I. Automatic Document Retrieval Systems

The conventional library classifies documents by numeric subject codes which are assigned manually (Dewey decimal system, Library of Congress system). Cross-indexing is provided in a card file by "subject" title, and author. Both the numberic index and the "subject" cross-indexing may be inadequate for retrieval. A book on the intersection of two subjects (e.g., "The Aerodynamics of Birds") or on a new subject (e.g., automata theory --- is it mathematics, computer science, logic?) is hard to classify and therefore hard to find, unless the librarian is asked where he filed it. A fully automatic system must duplicate the function of this librarian by extracting information from a natural language request and retrieving precisely those documents most likely to be needed by the requestor.

One method of subject classification that is used in automatic retrieval systems assigns to each document a list of subject identifiers, often called "keywords". This list can be treated as a binary vector by associating a position in the vector with each possible keyword in the retrieval system. The value in a vector position is one if the associated keyword is assigned to the document described by the fector, zero otherwise. Retrieval systems operated by NASA and by the National Library of Medicine (Medlars) use this type of subject classification [2]. In

both of these automatic retrieval systems, keywords are assigned to documents manually by subject experts.

An extension of keyword indexing represents each document as a positive weighted concept vector rather than a binary vector. Each classification concept is weighted to indicate its importance in the document. In the SMART retrieval system, these concepts and weights are assigned by automatic processing of the natural language text of each document or abstract. [3]

The user's query in an automatic information retrieval system can take several forms. The user may be asked to formulate his query using a restricted language. This language usually includes the set of keywords defined for the collection and sometimes the Boolean operations "and", "or", and "not". In the NASA system the user can assign values to each keyword instead of using the logical operations. Cross-referencing and hierarchal relationships among keywords can be used in both the NASA and Medlars systems to refine or expand the user's initial query. Both systems require the user to understand the indexing system in order to formulate effective search requests.

In the SMART retrieval system, the user is asked to phrase his query in natural language. The query is then processed in the same way as the document abstract, and a query concept vector is created. Logical relationships are not used in the query analysis.

SMART provides a fully automatic information storage and retrieval system of a relatively simple form. Document

abstracts are analyzed to construct representative concept vectors which are stored in the computer. When a user types a natural-language request into the system, it is converted to a concept vector representation in the same manner. Several types of automatic text-to-vector conversion have been used with the SMART system [3]. A list of common words to be ignored in constructing the concept vector is provided. A suffix dictionary is used to reduce all words to word stem form. A word stem thesaurus which treats each distinct word stem as a concept in the concept vector is used as an experimental standard. Frequency characteristics may be used to eliminate some concepts ("partial stem thesaurus), for example, words occurring less than five or more than 100 times in a given collection may be eliminated. This study uses a thesaurus which was constructed semi-automatically for the subject srea of aeronautical engineering. This "regular thesaurus" recognizes synonyms; that is, it converts words of the same meaning to the same concept, providing better retrieval performance than the stem thesaurus [4].

The degree of relationship between a query and a document is determined in the SMART system by some "distance function" of the query and document concept vectors. The most effective of the distance functions tested in SMART appears to be the cosine correlation, which measures the angle between concept vectors in n-dimensional space [4,5]. The cosine coefficient of two concept vectors ranges from 0

to 1, and is found by the formula

$$\cos(r,s) = \frac{\sum_{1}^{n} r_i s_i}{\sqrt{\sum_{1}^{n} r_i^2 \sum_{1}^{n} s_i^2}}$$

The process of determining the relationship of each document in the collection (or some subset thereof) to the user's query is called a "search" operation. The distance function is used to assign to each document a correlation coefficient indicating the relationship between the concept vector for that document and the query vector. The document identification numbers are then ordered by correlation coefficient and are assigned ranks from one to N (number of documents being searched) for evaluation purposes. The document most closely related to the user's query is assigned the rank 1 (considered the "highest" rank).

The retrieval algorithm is the goal of any information retrieval system. For each query, the system must produce a set of documents relevant to the requestor's need. In the SMART system the retrieval algorithm can be varied experimentally. The retrieval algorithm applies the search operation; it may select subsets of documents to be searched, and it may conduct several search operations in response to one user request. The details of the retrieval operation are the primary concern of this study.

One of the simplest retrieval algorithms, here called the "full search" algorithm, performs one search operation

using the entire document collection, and selects for retrieval the highest n documents in the ranked list resulting from the search operation. In an operating system each user could select this n; in the SMART system n is an experimental parameter.