I. The SMART Project — Status Report and Plans

G. Salton

1. Introduction

The SMART document retrieval system has been operating on a 7094 computer since the end of 1964. The system takes documents and search requests in English, performs a fully-automatic content analysis of the texts, matches analyzed documents with analyzed search requests, and retrieves those stored items believed to be most similar to the queries. The system has been used largely as an experimental tool for language analysis and for the evaluation of the effectiveness of many different types of analysis and search procedures. Thus, the potential benefits, as well as the complications which would have arisen from an operational implementation in a user environment were initially given up in favor of a system which could be operated in a controlled laboratory environment. As a result, it has been possible to conduct several hundred analysis and search experiments, using document collections in the areas of computer science, documentation, and aerodynamics, and considerable information is now known about the effectiveness of fully automatic retrieval procedures, and about the design of automatic information systems.

The general evaluation results obtained with the aforementioned collections are summarized in Section III of the present report. More detailed failure analyses and more thorough descriptions of the test environment and test results are contained in a subsequent report in this series (No. ISR-13) to be issued in the fall of 1967.

Even though the original SMART experiments were conducted in a laboratory-type environment, the basic aim of eventual conversion of the system into a prototype for future fully-automatic information systems has been pursued for the following reasons:

1) some criticisms concerning the doubtful value of experimental results can only be dispelled by demonstrations in an operational environment;

2) the effectiveness of the automatic search and analysis methods is likely to be improved if the user can be made to participate in the search process, since it then becomes possible to generate a search strategy geared to the information needs of individual users;

3) the development of equipment organizations which make it possible, more or less simultaneously, to provide service to a number of different users who may be located far away from the central equipment, offers reasonable prospects for an early implementation of real-time information systems.

Accordingly, an increasingly large number of search experiments has been performed in which a user environment is simulated by running iterative searches based on user feedback. Specifically, search strategies have been developed and evaluated which use feedback information obtained as a result of an initial search, to alter the search requests so as to bring them in close coincidence with the users' needs. These new requests can then be altered as before by new feedback information, and the procedure can be iterated as many times as required.

Since the user is held responsible for providing the information to be returned to the system, it is obviously not possible to perform a full

search of all stored items, except where very long search times are tolerable. Fast cluster search strategies have therefore been implemented and tested, based on a document grouping method, which restrict the actual search to only certain document groups for each given request. The cluster search experiments are described in Sections IV to VII of the present report, and the user feedback operations in Sections VIII to XII.

The main experimental results are summarized in the present section, and future plans and projects are outlined.

2. Experimental Results

The SMART experiments outlined in this report have all been performed using one or more of the following document collections: a collection of 780 abstracts of documents in the computer literature (IRE-3), a collection of 200 abstracts in aerodynamics (CRAN-1), and a collection of 82 short papers in the field of documentation (ADI). Manually assigned keyword lists were available in addition to the abstracts for the aerodynamics collection, and the ADI collection was available in abstract as well as in full text form. Most experiments were run in parallel over all three collections, except, of course, the comparisons with the manual indexing and the full text processes which could be performed only for the specific collections for which the necessary input was available (CRAN-1 and ADI, respectively).

Except for certain minor deviations, a ranking of the various analysis and search procedures in decreasing order of effectiveness produces the same output for all three collections. That is, methods which

are effective in one context are effective also in the others, and contrariwise, the less effective procedures turn out to be generally less effective. Specifically, the procedures based on synonym recognition, weighted content identifiers, cosine correlation to match documents and search requests, and document abstract processing are always more effective than methods using simple word stem matches (without synonym detection), nonweighted terms, correlation methods based only on the number of matching terms, and analysis procedures which consider only the titles of the documents being examined.

The following specific conclusions appear to hold generally for technical document collections, and are not likely to be reversed by future tests with larger collections in different environments:

a) document length: document abstracts are more effective for automatic content analysis purposes than document titles alone; further improvements appear possible when abstracts are replaced by larger text portions; however, the increase in effectiveness is not large enough to warrant the conclusion that full text processing is always superior to abstract processing;

b) term weights and matching functions: weighted content identifiers are more effective for content description than nonweighted ones, and the cosine correlation function is more useful as a measure of document-request similarity than the overlap function; service can be improved by using sophisticated request-document matching functions;

c) word normalization procedures are most effective if the vocabulary is redundant and relatively non-technical;

