligence techniques to further enhance the system. Our initial system, then, will be a "front-end," a user interface, and a series of displays that will allow an indexer easier, faster access to the tools necessary for indexing.

This system is portable, since it is being written in RATFOR, a high-level language that translates into Fortran. It should be possible to interface it, as a front end, with existing MAI systems, to use the outcome of MAI schemes as the "intelligence" for the displays, or to access another thesaurus in place of the EDB thesaurus. When one begins to think of document indexing, as well as document retrieval, as an interactive process, then the information locked within the data-base can be assessed and presented upon demand to the indexer as it now is to the user, but in a chameleon-like variety of forms to trigger thought processes in the indexer. The indexer becomes, in essence, a "user," also, of the system.

TIC's Indexing System

For TIC's system, the sequence of indexing steps is shown in Fig. 1. Following across the page, the processes depicted include:

1) A document surrogate (title, abstract, author, etc.) is received on tape already indexed to some degree, or as a physical document.

2) The tape is run through a batch process that checks keyword assignment against the on-line EDB thesaurus as well as the accuracy of selected fields.

3) The printout (or the physical document) is routed to subject specialists reponsible for the content of the document or of document surrogates. They make handwritten corrections, enter keywords, generate an abstract, etc.

4) The corrections or additions are sent to a descriptive cataloguer who enters them into a batch programming procedure.

5) The files so generated are run again through the check program.

6) The rejected records are sent back to the indexer.

7) Steps 4) to 6) are repeated until the tape is correctly indexed.

8) The final EDB tape is entered into RECON.

We propose to modify the procedure as shown in Fig. 2. The main features will be:

1) To allow the indexer direct access to the original tape instead of printing out the tape after the checking procedure. The checks will be performed as part of the indexing process flow.

2) The batch preprocessing of the document surrogate on tape will place individual records in files for the subject specialists so each can access surrogates in his area of interest. As part of this process category and keyword predictions will be attached to each record.

3) The indexing programs will possess the following features:

a) The use of a screen editor which can be called from a menu-driven program. An indexer can chose to edit a title or abstract, adding keywords or categories in any order.

b) The supplanting of manual thesaurus lookup with the display of any thesaurus term on the CRT screen. Additionally, information from a permuted index of the thesaurus will be available.

c) The option of entering keywords with a touch of a light pen to a screen or by keyboarding terms. These functions will replace the descriptive cataloguer entering the indexer's handwritten keyword lists.

4) Any additional fields will be added by a descriptive cataloguer.

5) A RECON tape will be generated, as a final product, as is currently being done.

METHODOLOGY

The design of this system encompasses techniques from computer science (CS), information retrieval (IR), computational linguistics (CL) and artificial intelligence (AI). Separating these overlapping disciplines in this way allows a conceptualization of the man-machine boundary that permits system development simultaneously on several fronts commensurate with the project goals. That is, an interactive indexing-editing system can be built in parallel with research on how to build "intelligence" into it. In the broadest view, the ultimate goal for any IR system has to be the effective retrieval of document citations for the end user, and the "intelligent" features under consideration will lead towards more precise indexing, better communication between indexer and user, and ultimately more effective

retrieval. In the short term, however, there is the practical consideration of eliminating a backlog of work in a cost effective manner.

From a theoretical perspective, Smith [4] applies O'Connell's [5] three stages of evolution of a technology to the development of a retrieval system; this breakdown is equally applicable to the development of computer assisted aids to indexing:

"The first stage in the evolution of technologies is one in which what is being done now can be done cheaper, faster and better with the help of technology than without.

The second stage occurs when we can do things to match the new capability that the new technology gives us.

The third stage occurs when we change our behavior and our ways of doing things to match the new capability that the new technology gives us."

This breakdown also roughly parallels our use of tools from CS, IR, CL and AI. Thus far, we have worked primarily in the first stage with standard computer science techniques. We are emulating current indexing behavior with a menu-driven system of displays to be used in a browsing mode by the indexer. He can explore text, thesaurus and permuted index displays via terminal and make intelligent choices about document content just as he would have done manually. This involves the display programs discussed in the introduction as well as the creation of programs and files for efficient storage and searching of terminology, terminal handling routines, tape I/O, etc.

The secondary indexes reflect the second stage. Using IR and CL techniques to analyze the text of title and abstracts, keywords and categories are suggested to the indexer. The tapes of citations will be run

through a batch preprocessor to attach keyword and category suggestions to each record for the indexer to accept or reject, or to trigger other possibilities in his mind. This work relies on theory and experimental evidence from Salton [6], Van Rijsbergen [7], Sparck-Jones [8], Stiles [9], and others on automatic statistical and linguistic text analysis.

