DEPARTMENT OF COMPUTER SCIENCE

CORNELL UNIVERSITY

INFORMATION STORAGE AND RETRIEVAL

Scientific Report No. IRS-13

to

The National Science Foundation

Reports on Evaluation Procedures and Results 1965-1967

Ithaca, New York
December 1967

Gerard Salton
Project Director

Department of Computer Science

Cornell University

Ithaca, New York 14850

Scientific Report No. ISR-13

INFORMATION STORAGE AND RETRIEVAL

to

The National Science Foundation

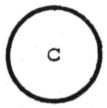
Reports on Evaluation Procedure and Results 1965-1967

Ithaca, New York

January 1968

Gerard Salton

Project Director



REPORTS ON EVALUATION PROCEDURE AND RESULTS 1965-1967

TABLE OF CONTENTS

	Page
SUMMARY	xi
I. KEEN, E. M. "Test Environment"	
1. Introduction	I-1
2. Document Collections and Search Requests	I-1
3. Relevance Decisions	I-5
4. Text Experiments	I-11
A) Experimental Procedures	I-11
B) Variables Tested	I-13
C) Vocabularies and Index Language Devices	I-17
5. Relevance Grade Test Results	I-24
6. Request and Collection Comparisons	I-28
A) Request Preparation	I-28
B) Specific and General Requests	I-31
C) Collection Comparisons	I-41
References	I-48
II. KEEN, E. M. "Evaluation Parameters"	
1. Introduction	II-1
2. Purposes, Viewpoints and Properties of Performance Measures	II-1
3. Measures for Ranking Systems	II-6

Staff of the Department of Computer Science

Cornell University

Kenneth M. Brown
Richard W. Conway
Patrick C. Fischer
Sally Grove
Juris Hartmanis
John E. Hopcroft
Eleanor R. Ide
E. Michael Keen
Joann Newman
Gerard Salton
Gerald Siegel
Roland A. Sweet
Robert J. Walker
Peter Wegner
Burton Weiss
Donna Williamson
Robert E. Williamson

Project Staff in the Division of Engineering and Applied Physics

Harvard University

Jeffrey Bean
Sally Hobbs
Michael Lesk
E. Ricardo Quinones

TABLE OF CONTENTS (continued)

Page

II. continued

A) Single Number Measures II-6

B) Varying Cut-off Performance Curves II-7

C) Comparison of Single Number and Curve Measures . II-13

4. The Construction of Average Precision Versus Recall Curves II-18

A) Averaging Techniques II-18

B) Cut-off Techniques II-21

C) Extrapolation Techniques for Request Generality Variations II-31

D) Extrapolation Techniques for Evaluation of Cluster Searching II-40

5. Measures for Varying Relevance Evaluation . . . II-43

6. Measures for Varying Generality Comparisons . . . II-46

7. Techniques for Dissimilar System Comparisons and Operational Testing II-51

8. The Comparison of Specific and General Requests and the Viewpoints of the "higher precision" and "high recall" user II-53

9. The Presentation of Data as Individual Request Merit . II-63

References II-67

III. KEEN, E. M.

"Search Matching Functions"

1. Introduction III-1

2. Matching Procedures used in Manual, Mechanized and Automated Systems III-1

III. continued

A) Manual Systems III-1
B) Mechanized Systems III-5
C) Automated Systems III-8

3. SMART Test Results — Matching Functions III-9

A) Description of Functions III-9
B) Retrieval Performance Results III-11
C) Analysis of Performance III-17

4. SMART Test Results — Weighting Scheme III-29

A) Description of Weighting Scheme III-29
B) Retrieval Performance Results III-30
C) Analysis of Performance III-32

5. Conclusions and Suggested Further Studies . . . III-48

References III-58

IV. REITSMA, K. AND SAGALYN, J.
"Correlation Measures"

Abstract IV-1

1. Introduction IV-1

2. Weighted versus Logical Description Vectors . . . IV-2

3. The Correlation Coefficients IV-5

A) The Inner Product IV-6
B) The Cosine Coeffieient IV-7
C) The Hypersine Coefficient IV-7
D) The Overlap Coefficient IV-8

TABLE OF CONTENTS (continued)

