I. A PROGRESS REPORT ON SMART

G. Salton

1. Introduction

The SMART project was initiated in the fall of 1962 with the objective of designing, and implementing on a computer, a fully automatic document retrieval system, capable of processing documents and search requests available in English, and of retrieving those documents most nearly similar to the corresponding search requests. The system was to be used not only for retrieval purposes, but also as a means for evaluating the effectiveness of the search and analysis techniques.

The emphasis on fully-automatic procedures, resulting in the design of a system in which natural language texts are processed without any prior manual analysis, was motivated primarily not by certain crisis situations in the information field, or by making appeal to the so-called "information explosion." Rather, it was recognized that, in the long run, the manual indexing of every information item to be processed could not be considered a practical alternative to automatic operations, in part because the increasingly large numbers of required subject experts could not be expected to become available when needed.

The other main emphasis, on systems evaluation, stems from the fact that very little is known about the effectiveness of automatic retrieval systems. It therefore becomes essential to investigate the appropriateness of a variety of evaluation measures, and to apply these measures in experimental situations before operational systems would actually be designed and implemented.

The SMART system which eventually emerged can be operated on the IBM 7094 computer, both as an experimental automatic retrieval system, and as a testing device for determining performance characteristics of a large variety of automatic analysis and search techniques. The system is characterized by the following properties:

(a) it is designed for eventual operation in a time-sharing environment, in such a way that a multiplicity of users are given simultaneous access to the files;

(b) the search function can be undertaken iteratively under user control, by performing several partial searches to approach the desired subject area, rather than a single one-shot process;

(c) several different analysis and search procedures are incorporated in the system, including in particular a number of stored intelectual aids for vocabulary normalization;

(d) the computation of sophisticated correlation measures between documents and search requests makes it possible to present _ranked_ document output in answer to the search requests in decreasing order of the correlations between documents and requests;

(e) evaluation procedures are available which are specifically designed to be used with the type of automatic retrieval system that furnishes ranked document output.

An initial version of the SMART system became operational on an IBM 7094 during the summer of 1964. In the next few paragraphs the research undertaken with this system is summarized briefly, and future plans are outlined in Part 3.

2. The SMART Project (1964-1965)

The general characteristics of the SMART system have been described in some detail in a number of previous publications.[1,2] A set of dictionaries was initially prepared for the computer literature, and the first experiments were performed with two collections of document abstracts each comprising about 400 abstracts of documents in the computer area. Tests were conducted to determine the effectiveness of a variety of manually constructed intellectual aids, such as synonym dictionaries, hierarchical subject arrangements, dictionaries of concept groups ("phrases"), syntactic analysis procedures, and so on.

The following design features of the SMART system appear to be novel and may be of particular interest:

(a) the synonym dictionary (thesaurus) is stored in tree form, and an efficient look-up process allows very rapid dictionary consultation (of the order of 2 or 3 seconds for texts of several hundred words);[3,4,5]

(b) a sophisticated suffix cut-off process, incorporating a large number of English morphological rules operates in conjunction with the dictionary process to produce the correct word stems for all original text entries;[4]

(c) a complete set of list-processing methods has been programmed (largely in machine language for reasons of program efficiency) to permit a variety of tree tracing options using the concept hierarchy;[6,7]

(d) word grouping procedures permit the construction of so-called "statistical phrases," consisting of pairs or triples of concepts which appear to be related;[8,9]

(e) a tree matching procedure, coupled with an automatic syntactic analysis process, permits the recognition of a large variety of semantically equivalent, but syntactically quite different constructions;[10,11,12]

(f) the system operations are organized in such a way that a programmed supervisory system is used to call upon the proper subroutines as required at any given instant; this makes it possible to update the system at any time by deleting obsolete routines and adding new ones.[13,14,15]

The programmed system is thus based on the manipulation of a variety of different data structures, including linear word strings, tree and graph structures, and reasonably complex vector and matrix operations. The programs themselves are largely written in machine language (FAP), since no existing programming system could accomodate the variety of required data structures.[16,17]

Work in systems evaluation has proceeded along several lines of research. New recall-like and precision-like evaluation measures have been derived, which are especially adapted to the evaluation of systems producing ranked document output as a result of the retrieval operations.[18] These measures have then been applied to the evaluation of automatic search and analysis techniques, using the abstract collections previously referred to. Results obtained with these relatively small collections indicate that synonym recognition is useful in improving retrieval performance; that hierarchical expansions are often effective in broadening or narrowing the coverage of search requests; that phrase procedures using combinations of concepts are far more powerful than procedures based on individual concepts; that weighted concepts are more discriminating than concepts restricted to weights of 0 or 1; that analysis methods which use complete document abstracts are more effective than those based on document titles only, and so on.[19,20]

Procedures have also been implemented to simulate the type of iterative search process, using several differ nt search procedures, which may become common in a real-time system where customers obtain direct access to the equipment.[21,22] Preliminary results indicate that in most cases, an iterative search process based on user feedback is much more useful than one-shot procedures.[19,21,23,24]

The first operating version of the SMART system suffers from a variety of size restrictions, notably concerning the total number of documents which can be processed (less than 500), the length of each document (less than 1000 words), the number of concepts included in the synonym dictionary (less than 1000), and so on. These restrictions are a result of the original progr m organization which is based on internal operations (with dictionaries and full documents stored entirely in core storage) to gain a considerable speed advantage.

In order to permit the SMART system to serve also for the manipulation of document collections of more realistic size, an "extended" SMART system has been designed during the first half of 1965, as described in Sections II to XVII of Report ISR-9. This new system should permit operations under more life-like conditions, and should make it possible also to apply the evaluation procedures to collections of practical interest such as the NASA, DDC, and MEDLARS systems.

