I. The SMART System -- Retrieval Results and Future Plans

G. Salton

1. Introduction

The SMART system is a fully-automatic document retrieval system, capable of processing on a 7094 computer search requests and documents available in English, and of retrieving those documents most nearly similar to the corresponding queries. The machine programs, consisting of approximately 150,000 program steps, can be used not only for language analysis and retrieval, but also for the evaluation of search effectiveness by processing each search request in several different ways while comparing the results obtained in each case.

The initial emphasis in the experimental runs performed with the SMART system was placed on the use of a large number of fully automatic language analysis procedures, including dictionary look-up as well as statistical and syntactic methods, and on the evaluation of the relative effectiveness of each procedure for indexing and search purposes. At the time of this writing, extensive experiments have been performed with four document collections in three subject areas : documentation, computer science, and aerodynamics. Notwithstanding the apparent diversity in the subject matter treated, the search results were found to be basically the same in each of the three areas, in the sense that procedures which appear to operate well in one area also exhibit a superior performance in the others. Furthermore, a comparison of the automatic text analysis methods with the standard manual keyword search process shows that many automatic procedures are fully as effective in retrieving useful materials and in rejecting useless ones as are the better known manual procedures.[1]

Since an information system, whether manual or automatic, may be expected to service a large variety of customers, each of whom may have different needs and different background, it is unreasonable to suggest that a single search of some part of the collection would prove equally useful for all customers at all times. Accordingly, more emphasis has been placed in the recent past on search experiments using storage organizations and search strategies which make it possible for the user to influence the search results by submitting to the system appropriate feedback information. A given search is then undertaken iteratively by processing the same search request several times, while altering the search conditions for each iteration. Such iterative retrieval techniques are particularly well adapted to automatic time-sharing equipment where customers can communicate directly with the system by means of suitable input-output equipment.[2,3]

Many different user feedback strategies have been considered experimentally [4], as well as a variety of search strategies. Some search strategies, based on the construction of groups of related documents, and groups of related search requests seem particularly promising, since they make it possible to obtain effective retrieval performance by comparing a given search request against only a small number of selected documents, instead of performing a full search of the collection.[5,6]

The procedure making use of document groups, or clusters, is based on the identification of certain document subsets similar in some sense to the given request. The search is then limited to only those documents included in the previously identified subsets. The query clustering process, on the other hand, depends on the accumulation of groups of requests previously processed through the system. In that case, the search strategy for a given query can be made to depend on the strategies previously found useful for similar types of queries. In either case, only a small portion of the collection is actually involved in the search process, and the actual loss in search effectiveness, compared with a full search is found to be small.

In the next part, a few of the principal evaluation results obtained with the SMART system are summarized, and some of the future research plans are discussed in part three of this section.

2. Experimental Results

The initial experiments conducted with the SMART system were specifically designed to answer certain fundamental questions concerning the design of information systems : can automatic text processing methods be used effectively to replace a manual content analysis; if so, what parts of the documents are most appropriate for incorporation into the analysis; is it necessary to provide vocabulary normalization methods to eliminate linguistic ambiguities; should such normalization be handled by means of specially constructed dictionaries, or is it possible to replace thesauruses by statistical word association methods; what dictionaries can be used most effectively for vocabulary normalization; is it important to provide hierarchical subject arrangements, as is done in library classification systems; alternatively, should syntactical relations between subject identifiers be preserved; does the user have an important role to fulfill in controlling the search procedure.

These and many other questions are answered by the following rules derived from the evaluation results, and described in greater detail in the remainder of this report.[1,4,6] In each case the evaluation is made in terms of two measures, known as recall and precision, which reflect, respectively, the ability of the system to retrieve wanted material, and its ability to reject nonwanted items:

1) The use of document titles alone for purposes of information analysis results in poor retrieval performance compared with the use of abstracts or full text.

2) The use of information identifiers which are weighted in accordance with their presumed importance leads to large-scale improvements in retrieval effectiveness, compared with the use of unweighted terms.

3) Dictionaries providing synonym recognition are of considerable help in improving retrieval performance, particularly when they reflect the properties of the vocabulary under consideration.

4) Absolute accuracy in the analysis of every single item is not so important as the accumulation of a maximum number of correctly analyzed items. If a choice exists between a method which can produce one guaranteed correct content iidication (syntactic analysis), and another which produces five indicators of which four are probably correct (statistical phrase process), the second is generally to be preferred.

5) Simple phrase generation methods lead to a definite improvement in recall at the expense of some initial loss in precision in the low recall region.

6) Deep indexing procedures which supply new information identifiers of which some are useful but many are not usually improve recall but depress precision.

7) Statistical concept-concept associations can be used to improve recall performance particularly for collections for which a well ordered synonym dictionary does not exist.

8) Keyword matching systems based on manually assigned index terms are found (at least for one well-known document collection in aerodynamics) to be not substantially superior to raw word matching techniques, and to be actually inferior to statistical word associations and to thesaurus methods.

9) Iterative search techniques, based on feedback information supplied by the user as a result of previous retrieval procedures, appear to offer major promise for more effective search operations.