The third stage represents the use of CL and AI techniques. Although there is superficially some overlap between IR and AI, the philosophical slants of IR and AI vary sufficiently to lead to diverse methodologies, particularly if we follow van Rijsbergen's lead in using Lancaster's definition: "[an IR system] does not inform (i.e., change the knowledge of) the user of the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request." This definition excludes most work in AI such as query systems, pattern recognition, representation, etc., which will be the heart of the stage 3 work. This third stage is first evaluative and then prescriptive. Does the suggestion of terms lead to greater consistency across indexers or does it just bias them? And does this consistency or bias lead to better retrieval from the RECON users perspective? Can the thesaurus be enriched by these suggested terms? How much of the human component is efficient and how much can MAI assist? As words are extracted, if they are added to the thesaurus keywords, are they more effective in document recall performance than the thesaurus alone?

These stages will be examined in more detail below.

STAGE 1

Although the bulk of work at this stage involves system design, including the writing and implementation of computer programs, the first step had to be a consideration of the indexing process itself. What does an indexer actually do, both intellectually and physically? We will describe this under the broad rubric of indexing behavior and draw on both published research and protocols taken from the TIC indexers.

Indexing Behavior

The primary function of indexing is to create representations of document content in the indexing language. This representation is accomplished in the Energy Data Base by assigning two elements to each citations:

1) One or more of 860 subject categories (Average 1.3) [10].

2) Any number of subject descriptors from the EDB controlled vocabulary [2]. What constitutes "good" indexing is universally expressed in the literature as that which promotes effective retrieval (e.g., Cooper [11]). The intellectual effort involved in this procedure is summarized by Digger [12]. Paraphrasing his breakdown and adding the steps relevant to TIC's system gives the following:

1) Scanning of the text (or abstract), title, author, and other cataloging material.
2) Assessing the nature of the document.
3) Identifying of the concepts.
4) Relating the concepts to user requirements.
5) Generating an abstract, if necessary.
6) Generating a title augmentation, if necessary, to replace the title with one that has information content, or to allow retrieval when a descriptor pair and title are combined into a subject entry [13].

7) Selecting the concepts to be indexed

   a) forming of concepts with respect to subject categories.

   b) translating concepts into thesaurus descriptors.

   Subsections a) and b) include descriptor splitting to prevent false coordination for users and descriptor flagging to identify descriptors as main headings or qualifiers [13].

8) Checking the work.

Most of these steps involve decision-making. When presented with terminal displays, as opposed to hard copy, data entry sheets and a pencil, how does the indexer choose to use them? The answer to this question will provide better definitions of the indexing process, of the information that is the most influential on indexing decisions, and how this information should be best presented.

Considering first the partially indexed magnetic tapes, we can define information fields from each record as falling into two classes.

1. Primary fields--These are the title, abstract, keyword and category fields which are always presented to the indexer. He can edit them and request a "link map" or path through the thesaurus or its permuted index which links terms and leads to those appropriate for concept definition.

2. Secondary fields--These can be requested to provide additional information. They are author, corporate author, category fields for other systems, and so forth.

Since the indexer can choose to go through the link map in any order, we propose to include a log that will keep track of the requests made for each of the fields to see what the most frequently used options are. This could lead in Stage 3 to a space state representation [14] where operators map the progress of a path from a start state to a goal state.

The System

The process will be as follows: Records will brought in sequentially and an output file will be created that will contain the changes the indexer has made to the original record. An indexer can operate in any of five modes. (This categorization is for purposes of discussion only--the indexer is aware only of one menu-driven procedure.)

1) Display mode. The primary or secondary fields can be displayed on the screen. The choice is menu driven.

2) Edit mode. Any of the four primary fields can be edited with a powerful screen editor that will provide such features as

a) positioning the cursor
b) deleting characters or words at or before cursor
c) inserting words or characters
d) searching for text string
e) replacing the character at the cursor
f) capitalizing words or letters or making them lower case
g) moving sections of text around
h) tabbing forward or backward on words
i) scrolling up or down

3) Help mode. There will be two forms of "help"--help with the editor commands and help with the main program control.

4) Link mode. When in the display mode, an automatic check will be made of the stem of each word in the title and abstract against the thesaurus. If a match is found, the word will be highlighted on the screen. It will then be possible for the indexer to trace associations to that word with his link map from the permuted index or from the thesaurus. If he enters the thesaurus display, terms associated with a term in the title or abstract are offered to

the indexer for persual. These associated terms are: broader terms, related terms, narrower terms, definitions, scope notes, date updated, used for and use terms. It is possible to move through this concept space by keyboarding the next term to be examined and, eventually, a touch of a term with a light-pen will enter this word as the next display. The Appendix is a sample of this procedure as currently implemented.

Since an abstract is a condensation of a document, the vocabulary in the title and abstract can only indicate the choice of words an author happened to use; it will not necessarily reveal all or even most of the concepts and ideas in the document. It is the indexer's problem to find his way from the natural language of the title and abstract to the allowed thesaurus terms, and the highlighting will give an entry point into this process. It was found in a similar experiment [15] mapping text into the INIS controlled vocabulary that after morphological analysis only 10% of the title and abstract text matched the thesaurus. Another experiment [16] showed that only 40% of the assigned descriptors contained in the text title and abstract of Petroleum Abstracts matched the Exploration and Production Thesaurus that was used to index that publication. Matching in neither case provided a substitute for manual indexing, but with our conception of the indexer as a component of the system, we believe that matching and highlighting will provide a significant entry point to the indexing process.