d) dictionaries providing synonym recognition produce statistically significant improvements in retrieval effectiveness compared

with word stem matching procedures; the improvement is largest for dictionaries obeying certain principles with regard to the word groupings which are incorporated;

e) phrase generation methods, whether implemented by dictionary look-up or by statistical association processes, appear to offer improvements in retrieval effectiveness for some recall levels by introducing new associated information identifiers not originally available; the improvement is not, however, sufficiently general or substantial, when averages over many search requests are considered, to warrant incorporation into automatic systems, except under special circumstances where suitable control procedures can be maintained;

f) the average performance of most hierarchy procedures does not appear to be sufficiently promising to make it reasonable to advocate their immediate incorporation in an analysis system for automatic document retrieval;

g) fully-automatic text processing is not substantially inferior to manual indexing methods; for large and heterogeneous collections, a clear advantage should result for the automatic text analysis, because of indexer variability, and the difficulties of insuring a uniform application of a given set of indexing rules to all documents; the computer process in such cases does not necessarily decay as the collections grow larger;

h) the order of merit of the tested procedures is approximately as follows:

   i) most effective: abstract processing with phrase and synonym recognition;

   ii) next most effective: weighted word stem matching and statistical word associations using abstracts for analysis purposes;

   iii) less effective: logical word stem matching disregarding term weights;

iv) least effective: title processing using only document titles for analysis purposes, and document - request matching based on overlap function;

i) feedback procedures based on relevance judgments submitted by the users as a result of an initial search operation appear effective in producing large-scale increases in performance during subsequent search iterations;

j) partial cluster searches using an examination of selected document groupings only rather than a complete match with all stored items seem to promise drastic reductions in search time, at only minor costs in recall and precision.

In the next few pages, plans are outlined for the implementation of additional search and retrieval experiments, using the laboratory environment provided by the SMART system. Furthermore, new design features are suggested which will make it possible to implement the automatic analysis and search system under conditions which approximate closely the presently accepted standards for operational information systems.

3. Future Plans

A) Additional Evaluation Experiments

The validity of the evaluation results previously outlined is difficult to challenge on technical grounds because of the controlled conditions under which the testing was carried out, and also because of the relative universality of the results over all the collections used. Whatever criticism may be voiced is generally based on more formal grounds: the small size

of the test collections, the artificial preparation of some of the requests, the relative output results based on comparisons between pairs of methods, rather than on absolute performance measurements.

The following new experiments to be carried out may respond, in part, to some of the above criticisms:

a) experiments with larger document collections, including a collection in the area of documentation generated at Euratom in Ispra for which request - document relevance judgments are available from several different users; this makes possible an evaluation of the effect of differences in user relevance judgments; in addition, the aerodynamics collection (CRAN-1) can be increased in size from 200 to 1400 documents, while keeping the query set constant to permit the performance of tests based on changes in the generality ratio (percentage of relevant items to total items in a collection);

b) experiments with collections tied to an operational environment, for example, by using document and request selections from the semi-automatic Medlars system, operating at the National Library of Medicine, in order to test how the fully-automatic SMART procedures compare with standard Medlars process carried out in a real-user environment;

c) experiments with collections in less technical subject areas, such as news articles published in Time, or the New York Times, for which search requests are also available from real-users; such an experiment makes it possible to ascertain whether retrieval performance in fact degrades as the subject field becomes more heterogeneous or less technical.

Further experiments are also planned in the areas of user feedback procedures and partial searches of the stored collection with fast response

times. Specifically, the presently used two-level searches, based on a match of each search request with a set of document group vectors, followed by a comparison with those individual document vectors only which are located in certain special document groups, can be extended to three-level or higher-level searches, where the document group vectors are themselves grouped into supergroups; these supergroups are then examined first, before selected lower-level groups are considered. In each case, an attempt is made to reduce the number of items actually examined without incurring a simultaneous loss in search effectiveness.

For systems for which operating experience has been accumulated in the form of requests already processed, it may be possible to use this experience to gain in the processing of new incoming items. Specifically, request clustering experiments can be performed based on groupings of requests previously submitted. New requests can then be matched against the request group vectors, and documents previously retrieved in answer to highly matching request groups can again be submitted to the users. If homogeneous user populations are present, then a request clustering system will permit the processing of new requests which are somewhat similar to old ones already in the system, at a minimal cost in search and retrieval time.