If these results are accepted as generally valid, one must conclude that future information centers will probably not be based on manual subject indexing, but will make use of some form of automatic text analysis. Among the techniques likely to be implemented in practice are synonym recognition and phrase generation methods made possible by the construction of suitable thesauruses and phrase dictionaries, and statistical term-term association procedures. Document identifiers may be expected to be based on document abstracts, or longer document excerpts, and weights will be assigned to improve retrieval performance. A variety of additional techniques, including hierarchical subject expansions and automatic syntactic analyses may be used under special circumstances, but their general applicability is still unproved.

3. Discussion and Future Plans

In discussing the evaluation results previously outlined, it is important to consider the context within which these results were obtained before their general validity is accepted. It is,in fact, possible to argue that the results are completely invalid because in many cases no real user need existed when the requests were formulated; because the searches were conducted in an artificial environment rather than within an operational system; because the collection sizes used were in all cases very small, consisting of less than 1000 documents for each collection; because the dictionaries used to perform the word normalization were in some cases not constructed independently of the collections; because some of the relevance judgments used to compute recall and precision may be suspect since they were not always generated by actual users of the system; because the original manual indexing available for the aerodynamics collection may not have been performed under ideal conditions; and because in a situation in which it is impossible to alter one given variable without also affecting many others, it is difficult to make positive statements whose general validity is unchallengeable.

In fact, the situation is not nearly so complicated as these objections appear to indicate. Most of the searches in fact exhibited a quite consistent behavior over a large range of experiments involving many changes of variables.

Thus concept or synonym dictionaries were constructed for three subject fields in several different ways, and dictionaries constructed from one sample collection were used on a different new collection with substantially similar results : synonym recognition was always found to be superior to raw word stem matches.

Relevance judgments, evaluating the usefulness of documents with respect to search requests were made variously by project members, by university students drafted for the purpose, and in the case of the Cranfield aeronautics collection by scientists and experts in the field. The same is true for the original request formulations. The output results obtained under all these different conditions were, however, substantially similar between different methods. Unquestionably, some of the relevance judgments used were incorrect, but if they were incorrect for one method, they were similarly faulty for the others, and the bias, if any, seemed to operate in the same direction in each case. Furthermore, the Cranfield relevance judgments, made by scientists under carefully controlled conditions, are subject to exactly the same challenges, as those made by students and staff.

The hand-indexing available for the Cranfield aeronautics collection was made by two or three trained indexers with some help from subject experts. Since the collection size was small an unusual degree of consistency would seem to have been maintained; furthermore, the degree of indexing was unusually deep, consisting of an average of over thirty terms for each document. If that indexing is not typical, then surely it is because normal keyword indexing cannot proceed under the same controlled conditions for large collections, and the search results for larger collections may be expected to exhibit an even clearer advantage for the automatic procedures.

Still, when all is said and done, it is clear that some of the aforementioned objections can only be stilled by operating with larger than token collections, and hopefully by tying the experimental system into a real user environment. The following SMART experiments are therefore planned for the future :

1) experiments with larger document collections, both hand-indexed, and unindexed;

2) experiments in different subject areas, possibly including social science topic areas, and news articles, rather than only physical science material;

3) experiments in a real-user environment in which people with actual need propose the search queries, and make relevance judgments;

4) experiments with iterative search techniques in which user feedback information is used to conduct improved searches;

5) experiments with multi-level searches for which search efficiency is maintained even though only a small part of a given collection is actually searched;

6) experiments with storage organizations using document and request groupings to optimize search efficiency;

7) heuristic search strategies previously found useful to perform new required searches under similar conditions;

8) real-time search experiments in which users communicate directly with the system, under operational conditions.

It is not expected that the basic evaluation results already obtained will be substantially affected by these new environments; however, additional information will be gained, particularly about operational conditions, which will hopefully be useful in improving the design of actual automatic information systems.

References

[1] M. E. Lesk and G. Salton, Design Criteria for Automatic Information Systems, Report No. ISR-11 to the National Science Foundation, Section V, June 1966.

[2] J. J. Rocchio, Document Retrieval Systems -- Optimization and Evaluation, Harvard University Doctoral Thesis, Report No. ISR-10 to the National Science Foundation, March 1966.

[3] J. J. Rocchio and G. Salton, Information Search Optimization and Iterative Retrieval Techniques, AFIPS Conference Proceedings, Vol.27, Part 1, Spartan Books, Washington, D. C., 1965.

[4] W. Riddle, T. Horwitz, and R. Dietz, Relevance Feedback in Information Retrieval Systems, Report No. ISR-11 to the National Science Foundation, Section VI, June 1966.

[5] J. D. Broffitt, H. L. Morgan, and J. V. Soden, On Some Clustering Techniques for Automatic Information Retrieval, Report No. ISR-11 to the National Science Foundation, Section IX, June 1966.

[6] V. R. Lesser, A Modified Two-Level Search Algorithm Using Request Clustering, Report No. ISR-11 to the National Science Foundation, Section VII, June 1966.