5) Enter mode. When a keyword is chosen for entry into the final keyword list, the indexer will be able to enter it from any position on the screen by a touch of the light pen or by keyboarding it into the file.

Morphological Analysis

If a manual check against the thesaurus were being performed, it would be a simple matter to disregard morphological differences that do not affect meaning. Morphology, or the study of word formation, is the least debated aspect of grammatical theory (as opposed to syntax and semantics which will be our concern in stages 2 and 3), hence, it will be the basis of our stage 1 attempt to use the system to give additional clues about directions to search in the link map. Looking again the Appendix, it can be seen that "condensers" provides a match with the thesaurus while "condenser" does not. While a person could easily map "condenser" onto "condensers," a computer cannot, nor could it relate these forms to "condensing" and "condensed." Hence we are developing a stemming routine based on the work of Lovins [17] which will be used to stem the thesaurus, as well as the text of each document. This will have several benefits.

1. It will provide more matching clue words with which to build the link map.

2. It will allow better statistics in stage 2 since occurrences of "condenser" and "condensers" will be counted together and not as unique character strings.

3. This compression will reduce the size of our word files so they will require less storage space and can be searched more efficiently.

4. It can be used in stage 3 with phrases to form conceptual relationships of interest. For instance, "physics applications," "applications of physicists," "applied physics," "physical applications," "application of physical," etc., might be considered as instances of the same concept.

Spark-Jones [8] reports that "stems never perform worse than word forms and sometimes perform better." The use of stemming varies, depending on the

subject matter in the database and the extent to which human decision making is tolerated in the stemming process.

The EDB database is, of course, devoted to all aspects of energy--technological, economic, social and political. This has a definite impact on what stems must be recognized, and Lovins' lists of suffixes will be expanded accordingly. For example, "acetoacetates" and "acetoacetic" should be considered as equivalent by a stemming routine--as, indeed, they are by Lovins' suffix list. However, "autoradiography" and "autoradiolysis" should also have the same stem, but -ography and -olysis are not present in Lovins' lists. We will modify the list to better reflect the contents of the EDB database. We are also considering creating an algorithm for removing certain prefixes. As Lovins' algorithm works quite well, such a prefix-stripping algorithm will be patterned after the suffix algorithm. Only a few prefixes will be candidates for removal--"antineutron" should go to "neutron," exoskeleton" to skeleton," and "ultracold" to "cold." Prefixes such as "de-," "ex-," and "un-" will not be removed, in general.

Stemming algorithms may err by identifying words that should be different (such as "aerial" and "aeration," which both stem to "aer-"), or by failing to identify words that should be equivalent. Because of the interactive nature of this system, the first of these is not particularly serious--the indexers can easily reject a suggested index term. The second type of error is more serious, since indexing terms are liable to be missed. Since decreasing one type of error tends to increase the other, we choose to err on the side of excessive identification.

Totally automated systems require near-perfection of a stemming algorithm, a standard that is rather difficult to meet. By reserving final decisions on

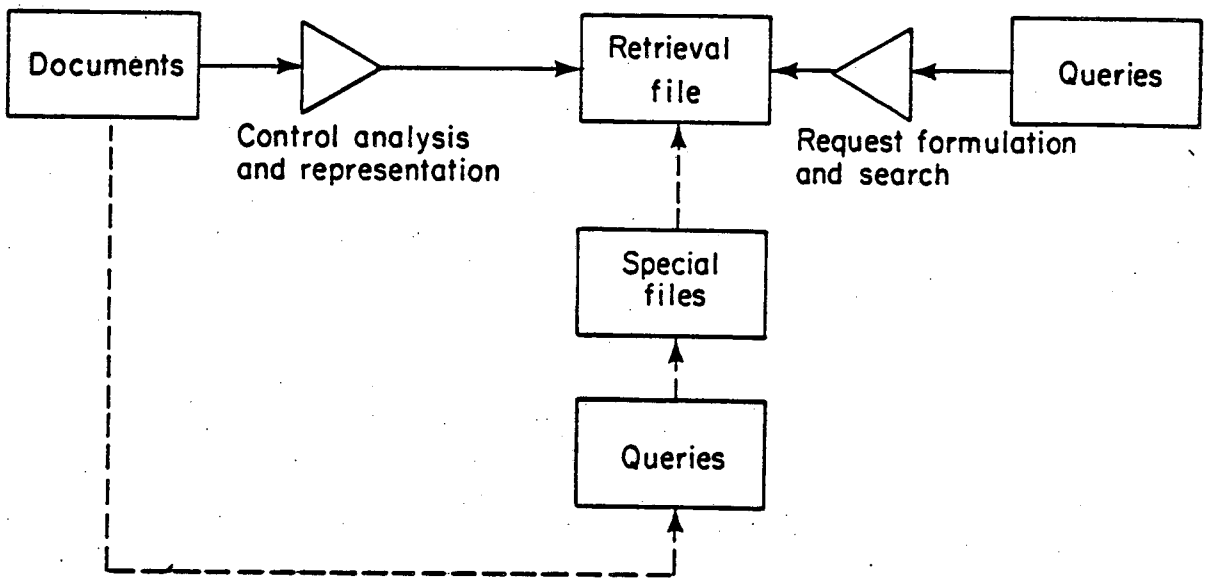
the acceptability of indexing terms for a human indexer, we can have a very successful system with somewhat less than perfect stemming.