Additional user feedback experiments are also planned including, in particular, the following:

a) the negative feedback method described in Section IX of the present report must be checked out with actual document collections; this process attempts to retrieve relevant docu-

ments from a collection even though the user is able to supply only negative information in the form of specifications for nonrelevant items previously retrieved;

b) further feedback experiments based on citation or author information of documents related to the user's needs can be carried out; in particular, an initial search request might be replaced by a list of documents which are known to be germane to a user's need, the instructions being to retrieve documents similar to those initially supplied;

c) ambiguous requests, identified by the submission of feedback information falling into different subject categories – some documents being promoted during subsequent search operations and others being demoted – can be broken down automatically into several non-ambiguous requests, each being answered by a homogeneous document set;

d) the user feedback system which presently operates separately, and apart from the fast-response document clustering system, must be incorporated into a complete user-oriented search system to simulate the conditions which users are likely to meet under actual operating conditions; specifically, such a system includes provision for iterative, user-controlled searches based on feedback information, while at the same time insuring response times not in excess of a few minutes to ensure the user's cooperation during the search.

B) New Real-Time Operating System

The original SMART system, presently operating under a batch-processing monitor on an IBM 7094 and a CDC 1604, was designed as an experimental tool, and includes programs which are deliberately experimental in nature. In fact, it was believed more important initially to analyze search results in great detail, and thoroughly to understand the retrieval

situations pertaining to many different types of running conditions, than to come forward with a few pieces of relatively poorly understood material relating to real-life situations.

While the current analysis tasks cannot as yet be considered to be terminated, as shown in the preceding paragraphs, the time is nevertheless at hand when a batch-processing type system should be supplemented by a system capable of furnishing real-time responses to user requests, and of providing individual service to individual users. In addition to furnishing an automatic retrieval tool which corresponds more directly to the type of situation which users may face in the not too distant future, an automatic real-time system will make it possible to carry out a number of new studies which cannot easily be superimposed onto the current SMART operating system. The following features, in particular, depend to some extent on real-time capabilities:

a) the implementation of a limited conversational system which would enable the users to supply information relating to user needs, and the system to respond to user indications by performing appropriate tasks and returning appropriate comments;

b) the implementation of a display system capable of furnishing to the user, selected portions of the stored files, including dictionary excerpts, and excerpts of the stored document information, to be used during the search and retrieval operation;

c) the monitoring of system parameters which are likely to influence operating strategies and costs, such as the response time, the fraction of time usefully devoted to retrieval tasks by the users and by the computer, the fraction of time spent by

the users waiting at the consoles, and so on;

d) the simultaneous accommodation of several users, some of whom may be located at some distance from the central equipment, and all of whom may require access to the same central data store;

e) the processing of background retrieval tasks, for example, in the form of standing requests, or standing user profiles, as required in a system for the selective dissemination of information;

f) the provision of limited facilities for question - answering (fact retrieval) in addition to the document retrieval system, for example in the form of stored tables of data, and facilities for handling requests answered by the tabular information.

These tasks can be implemented even without full time-sharing capabilities, by simply providing a small number (two or three) of appropriate input-output stations, and facilities for attaching these stations to the central equipment. Initially, only one user would be serviced at any one time, depending on the particular scheduling algorithm in use. The file storage could initially be implemented by disk files, for document collections of limited size, and might eventually be expanded to a tape strip store (such as, for example, a data cell) providing access times of the order of a minute to document files of from 50,000 to 100,000 items.

If the initial implementations prove viable, the real-time system could eventually be expanded to a full time-sharing facility, where several different user classes receive more or less simultaneous service, while sharing all computer facilities including memory space, stored files, input-output consoles, and the like. In that case, a monitor system might be

provided to handle many different tasks of varying priority.

Two principal criteria for choosing the particular task to be handled at any given instant could then be the priority status for that task, and the state of availability of the required files in the fast-access store. As the importance of the priority status is increased and that of availability is decreased, the system approximates an individual request time-sharing system where the user can expect interruption of all current tasks in order to free the resources for current request processing. On the other hand, as the importance of file accessibility takes on major importance, at the expense of priority, the system operates as a batch processor in which current tasks with accessible files are completed before any new task is commenced.

With the real-time implementation of the automatic SMART analysis and retrieval procedures, it should be possible to transform the present experimental laboratory type system into a prototype of an information handling facility of the kind to be expected in a real user environment in the not too distant future.