APPENDIX A

THE SMART SYSTEM

The SMART automatic document retrieval system currently running on the IBM 7094 digital computer at Harvard University was used both as a simulation environment and data base generator for the experimental results presented in this thesis. As the SMART system has been thoroughly documented (references 1-3), only a brief summary of its main features is outlined here.

A. Content Analysis Techniques

The indexing function of the SMART system is capable of incorporating a number of automatic content analysis techniques. Documents are entered into the system in the natural language (with a minimal number of keypunching conventions) and passed through a dictionary lookup phase. The lookup operates with a stem-suffix splitting algorithm (which incorporates spelling rules), and word stems are matched against entries of a stored dictionary. A variety of dictionaries may be used in the system ranging from a simple one to one encoding (keyword dictionary) to a dictionary which produces a many to many thesaurus-type mapping. In addition to providing a semantic encoding for the detected stems, the lookup process has provisions for providing syntactic stem codes based on both the stem and suffix dictionaries. After the initial lookup phase, a coded

sentence by sentence text image is available for additional content analysis. The principal processes available at this stage are phrase identification procedures which may be based on an automatic syntactic analysis or on a simple term-term co-occurrence detection scheme.

At the conclusion of the semantic coding process, the sentence by sentence text image is compressed into a weighted property vector index image. Property weights are derived by a summation over the encoded text image so that the weight of a given component of the vector index image of a document is representative of the frequency of occurrence in the document of the features mapped into that component. To reflect the multiple mappings incorporated in the thesaurus transformation, each input term is mapped with a constant total weight. Thus a term which is encoded into a single thesaurus category contributes a weight w (w is a scale factor equal to 12 in the current system). If the input term maps into k categories, each category receives a contribution of w/k to its final weight. An occurrence of the term "band" of Figure 2.1 (chapter 2), for example, would contribute a weight of 12 to category 30, while an occurrence of the term "carrier" would contribute weights of 6 to concepts 61 and 316. This technique prevents ambiguous terms (terms which map into several categories) from distorting the concept weights of the final index vector. While the property vector is the primary index language of the system, a number of alternative component weighting schemes are possible, including the option of ignoring all frequency derived information (which produces, in effect, a set-represented text image).

Document index images generated by this process may be subjected to a number of additional modifications. A variety of transformations based on a pre-specified hierarchical structuring of the elements of the index language is provided. Alternatively, relations among index terms derived from statistical associations in a given collection may be used for modifying index images. The system, therefore, may provide a variety of representations for input documents based on the initial dictionary lookup and subsequent transformation rules defined on the index language. (Note that the index images used experimentally for this thesis were generated by a lookup using version 2 of the SMART thesaurus with no phrase detection and no additional semantic transformations.)

B. Search Request Formulation

Search requests in the SMART system are introduced directly in the natural language and may be treated exactly as are document texts. Requests, therefore, may be subjected to all or any subset of the content analysis procedures available for document processing. In addition to varying the index image of a search request by the sequence of analysis procedures to which it is subjected, a number of additional query modification procedures[4] (including the relevance feedback technique discussed in chapter 3) are being considered for inclusion into the system.

C. Query-Document Matching

The flexibility provided by the computer allows the SMART

sytem to incorporate a variety of query-document matching functions. In addition to the cosine correlation measure for vector mode images, several correlation measures suitable for set mode index images are provided. Additionally, new index image comparison functions may be easily programmed for experimental purposes.

D. Evaluation

Assuming that relevance judgments are available for test search requests, SMART has the capability of automatically generating a variety of evaluation measures for each retrieval operation, including those discussed in section 4 of chapter 5. Additionally, the system provides sufficient output such that auxiliary programs can produce system evaluations over a number of searches, i.e. parameter averages, average precision vs. average recall plots, etc. Novel evaluation algorithms may, therefore, be applied to a number of searches by construction of computer programs to process such data.

REFERENCES

1. Salton, G., "Information Storage and Retrieval", Reports ISR 7, ISR 8, and ISR 9, The National Science Foundation, Harvard Computation Laboratory, June 1964, Dec. 1964, and Sept. 1965

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

3. Salton, G., "A Document Retrieval System for Man-Machine Interaction", Proceedings of the ACM 19th National Conference, Philadelphia, Pa., 1964

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