STAGE 2

There is a parallel between the evolution of on-line searching and the development of systems for interactive indexing. s machines became capable of storing and searching large amounts of data, it became possible for users to interactively browse, formulate and change search requests on-line and get nearly instant system response. With batch searching, the computer was made to minic the manual search process, albeit faster and with a greater volume of documents. The development of on-line searching capabilities, and the ability to refine a search under way (e.g., MEDLINE [18]), represent stage 2 in the user domain as on-line interaction of the indexer with classification and vocabularies does in the indexer domain.

Additions to Smith's basic model [14] of an IR system (see page 14) shifts the role of the indexer from a creator of the retrieval file to a user as well. Such an interactive system gives the indexer access to collection statistics, words weights, on-line thesauri, etc.

We propose to achieve the dotted portion, which allows the indexer to query the retrieval file, through special files constructed from the history of the database. Bennet [19] in his Negotiated Search Facility allows the indexer access to previously indexed documents that use the terminology under consideration. This technique, while viable in a small test collection, or grist for stage 3 experiments, might overwhelm an indexer with information if he had access to RECON in its current configuration during the indexing process. Rather, the information will be stripped from RECON, condensed, and presented to the indexer as described below.



Automatic Text Analysis

Automatic text analysis breaks down generally into statistical and linguistic approaches, the latter including morphological analysis, syntactic and semantic methods. Given the magnitude of the Energy Data Base (>1,000,000 documents), the number of subject categories (861), and the size of the controlled vocabulary (>20,000 main terms), statistical techniques seemed the most accessible, particularly when combined with morphological and eventually syntactic analysis. The basis of our work is that pioneered by Luhn [20] who used frequency counts of words in the document text to determine which words were significant in representing the document content. He compiled a list of "keywords" for each document. This technique has been refined and widely used

in subsequent years for such pruposes as thesaurus construction, devising measures of word associations for retrieval, etc.

Statistical Analysis

We will initially use word frequency statistics from the title and abstract of documents for the prediction of EDB subject categories. This falls into the realm of automatic classification. As van Rijsbergen [7] points out, all classification is for a special purpose; we are not concerned that this time with a "best" classification but rather with accurately fitting documents into the preexisting EDB subject categories. Success is thus measured by the degree to which we succeed in these predictions. Most classification is eventually judged by its performance from the users perspective during retrieval, and, indeed, if the indexer is considered as user then the success is the degree to which the categories are predicted for him.

In terms of system design, there are two functions for this prediction:

1) Routing of documents to individual indexers. The 861 subject categories are hierarchally arranged with 40 first level categories, 302 second-level categories and 499 third-level categories [10]. Each indexer covers the subject matter of a number of first level categories, so it should be possible to accomplish this routing with a high degree of accuracy.

2) A prediction of category or categories to be attached to each record for the indexer's consideration, along with the statistical weights representing our degree of belief in the prediction.

An Experiment

In order to test the efficiency of this method, two first level categories and the associated 2nd and 3rd level categories were selected from the EDB category list--Physics (640000) and Petroleum (020000). A series of computer

programs were written and 12 issues (6 months) of EDB tapes with approximately 23,000 document citations were processed to form a training set.

For each record, all 1, 2 and 3 word phrases in the title and abstract were extracted as well as the keywords that had been assigned to each document. These were used to form three files, since experiments by Spark-Jones [8] have shown that title and abstract, when automatically extracted and used as index terms, have different retrieval efficiencies. Only documents in which exactly one category was assigned were used. Although this cut down the number of documents in each category, it provided a "clean" training set. To use the documents with multiple category assignments has an inherent problem. Assume a citation had been assigned to both a category on coal gasification and on chemical synthesis. The vocabulary in the title and abstract would reflect both categories and every word would be added to the list describing each category. Hence, there would be confounding since chemical synthesis terminology would be added under coal gasification and vice versa. While we realize we are adding bias to the statistics by selecting only single category documents, the confounding would probably give worse results. We wish to test this hypothesis at a future time.

Our initial training set is quite small and will be increased. It consists of 18 categories with the number of documents in each category ranging from approximately 50 to 300. The research question is how well the vocabulary in these 18 categories can predict a category for a new document. As a test, an EDB issue not in the training set was searched for documents surrogates that had been assigned to one of the 18 categories. Phrases were extracted and the words checked against an inverted index that has the frequency of each word in each category.

