I. THE SMART SYSTEM - AN INTRODUCTION

Gerard Salton

1. Introduction

The first eleven sections of the present report are devoted to a detailed description of the SMART document retrieval system.† This system is designed to process English texts and search requests, and to identify those items of information which are similar in some sense to a stated query. The original texts and search requests are analyzed entirely by machine using a variety of statistical, syntactic, and semantic procedures. No fixed processing sequence is provided; rather it is possible to use a large number of different methods both for the analysis of information and for the comparison of stored data with search requests. This organization makes it possible to reprocess a given search request and a given document collection many times, while making adjustments in the processing methods. The system is therefore useful in simulating a practical retrieval situation where the user can interact with the system, and can obtain an adequate response to each request by successively adjusting the search specifications. Furthermore, by comparing the results obtained under different processing conditions, the system helps in evaluating the efficacy of the various available procedures for the analysis and matching of stored information.

---

†An overall description of the SMART system is included in a previous report in this series.[1]

In the present section the main characteristics of the SMART system are briefly reviewed, and an attempt is made to provide a guide for perusal of the remaining part of this report.

2. Characteristics of the SMART System

Within the past few years a great deal of effort has been devoted to the generation of methods for information analysis and to the design of experimental automatic information retrieval systems. Procedures have, in particular, been developed for the production of automatic indexes and simple abstracts of texts, for the automatic syntactic analysis of sentences in the natural language, for the automatic generation of simple answers to restricted types of queries, for the automatic comparison of user interest profiles with information identifications, and for the automatic construction of many types of listings in many different formats.

As a result of these multiple activities, there now exist a large variety of different methods to perform a given information processing task, and it is difficult to decide in any given situation which of the many possible approaches is most usefully taken. The problem is compounded by the fact that an abstract model to aid in the determination of the necessary and sufficient characteristics of a good retrieval system is not readily available. Furthermore, results obtained from most present experimental systems are in general applicable only to special situations and restricted types of data.

The retrieval system (SMART) described in the present study is designed to overcome the limitations of other experimental systems, not only because it includes many more different features than other systems, but also because its design is flexible enough to permit changes and adjustments to be made as new information becomes available, or as new tests are to be performed. The system, in fact, makes it possible to process the same body of data in many different modes by calling into play different methods for the determination of information content, different criteria for matching items of stored information, and different ways for specifying the information requests.

This flexibility is achieved by designing the system around a central supervisor, called CHIEF, which decodes the input instructions and calls on the various subroutines as needed. The supervisor performs three principal functions: during the initialization phase, the supervisor reads from the input tape the list of processing options and the parameters which affect the current run, and performs the necessary decoding; during the second stage, the current document collection, consisting of a mixture of English texts in BCD coding and of preprocessed documents in binary, is read in and processed as required by the active processing options; during the last stage, the retrieval operations proper are performed.

At the present time, nine basic input operations are recognized, and the processing is controlled by 35 processing options and by five variables affecting the value of various parameters. The processing options fall into four basic categories: alphabetic dictionary

procedures, operations using the semantic concept hierarchy, syntactic procedures using a phrase dictionary and structural matching methods, and statistical association operations based on co-occurrence characteristics of terms within a collection of documents or within the sentences of a given document. In addition, there exist a variety of general processing methods, and extensive dictionary updating routines.

Five basic dictionaries or tables are used by the system: an alphabetic stem dictionary designed to supply each word stem with a number of syntactic and semantic codes, an alphabetic suffix table to obtain syntactic codes for word suffixes, a numeric concept hierarchy to represent various hierarchical relations between semantic categories, a dictionary of "criterion" phrases exhibiting syntactic relations between the components of a phrase, and a dictionary of "statistical" phrases in which the components are related by co-occurrence characteristics within the sentences of documents.

The processing is presently restricted to written texts of 1000 words or less, and the available dictionaries include mainly terms used in the computer literature. These limitations are not, however, of a fundamental nature and may be removed in future versions. Future extensions may also include the use of bibliographic citation data for purposes of content analysis, and an adaptation of the system to time-sharing operations in which a number of requests are processed simultaneously and a conversational system is built-in between user and equipment.

The SMART system as presently available on the IBM 7094 computer is described in detail in the remaining sections of this report. In the next few paragraphs, the main features incorporated into the system are briefly mentioned, and extensive references are made to the more complete descriptions appearing elsewhere in this report.

3. SMART Documentation and Principal Systems Features

The organization of the systems documentation consisting of Secs. II to XI of the present report is summarized in Table 1. The discussion follows a logical sequence rather than the actual processing sequence used in SMART. Thus, the updating procedures for the various dictionaries are described together with the lookup and other operations affecting each given dictionary, even though, in fact, updating operations for all dictionaries are performed in a single updating run, ahead of any of the other processes.

The overall systems description is contained in Secs. I and II; alphabetical stem and suffix dictionary procedures are described in Secs. III and IV; operations affecting the hierarchical concept dictionary are similarly covered in Secs. III and V. The syntactical processing is discussed in Secs. VI, VII, and VIII, and statistical methodology is covered by Secs. IX and X. Finally, the constructions of the "vacuous" dictionary and various housekeeping tasks are included in Sec. XI.