Page

IV. continued

E) The Maron-Kuhns Coefficient IV-9

F) The Parker-Rhodes-Needham Coefficient . . . IV-11

G) The Stiles Coefficient IV-13

H) The Average Coefficient IV-15

I) The Reitsma-Sagalyn Coefficient IV-16

4. Method of Evaluation IV-17

5. Experimental Results IV-19

6. Discussion IV-22

References IV-26

Tables IV-27

V. KEEN, E. M.

"Document Length"

1. Introduction V-1

2. SMART Test Comparisons V-3

3. Effect of Changes in Document Length V-4

4. Test Results V-13

A) Abstracts versus Titles V-14

B) Abstracts versus Full Text V-28

C) Abstracts versus Indexing V-40

5. Individual Requests and Discussion of Results . . V-48

6. Conclusions V-58

References V-60

Page

VI. KEEN, E. M.

"Suffix Dictionaries"

1. Introduction VI-1
2. Description of Suffix Dictionaries VI-1
3. Retrieval Performance Results VI-4
4. Performance Analyses VI-9
5. Conclusions VI-20

References VI-22

VII. KEEN, E. M.

"Thesaurus, Phrase and Hierarchy Dictionaries"

1. Introduction VII-1
2. Description of Thesaurus Dictionaries VII-1
3. Description of Phrase Dictionaries VII-3
4. Description of Hierarchy Dictionaries VII-8
5. Retrieval Performance Results VII-10
   A) Thesaurus Dictionaries VII-10
   B) Phrase and Hierarchy Dictionaries VII-27
6. Summary of Results VII-37
7. Performance Analyses VII-43
8. Further Studies Required VII-55

References VII-58

TABLE OF CONTENTS (continued)

Page

VIII. DATTOLA, R. T. AND MURRAY, D. M.

"An Experiment in Automatic Thesaurus Construction"

Abstract VIII-1

1. Introduction VIII-1

2. The Construction Algorithm VIII-2

A) Clustering the Document Collection . . . VIII-3
B) Formation of Initial Classes VIII-3
C) Formation of Merged Classes VIII-6
D) Formation of Final Classes VIII-9

3. Evaluation VIII-11

A) Evaluation of the Classes VIII-11
B) Retrieval Evaluation VIII-15

4. Analysis of Results VIII-17

A) Overlap VIII-22
B) Unique Concepts VIII-22
C) Homogeneous Concept Classes VIII-23
D) Dividing Weights VIII-24
E) Cranfield Collection VIII-24
F) Comparison of Other Methods VIII-25

References VIII-26

IX. LESK, M. E.

"Word-Word Associations in Document Retrieval Systems"

1. Introduction IX-1

2. Method IX-1

3. Results IX-5

4. Retrieval Experiments IX-18

TABLE OF CONTENTS (continued)

Page

IX. continued

5. Conclusions IX-50

References IX-52

X. KEEN, E. M.

"An Analysis of the Documentation Requests"

1. Introduction X-1

2. Request Preparation X-1

3. Characteristics of the Requests X-2

A) Length X-2

B) Important Request Words X-3

C) Multiple Need Requests X-3

D) Unclear Requests X-5

E) Difficult Requests X-6

4. Relevance Decisions X-8

5. Request Performance X-9

A) General Performance Analysis Methods X-9

B) Variation in Generality, Length and Concept Frequency X-10

C) Comparison of Requests of the Two Preparers . . X-20

D) The Recognition of Important Request Words . . X-26

6. Performance Effectiveness and Search Procedures . . X-34

References X-41

APPENDIX A

"Recall-Precision Tables" A-1

APPENDIX B

"Original and Modified ADI Queries" B-1

Summary

The present report is the thirteenth in a series covering research in automatic storage and retrieval conducted by the Department of Computer Science at Cornell University with the assistance of the Division of Engineering and Applied Physics at Harvard University. The present report contains a detailed analysis of the retrieval evaluation results obtained with the automatic SMART document retrieval system over the last few years for document collections in the fields of aerodynamics, computer science, and documentation. The various components of fully automatic document retrieval systems are covered in detail, including the form of input (section V), automatic content analysis methods (sections VI, VII, VIII, and IX), and the matching procedures used to compare documents and search requests (sections III and IV). The complete test environment and the parameters which enter into the evaluation process are also described (sections I and II).