The weighting functions used to perform this evaluation are varients of those suggested by Field [21] who derived a measure of association between text words and classification. Using abstract words only, we tested the following coefficients:

$$w_{ij} = \frac{f_{ij}}{\sum_i f_{ij}} \cdot \frac{f_{ij}}{\sum_j f_{ij}} \qquad (1)$$

$$w'_{ij} = \frac{f_{ij}}{\sum_i f_{ij}} \cdot \frac{\log f_{ij}}{\log \sum_j f_{ij}} \qquad (2)$$

where i is the word index, j the category index and f_{ij} is the frequency of a word in a category. The first term normalizes by word; the second by category. w_{ij} is calculated for each abstract word and then a category weight of conditional probability for that category is generated.

$$P(c_j | k_i) = \sum_i w_{ij}$$

i.e., the probability of category j given the set of words k_i.

A preliminary test of this method using the very small training set with unstemmed vocabulary correctly predicted categories for 86% of the 300 documents we analyzed with weighting function (1). This crude methodology gives better results than those obtained by Field, or Hamill and Zamora [22] in a similar experiment assigning categories to documents according to Chemical Abstract groupings. It encourages us to pursue these experiments with the following types of analyses:

1) Apply the stemming algorithm to the terms in the inverted index and then stem the document terminology before processing it. This will greatly

reduce the size of the file. Additionally, we will look at phrases vs single words in prediction, in both the stemmed and unstemmed form.

2) Experiment with the number of words per category needed for prediction. There is a large body of work relating word frequency to prediction both from the perspective of retrieval performance (e.g., Svenonius [23]) and classification (e.g., Salton [6]). Although the assignment of documents to categories falls into classification, the structure of the category system makes the work on retrieval effectiveness relevant.

Words can be divided into content-bearing words and non-content-bearing words. Such words as "and," "the," and "energy" can be put into a stop list since their presence in a title or abstract will not convey any information for this document collection. Among the content bearing words, it has generally been found in retrieval experiments that if high precision is desired, narrow, or infrequently index occurring terms are more effective than broad ones, whereas if all documents on a subject are required broad indexing leads to more relevant documents. The mapping of these findings onto our classification scheme suggests some experiments.

The EDB subject categories are hierarachically arranged and the category system can be decomposed into subsystems that can be independently analyzed. Since the subject categories were devised to be independent, there should be only weak interactions between them. Rather than performing a one-step classification procedure by constructing an inverted index from the entire training set on every word with its associated categories and frequencies and calculating the weights to predict categories from that, one could utilize the hierarchical structure to construct a number of inverted indexes that are sequentially searched. This permits a hierachical prediction using different subsets of the total vocabulary that was extracted for the training set. This

category decomposition allows a vocabulary decomposition where files can be constructed on the depth of content of vocabulary. This is the point of contact with the retrieval studies. The vocabulary to predict the first level will be broad and superficial, i.e., the most predictive words will probably be very general, such as "stars" or "coal." To correctly place a document in one of the 40 first level categories possibly can be done with a small file of these highly predictive, general terms. We propose to merge the vocabulary at all levels for first level prediction and experiment with this concept by predicting categories with the 100, 500, 1000, 2000 most frequently used words in each category. Then we will test whether the prediction of the 2nd and 3rd level categories is better without some fraction of the most frequently used words. That is, will the higher precision of retrieval as shown by the more infrequent words be mirrored by a higher percentage of correct classification in the 2nd and 3rd levels?

3) Experiment with other prediction algorithms. For example, Shannon's expression for entropy, since entropy can be viewed as a measure of uncertainty or disorder seems a likely candidate.

Syntactic Analysis

Syntactic analysis occurs in both stages 2 and 3. The simple rules that we use to break text into two and three word phases are a primitive form of stage 2 syntactic analysis, while natural language parsing used in computational linguistics for AI language understanding represents the other extreme. Our needs lie somewhere in the middle, due to a number of limiting factors:

1) The indexer is not going to be "conversing" with his documents. Hence sophisticated AI techniques, the construction of grammatical parsers, formalized semantics and complex inference rules can be avoided.

2) We do not have to analyze general text. Titles and abstracts of technical documents are a special, simplier form of text with less ambiguity of meaning than the context dependent semantic information that is so difficult to handle in a newspaper or in dialogue.

3) We have a limited purpose system in that our goal is to produce suggested category and keyword descriptors. Hence, syntactic analysis need not be very deep or profound. If we can extract simple phrases to be matched against the thesaurus or against lists of vocabulary derived from frequent usage in categories, the analysis will be a success. We suspect that the program could even miss some portion of these phrases with negligible effect on indexing success. This will require some testing.

4) Again, it is the presence of the indexer as a critical component of the system that allows us to accept error rates that could not be tolerated by a completely automatic system. The discussion of this point earlier (with regard to stemming) applies here as well.