The new system can accomodate up to 2^{18} (about 262,000) documents; the number of concepts is effectively unrestricted (up to 2^{18}); the size of the hierarchy is unrestricted (except that the branching ratio of each individual node is assumed not larger than 20); the length of each document is also unrestricted for all practical purposes (a limit exists only because of the limitation on the total number of concepts which may be generated). It may be expected that this extended SMART system will prove useful immediately in extending the retrieval experiments to new search methods and to more interesting operating environments.

3. Future Projects

The extended system makes it convenient for the first time to use a variety of statistical word association methods for the analysis and retrieval of information. These statistical procedures are based on term-term, and document-document correlation matrices which could not, because of their large size, be processed in the original system.

The following projects are under consideration in connection with the extended SMART system:

(a) experiments using larger document collections and also different subject fields (arrangements have been made in this connection with the Aslib-Cranfield project for the transfer of a collection of 1200 hand-indexed document abstracts in aeronautical engineering together with the necessary subject dictionaries; experiments with this collection will make it possible to compare the efficacy in retrieval of the manually assigned index terms with that of the automatic text manipulations incorporated into SMART);

(b) experiments using new processing methods, including statistical term and document associations, bibliographic coupling procedures, and author and journal images;

(c) experiments in a real-time environment with actual user populations, assuming the availability of console input and a time-sharing computer organization;

(d) experiments using combinations of analysis methods, and the iterative search techniques previously developed, to test the effectiveness of multiple searches compared to the single search operations now used with all large, operating systems;

(e) experiments using documents in other languages, such as French, together with translated versions of the dictionaries, to determine whether multilingual collections can be processed;

(f) experiments using different sections of each document for retrieval purposes, such as section titles, figure captions, summaries, reviews, and so on.

In addition, the efforts on systems evaluation, up to now largely confined to the generation and computation of various recall and precision measurements, may be expected to be continued, and to be extended also to cost and other parameters, if more realistic operating conditions can be achieved.

REFERENCES

1. G. Salton, "A Document Retrieval System for Man-machine Interaction," Proceedings of the ACM 19th National Conference, Philadelphia (1964).

2. G. Salton and M. E. Lesk, "The SMART Automatic Retrieval System - An Illustration," Communications of the ACM Vol. 8, No. 6 (June 1965).

3. E. H. Sussenguth, Jr., "The Use of Tree Structures for Processing Files," Communications of the ACM, Vol. 6, No. 5 (May 1963).

4. M. Cane, "Dictionary Look-up and Updating Procedures," Report ISR-7, Section IV, Harvard Computation Laboratory (June 1964).

5. M. Cane, "Dictionary Look-up and Set-up Procedures," Report ISR-9 Sections V and VI, Harvard Computation Laboratory.

6. G. Shapiro, "Processing of the Concept Hierarchy," Report ISR-7, Section V, Harvard Computation Laboratory (June 1964).

7. G. Shapiro and M. Razar, "Hierarchy Set-up and Expansion Procedures," Report ISR-9, Section XV, Harvard Computation Laboratory.

8. M. Lesk and T. Evslin, "Statistical Phrase Processing," Report ISR-7, Section IX, Harvard Computation Laboratory (June 1964).

9. G. Shapiro, "The Statistical Phrase Program," Report ISR-9, Section VII, Harvard Computation Laboratory.

10. E. H. Sussenguth, Jr., "The Sentence Matching Program - GRAPH," Report ISR-7, Section VII, Harvard Computation Laboratory (June 1964).

11. G. Salton, "Automatic Phrase Matching," Report ISR-8, Section V, Harvard Computation Laboratory (December 1964); also presented at the International Conference on Computational Linguistics, New York (May 1965).

12. G. Salton and E. H. Sussenguth, Jr., "Some Flexible Information Retrieval Systems Using Structure Matching Procedures," Proceedings of the AFIPS Spring Joint Computer Conference, Washington (April 1964).

13. M. Lesk, "The SMART System - General Program Description," Report ISR-7, Section II, Harvard Computation Laboratory (June 1964).

14. M. Lesk, "The SMART Automatic Text Processing and Document Retrieval System," Report ISR-8, Section II, Harvard Computation Laboratory (December 1964).

15. T. Evslin, "A General Discussion of the SMART System," Report ISR-9, Section II, Harvard Computation Laboratory.

16. G. Salton, "The Representation of Structured Information," Proceedings of the IFIP Congress - 65, Spartan Books, New York (1965).

17. G. Salton, "Data Manipulation and Programming Problems in Automatic Information Retrieval," ACM Working Conference on Pragmatics and Programming Languages, San Dimas, August 1965.

18. J. J. Rocchio, Jr., "Performance Indices for Document Retrieval Systems," Report No. ISR-8, Section III, Harvard Computation Laboratory (December 1964).

19. G. Salton, "The Evaluation of Automatic Retrieval Procedures - Selected Test Results Using the SMART System," Report ISR-8, Section IV (December 1964); also to be published in American Documentation (1965).

20. C. Harris, "Analysis of Student Requests," Report ISR-9, Section XX, Harvard Computation Laboratory.

21. J. J. Rocchio, Jr. and G. Salton, "Automatic Search Optimization and Iterative Retrieval Techniques," Report ISR-9, Section XXIV, Harvard Computation Laboratory.

22. J. J. Rocchio, "Relevance Feedback in Information Retrieval," Report ISR-9, Section XXIII, Harvard Computation Laboratory.

23. M. Lesk, "A Program to Evaluate the Iterative Search," Report ISR-9, Section XVIII, Harvard Computation Laboratory.

24. J. J. Rocchio, "The Merged Output Results," Report ISR-9, Section XIX, Harvard Computation Laboratory.