

II. User Interaction With An On-Line Retrieval System

The full search retrieval algorithm returns to the user the n documents with concept vectors "closest" to the query vector as measured by the angle between vectors (cosine correlation). If the user's original query is an accurate and complete description (in "concepts") of his need, and if the documents relevant to the user are clustered "close" together in the space of concept vectors, this algorithm can isolate these few relevant items from a large collection of irrelevant material. However, neither of these conditions is common in practice. It is evident from experiments with the SMART system that a user familiar with the subject area but unaware of vocabulary and word frequency effects on the search process is unlikely to formulate an initial query that provides optimum retrieval [6]. It is unreasonable, however, to expect each user to understand the fine details of the document classification system.

Further, there is evidence in the experimental document collection used here that the documents judged relevant by the users are not always clustered neatly in the concept vector space. Even with full knowledge of the document collection it is often impossible to formulate a single query that will rank all relevant documents above all non-relevant documents. This may indicate flaws in the text-to-vector mapping used for this study. However, the needs of the human users of document collections are so diverse that a subject classification system appropriate for all queries may not exist, or may be impractical to implement.

Since the user's original query is often inadequate, some sort of user interaction with the retrieval operation is desirable. The user of a manual retrieval system such as a library might at first ask a general and unclear question. The librarian, using his knowledge of the document collection, might then ask the user a few questions and show him a few books in an attempt to pinpoint his needs. Recent technological developments encourage the investigation of similar types of user feedback in automatic retrieval systems. Large capacity random access memory devices allow the storage of natural language document titles and abstracts. On-line low speed terminals and time-sharing techniques may be used to provide real time interaction with many users at once, at several convenient locations.

Two major considerations arise in such an on-line system. In the present batch-processing systems, such as NASA and Medlars [2] immediate response to the user is not necessary. In an interactive system the computer time required to process a single query takes on a new importance. The low input-output speeds of those terminals appropriate for interactive applications introduce a second limitation. [7] For example, typing out a single document abstract on a typewriter terminal could easily consume more time than the

computer retrieval operation. An interactive document retrieval system therefore requires an efficient retrieval algorithm and a minimum of necessary interactive input and output.

Several methods of user interaction have been tested in the SMART system using the document collection employed in this study (the 'Cranfield 200' collection described in Section IV). Results of this investigation^[6] are summarized below.

The interactive strategies tested can be divided into pre-search and post-search algorithms. In pre-search interaction, information is presented to the user and a new query is constructed by him before the search operation takes place. The "repeated concepts" algorithm asks the user to choose one or more of his query terms to be repeated for emphasis. The "word frequency" technique displays for the user the frequencies with which his query terms occur in the document collection. The user is then invited to eliminate or change query terms that are too common or too rare to be useful for retrieval. Both of these displays help the uninformed user to take advantage of the effects of word frequency in a retrieval system using frequency-weighted vectors for document classification. The "thesaurus display" supplies synonyms and terms related to the terms of the initial query from a stored thesaurus appropriate to the subject area. The thesaurus used for

this display in reference 6 is the "regular thesaurus" described in Section I of this report. Since the same thesaurus can be incorporated automatically into the SMART system, manual and automatic thesaurus procedures are compared in reference 6. The automatic application of the thesaurus to document and query vectors gives better retrieval results than the manual thesaurus display, except at low recall levels. The "source document display" exhibits concepts assigned to a relevant document known to the user before retrieval. When this display is used in addition to the automatic thesaurus, results are better than with automatic thesaurus alone.

Post-search techniques display the partial results of an initial search operation so that the user can reformulate his query and request another search. These algorithms may be iterated as often as the user desires. All post-search algorithms share a common disadvantage, the computer time required for several search operations. The time is well spent, however, for all post-search techniques investigated give better retrieval than automatic thesaurus display. "Title display" which displays the titles of the first n (in this reference $n=5$), documents retrieved by the initial query, provides better retrieval than thesaurus display except at high recall. "Abstract display", which displays n full abstracts, requires more output time and more time for user thought, but gives consistently better

performance than title display. A variation of "relevance feedback", the technique investigated in this study, gives retrieval results nearly comparable to abstract display. Moreover, this report gives more effective variations of the relevance feedback algorithm than the version used by Lesk and Salton.* When pre-search and post-search information is combined, manual thesaurus display followed by abstract display gives better retrieval than either method alone. Adding word frequency information to the combination is helpful when the null thesaurus is used.

Estimates of the search cost per query show that abstract display, which gives the best overall performance of the methods tested, is the most expensive. The other post-search algorithms, title display and relevance feedback, are more costly than any pre-search method. Relevance feedback requires the least user effort of any post-search strategy. Lesk and Salton [6] recommend the following algorithms:

- a) For normal users needing high recall, automatic thesaurus followed by automatic relevance feedback.
- b) For highest precision when high recall is not required, word stem matching followed by title display.
- c) For experienced and patient users needing maximum performance, thesaurus display plus frequency information followed by abstract display.

*Lesk and Salton use the Q₂ Strategy with N equal to 5. See Sections VI-C and VI-D for more effective algorithms.

The Lesk and Salton study shows that relevance feedback is one of the most effective user interaction techniques. In relevance feedback, the user is given a small set of items retrieved using his original query. He is then asked to judge which items of this set are relevant to his needs. This information is used to automatically produce a new query for another search. This feedback process can be iterated as often as desired. Relevance feedback has a definite psychological advantage over abstract display; the user is not required to make sophisticated decisions in rephrasing his own query. Instead, he can supply much information to the retrieval system at little effort by saying in effect "I want documents on the same subject as this document". The stored abstract of a chosen relevant document contains a more detailed description of the subject than a user would care to type as a query. In the experimental collection, the document vectors have approximately ten times as many concepts as the query vectors, so the user submits a ten times more detailed "query" simply by typing a document identifying number.

The disadvantages of relevance feedback should be ~~reiterated~~. Like abstract display, relevance feedback requires the system output of document abstracts or information of comparable detail. Also, multiple searches of the document collection are made. Designers of retrieval systems must decide whether the extra output time and computer time is justified by the retrieval improvements obtained.