Much of the literature in MAI consists of small collections and while the results point to directions to explore, particularly in stage 3, they are not immediately applicable to our problem. An exception to this is the system of the Department of Defense Technical Information Center (DTIC) which has been operational since 1974 [3]. The document content is similar to TIC's (technical reports, journal articles, etc.) as is the size of the database. Currently 75% of their documents are automatically indexed, while 25% must be manually checked. Briefly, the basis of this system is syntatic analysis which chooses words and phrases, and checks them against dictionaries for recognition and syntactic type. This approach seems to hold promise for us and we wish to explore the feasibility of incorporating these techniques in

conjunction with thesaurus terms and category assignment into our system and using DTIC's computer programs, where possible, to carry this out.

STAGE 3

Stages 1 and 2 emulate human indexing behavior and introduce the concept of the indexer as user of the database. Stage 3 extends these concepts by optimizing the prediction of keywords through semantic analysis, provides semi-automated thesaurus expansion for better retrieval and explores another facet of indexer as user. Much of this work falls into automatic indexing, such as the SMART system of Salton [6] or the experimental collections studied and reviewed by Spark-Jones [8]. The emphasis in these experiments is on retrieval effectiveness and as Spark-Jones comments, "there have been few direct controlled comparisons between manual and automatic indexing; that is, ones in which other variables are not affected." However, the ultimate role of indexing is retrieval and though we can devise tests in stages 1 and 2 that will test how well we emulate human behavior, stage 3 will ultimately move headlong into retrieval experiments. But these experiments must focus not only on the traditional relevance, recall and precision measurements but on the link between the indexer and user, on the type of communication they can establish through dialogue or files. Retrieval strategies on the part of the user should be fed back to the indexer and the indexer should be able to make use of previously indexed documents. Walker [24] finds the "facilitating of effective communication between human generator and human user" to be the central problem in information science. If we now let the indexer as user refer not only to his being a user of condensed information in the database but a user from the retrieval perspective as well, then his insights on

retrieval through indexing, and on query formation should give information on these processes. Walker is addressing this issue in a system he is prototyping where a major objective is to gather data about the nature of problem formation and then modify the ways in which material is organized in the files and presented to the user. Another suggestion would be to use the vocabulary assigned to a document as input for a search that retrieves similar documents in the file. These are then available to a user at search time or an indexer as part of an experiment on communication between user and indexer.

These speculations might represent a leap beyond stage 3, however, and into another dimension. If we return instead to the previous lines of thought and extend them, a framework for these suggestions is the representation issue in AI. In general, representation refers to how knowledge is organized in a system, and it covers the range from what elements are chosen for document surrogates, to how keywords are related to the text, to how to represent the content of a query and the content of the data and relate one to the other. In our system, with indexer as user, we have used a simple matching procedure between words in text and words in the thesaurus; the query is implicit. But we could go beyond this level to a more complex representation using, for example,

1) Synonym dictionaries to expand choices for thesaurus terms.

2) Semantic networks [25] as a varient of the thesaurus representation. A thesaurus is a semantic network in the sense that the relationship between terms (nodes) are arcs labeled BT, NT, RT, UF or USE. But if each arc could now have a weight associated with it, a semantic space would be created that would give clues for indexing and retrieval of documents.

3) A more sophisticated morphological analysis [26] such as diagrams where pairs of consecutive characters are the attributes to be compared.

CONCLUSION

The motivation for this work came from the need to quickly develop an on-line editing system for indexers in a production environment while recognizing that the problems involved in an efficient system spanned information science and touched upon artificial intelligence and computational linguistics. When presented with the opportunity to develop such a system and to work with trained subject specialists the reaction from a theoretical perspective is to let one's mind wander over the possibilities and potentialities for research. They span:

1) indexing behavior, from how indexers use knowledge to how they interact with a terminal and editor,

2) linguistic analysis of natural language-morphology, syntax, and semantics,

3) statistical techniques for analyzing vocabulary,

4) representation systems,

5) retrieval experiments.

Some of these areas will directly and immediately have an effect on the indexer and indexing behavior. Others hidden to the indexers will give an incremental increase in performance of the system. Yet others will expand our knowledge of cognitive behavior, and the uses of information by indexers leading to a new look at methods of indexing and the impact of this indexing on users and retrieval systems.

REFERENCES

1. DOE/RECON: Manual for the Dial-Up User, TID-4587, TIC.

2. DOE Energy Information Data Base: Subject Thesaurus, TID-7000, NTIS.

3. Paul H. Klingbiel, Machine-Aided Indexing of Technical Literature, Information Storage and Retrival, vol. 9, pp. 79-84 (1973).

4. L. C. Smith, Artificial Intelligence in Information Retrieval Systems, Information Processing and Management, vol. 12, pp. 189-222 (1976).

5. J. D. O'Connell, E. C. Fubini, K. G. McKay, J. Hillier and J. H. Hollonon, Electronically Expanding the Citizen's World, IEEE Spectrum, vol. 6, pp. 30-40 (1969).

6. G. Salton, The SMART Retrieval System, Prentice Hall, Inc., Engelwood Cliffs, NJ (1971).

7. C. J. van Rijsbergen, Information Retrival, Butterworths, London (1979).

8. K. Spark-Jones and R. G. Bates, Research on Automatic Indexing 1974-1976, Report, Computer Laboratory, University of Cambridge (1976).