Section II is of primary interest, since it covers the main systems organization, as well as the current program limitations and

Process Type	Description	Section No.
Main System	Main systems characteristics, Organization of supervisor, Input and operating instructions, Operating characteristics.	I,II
Stem and Suffix Dictionary	Stem and suffix dictionary lookup, Morphological rules affecting suffix cut-off, Updating of stem and suffix dictionaries.	III,IV
Concept Hierarchy	Construction and updating of concept hierarchy, List processing methods for processing of hierarchical tree structure.	III,V
Syntactic Procedures	Editing system for syntactic procedures, Construction and updating of criteria tree file, Structural matching.	VI,VII,VIII
Statistical Procedures	Detection of statistical phrases, Term and document correlation procedures, Term and document clustering, Request processing.	IX,X
Housekeeping	General housekeeping routines, Construction and updating of "vacuous" dictionary	XI

Summary of SMART Documentations

TABLE 1

operating characteristics. Readers interested in the programming aspects, the linkage organization, and the operations of the supervisory system should become familiar with the material covered in this section. Of particular importance is a complete set of systems flowcharts included in Sec. II.

The collection of abstracts used to perform the current computer experiments is described in Sec. III. Also covered is the general organization of the alphabetical stem and suffix dictionaries, of the concept hierarchy, and of the vacuous stem dictionary. (The automatic procedures connected with these dictionaries are described in Secs. IV, V and XI, respectively.) Problems pertaining to the assignment of semantic and syntactic codes to dictionary entries are also included in Sec. III.

The detailed operations of the alphabetical stem and suffix dictionaries are described in Sec. IV. Specifically outlined are the internal (tree) formats used to store word stems and suffixes, and the procedures used to separate text words into stems and suffixes before performing the dictionary lookup operations proper. Provision is also made for addition and deletion of dictionary entries by means of a complete set of updating programs. The dictionary lookup itself is based on an attempt to find the longest match between a given argument and a dictionary entry; if a false stem is detected, a backtracking routine is used to locate the last previous acceptable stem. Stems are detected by a left-to-right, letter-by-letter scan; suffixes are found either by left-to-right or right-to-left scans, depending on

whether the corresponding stem was previously isolated or not. Elaborate precautions are taken in the suffix splitting routine to account for morphological changes, such as deletion of final "e," changes from "y" to "i," and doubling of final consonants. These procedures are completely described in Sec. IV.

The operations dealing with the concept hierarchy are covered in Sec. V. The hierarchy is stored as a list structure, and each concept number is provided with four associated link addresses pointing, respectively, to the parent on the next higher level within the hierarchy, to one of the sons on the next lower level, to one of the brothers on the same level, and to a possible cross-referenced concept. The methods permitting to gain access to the hierarchy as well as the chaining procedures to be used to traverse the hierarchy are detailed in Sec. V. A complete set of hierarchy updating programs is also provided. The addition of new concept numbers is performed in a relatively straightforward manner; deleted concepts cannot, however, be simply eliminated, since some of the attached nodes might then be left "hanging." Under the present operating procedure, tree nodes corresponding to deleted concepts are marked, and the sons of such deleted nodes are added to the set of brothers of their former parent.

The overall organization of the syntactic sentence matching procedure is described in Sec. VI. Two main programs are described in this section: the first is an editing program designed to convert the binary data produced by the dictionary lookup process into the binary

phrase processing, very similar in spirit to the criterion tree routine, except that the components of a statistical phrase are related only in that they co-occur within the same sentences of the documents, instead of exhibiting also syntactic relations. A table of statistical phrases is used as before to assign semantic markers to the matching sentences extracted from the documents. Incorporation of the procedures described in Secs. VI and IX respectively, make possible a comparison between phrase processing based on co-occurring words and phrase processing taking into account syntactic relationships.

Section X contains a description of the principal statistical procedures. Two basic inputs are used: a set of concept-sentence matrices in which each matrix element represents the frequency of occurrence of a given concept in a given sentence of a document, and a concept-document matrix consisting of the occurrence frequencies of concepts within the documents of a collection. Using these matrices, it is possible to compute similarity coefficients between concepts, based on co-occurrence characteristics within the same sentences of a document, or within the same documents of a collection; similarly, correlation coefficients can be computed for documents, based on the number of terms jointly assigned to them. The procedures used to compute term-term and document-document correlations, and term and document clusters, are described in Sec. X. The comparison of search requests with stored document identifications is also covered in that section.

A variety of housekeeping routines, largely internal to the SMART system, are described in Sec. XI. Also included is a description of the procedures used to construct the "vacuous" concept dictionary. This dictionary is used to assign dummy concept numbers to the text words, a new concept number being generated for each occurrence of a new word stem.

The documentation of the SMART system included in the present report should be sufficiently complete to enable an interested user to prepare the necessary input decks and to use the system on any 7094 installation with the appropriate tape complement and monitoring systems. The extensive flowcharts included in each section should be particularly useful in gaining an understanding of the system. A more intensive study of the actual programs will be necessary in order to effect internal programming changes.

REFERENCES

1. Salton, G., "A Flexible Automatic System for the Organization,Storage, and Retrieval of Language Data (SMART)," Information Storage and Retrieval, Report ISR-5 to the National Science Foundation, Sec. I, the Computation Laboratory of Harvard University (January 1964).

2. Kuno, S., "The Multiple-path Syntactic Analyzer for English," Mathematical Linguistics and Automatic Translation, Report No. NSF-9 to the National Science Foundation, the Computation Laboratory of Harvard University (May 1963).

3. Sussenguth, E. H., Jr., "Structure Matching in Information Processing," Information Storage and Retrieval, Report No. ISR-6 to the National Science Foundation, the Computation Laboratory of Harvard University (April 1964).