9. H. F. Stiles, The Association Factor in Information Retrieval, Journal of the ACM, vol. 8, pp. 271-291 (1961).

10. DOE Energy Information Data Base: Subject Categoris, TID-4584, TIC.

11. William S. Cooper, "Is Inter-Indexer Consistency A Hobgoblin?" American Documentation, vol. 2, pp. 268-278 (1969).

12. Jeremy A. Digger, A Study of the Intellectual Elements of Indexing for Information Retrieval, thesis submitted for fellowship of the Library Association (1973).

13. Guide to Abstracting and Indexing at the Technical Information Center, TID-4583, NTIS.

14. S. Amarel, Problem Solving and Discussion-Making by Computer: An Overview. In: Cognition: A Multiple View, P. L. Garvia, ed., pp. 279-329, Spartan Books, New York (1970).

15. Lynn Evans, Evaluation of the ISPRA Automatic Indexing Programs SLC-II. Final Report, INSPEC (July 1978).

16. Ray W. Graves and Donald P. Helander, A Feasibility Study of Automatic Indexing and Information Retrival, IEEE Translations on Engineering Writing and Speech, vol. EWS-13(2), pp. 58-59 (1970).

17. Julie Beth Lovins, Development of A Stemming Algorithm.

18. D. B. McCarn and J. Leiter, On-Line Services in Medicine and Beyond, Science, vol. 181, pp. 318-324 (1973).

19. John L. Bennett, On-Line Access to Information: NSF as an Aid to the Indexer/Cataloger. American Documentation, vol. 20, pp. 251-267 (1979).

20. H. P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development, vol. 1, pp. 309-317 (1957).

21. B. J. Field, Towards Automatic Indexing: Automatic Assignment of Controlled-Language Indexing and Classification from Free Indexing, Journal of Documentation, vol. 31(4), pp. 246-26 (1975).

22. Karen A. Hamill and Antonio Zamora, The Use of Titles for Automatic Document Classification, JASIS, vol 7(1), pp. 397-402 (1980).

23. E. Svenonius, An Experiment in Index Terms Frequency, JASIS, vol. 15, pp. 109-121 (1972).

24. D. E. Walker, The Organization and Use of Information: Contribution of Information Sciences, Computational Linguistics and Artificial Intelligence, JASIS, vol. 32(5), pp. 374-363 (1981).

25. Findler, N. V., ed., Associative Networks--The Representation and Use of Knowledge in Computers, Academic Press, New York (1979).

26. G. W. Adamson, and J. Borcham, The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles, Information Storage and Retrieval, vol. 10, pp. 253-260 (1974).

SEQUENCE OF CURRENT PROCESSING STEPS

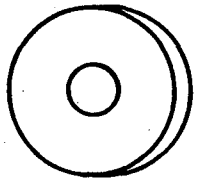
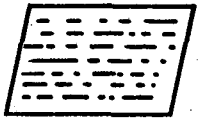
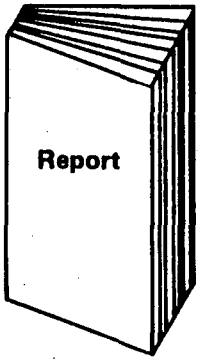
Source	Batch check	Indexer	Descriptive cataloguer	Batch check	Indexer	Batch check	Descriptive cataloguer	Recon tape
Tape:  GRA EI AIP etc	√	Printout  Manual proofing, additions, deletions	Keyboard changes	√	Verification by eye and hand at rejected records	√	Any further changes	
Document:  Report		Manual writing of keywords, categories, (abstract)	Enter other fields Keyboard all!	√		√		

Figure 1

SEQUENCE OF INTERACTIVE INDEXING

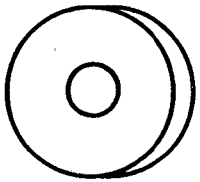
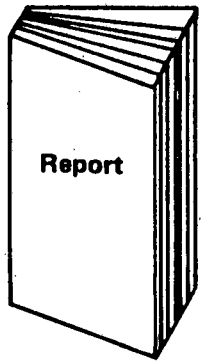
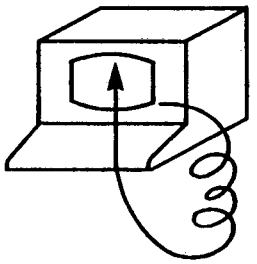
Source	Batch check	Indexer	Descriptive cataloguer	Batch pro-cess	Indexer	Batch check	Descriptive cataloguer	RECON tape
Tape:  GPA EI AIP etc Document:  Report			Enter title, (abstract), keywords other fields	√ √	 – Screen editor for text correction – Keyword suggestions – Category suggestions – Light pen entry	√ √	Enter other fields Any further changes	

Figure 2

A tape (GRA, AIP, EI, etc.) is read and the title and abstract fields are extracted.

Title: Gerdien condenser instrumentation for measuring high-latitude middle atmosphere electrical parameters. Special report

Abstract: Gerdien condensers for measuring electrical conductivity, ion mobility and charge number density were flown in recent rocket programs to investigate the high-latitude middle atmosphere. The instruments were launched in two coordinated programs (Aurorozone I and II) at Poker Flat, Alaska to study the effects of auroral energetics on electrical parameters and in a solar eclipse rocket program at Red Lake, Canada. The design of the Gerdien condenser instrumentation for the Aurorozone II program and the solar eclipse program is considered. In addition, electrical parameters measured for the two auroral programs are presented and discussed. The initial results from the measurements indicate that high-latitude middle atmosphere electrical parameters are strongly influenced by the auroral energetics. (Author)

All the words in the title and abstract are checked against the on-line thesaurus and highlighted if there is a match. (We are currently working on stemming routines so both "condenser" and "condensers" will be highlighted)

The highlighted terms are then checked against a permuted index of the thesaurus main terms, followed by a prompt to the indexer. Either all of the permuted index terms for a highlighted term can be displayed, or some selection from them.

```
Permuted index entries for all highlighted terms?
        Type y or n:

y

Entries for condensers
Condensers
Ice Condensers
Isolation Condensers
Steam Condensers
Vapor Condensers
```

```
Entries for mobility
Carrier Mobility
Electron Mobility
Hole Mobility
Ion Mobility
Ion-Mobility Detectors
Mobility
Particle Mobility
```

etc.

When the list is exhausted, an option is given to go on to the next record.

If the indexer sees any term he wishes to enter into the keyword list, a touch of the lightpen to the screen will perform that function. Likewise, he can delete keywords from the list or edit them.

```
Type return for more; ^C exits:
```

Title: Clinical investigation. Annual research progress report no. 15 (final), for fy 79

Abstract: Subject report identifies the research activities conducted by Fitzsimons Army Medical Center investigators through protocols approved by the Clinical Investigation Committee and registered with the Clinical Investigation Service during Fiscal Year 1979 and other known presentations and publications by the Fitzsimons Army Medical Center professional staff. The research protocols described were conducted under the provisions of AR 40-38, as amended, Clinical Investigation Program, AR 40-7, Use of Investigational Drugs in Humans, AR 70-25, HSC Reg. 40-23, Management of Clinical Investigation Protocols and Reports, Use of Volunteers as subjects of research and AR 40-38, as amended, Clinical Investigation Service, policies and procedures, to insure the medical being, preservation of rights and dignity of human subjects Who participated in these investigations. (Author)

```
Permuted index entries for all highlighted terms?
        Type y or n:

n
```

```
For some of the terms?
    Type y or n:
```

```
y

Please type one term at a time from the list
   Type FIN to end ----- LIST to review list

Drugs
Management
preservation
investigations
```

This list contains all the highlighted terms in the title and abstract.

```
 Enter word
drugs

Entries for Drugs
Antimitotic Drugs
Antineoplastic Drugs
Antithyroid Drugs
Drugs
Immunosuppressive Drugs
Psychotropic Drugs
Radiomimetic Drugs

 Do you wish thesaurus display?
       Type y or n:
```

At this point, it is possible to go into the thesaurus heirarchical display. If any of the BT's, NT's or RT's are desired in the keyword list, they can be entered by a touch of the lightpen to the screen.

y

BROADER TERMS	RELATED TERMS	NARROWER TERMS
		ANALGESICS
		Acetylsalicylic Acid
		Codeine
		Morphine
	Antiseptics	ANESTHETICS
	Chemotherapy	Cocaine
	Drug Addiction	Thiopental
	Pharmacology	ANTIBIOTICS
		Acrtinomycin
	----------	Penicillin
	DRUGS	Puromycin
	----------	Sulfadiazine
		Tetracyuclines
	Therapy	ANTINEOPLASTIC DRUGS
	Toxicity	Chlorambucil
	Vitamins	ANTIPYRETICS
		Cinchonine
		Quinine
		ANTITHYROID DRUGS
		Thiouracil

Since the narrower terms are longer than one screen-load, the following command displays the rest of them.

PLEASE ENTER NEXT COMMAND-- +nt

BROADER TERMS	RELATED TERMS	NARROWER TERMS
		Thiourea
		IMMUNOSUPPRESSIVE DRUGS
		ISOIAZID
		METHYLENE BLUE
		PSYCHOTROPIC DRUGS
		RADIOMETRIC DRUGS
		RADIOPHARMACEUTICALS

	DRUGS	

It is now possible to move through the concept space of the on-line thesaurus by entering main terms. For example, the RT's to DRUGS suggested Drug Addiction as an allowed term which the indexer chose to examine.

```
PLEASE ENTER NEXT COMMAND-- drug addiction

      BROADER TERMS                RELATED TERMS              NARROWER TERMS

                                   Drugs

DISEASES                           -------------------------
    Addiction                      DRUG ADDICTION
SOCIAL PROBLEMS                    -------------------------

PLEASE ENTER NEXT COMMAND-- return

 Enter word
fin

Type return for more. ^C exits:
```

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720