Unlike its predecessor (report ISR-12), the present report does not cover the iterative search procedures based on user feedback, or the partial cluster searches designed to speed up the search process, but confines itself to the treatment of the standard fully-automatic analysis and search procedures incorporated into the SMART system, and supplements the summary titled "Computer Evaluation of Indexing and Text Processing", previously published as section III of report ISR-12. Preliminary retrieval results for relevance feedback and cluster searching are contained in section V of report ISR-12; a more definitive treatment of user-controlled

search and retrieval methods follows in subsequent reports in this series.

The analysis of information search procedures and the measurement of retrieval performance are tasks for which a clearly established methodology does not as yet exist. For this reason, it becomes necessary to consider several different procedures, each designed to reveal a different aspect of retrieval evaluation: the total system viewpoint, the viewpoint of the user who insists on high precision, and that of the user who requires high recall. In each case, the aim of the studies included in the present report has been to reach conclusions which may be of practical help to the designer of automatic information systems. The analysis used is thus "insight oriented" rather than "proof-oriented" in the sense that a selective manual analysis of a few typical requests is used in order to gain an understanding of the general process. Formal statistical significance computations of the evaluation results are contained in section III of report ISR-12.

All but three of the ten sections of this report have been prepared by E. M. Keen. Section I is devoted to a detailed examination of the test environment used, including a description of the document collections and search requests, of the manner of obtaining relevance decisions, of the variables entering into the evaluation process, and of the differences in the query and document make-up for the three collections in use.

The evaluation parameters are examined in detail in section II. The viewpoint used in generating the performance measures is first described. This is followed by an introduction of some performance measures which are particularly useful for systems producing a ranked document output in decreasing order of correlation between documents and search requests. Two types of evaluation measures are used including global measures where a

single number serves as an indication of system performance, and continuous recall-precision curves. The construction of the recall-precision curves is described in detail in section II, as are the methods used for producing curves averaged over many search requests. Extensions of the basic evaluation techniques are also discussed to cover cases where variable relevance grades are assigned to documents, and to indicate the problems inherent in a comparison between experimental and operational retrieval systems.

Detailed test results, covering the correlation methods used to compare analyzed documents with analyzed search requests are given in section III. The analysis includes, in particular, a comparison between the "overlap" coefficient which represents a measure proportional to the number of matching terms, and the "cosine" coefficient which takes into account also the total number of terms present in a given document. In each case, the terms are either weighted in accordance with their presumed importance, or unweighted. The conclusion reached is that the cosine correlation used with the weighted content identifiers produces a superior retrieval performance in comparison with the other possible correlation procedures.

Ten additional correlation measures are examined in section IV by K. Reitsma and J. Sagalyn, using the ADI collection for test purposes. The coefficients used include, in particular, the cosine function, the overlap measure, the inner product, the Maron-Kuhns measure, the Parker-Rhodes-Needham measure, and others. Overall, when all recall levels are taken into account, the cosine measure again produces the best results.

The effects of varying the length of the documents to be analyzed automatically, and of using a variety of stored dictionaries in the course of the analysis are examined in detail in sections V, VI, and VII by E. M. Keen. Variations in document length are covered in section V, including a comparison for all collections of the use of titles only with the use abstracts, and of abstracts with full text for the ADI collection. Titles are found to be unsatisfactory for high recall searches with all collections. At the high precision end of the spectrum, good titles are sometimes almost equivalent to poor abstracts. In general, however, abstracts are superior to titles as a source for generating content identifiers. The abstracts are found only slightly inferior to full text for the ADI collection, suggesting that the increased expenditure of entering full text is probably not warranted.

Two types of suffix cut-off procedures designed to reduce language variability by transforming full word paradigms to word stems are evaluated in section VI. The suffix 's' dictionary produces common word forms for items which differ only in a final 's' whereas the word stem dictionary reduces a complete family of related words to a common word stem. The stem dictionary is found to be superior as a retrieval tool to the suffix 's' dictionary for the IRE and ADI collections; the suffix 's' dictionary is slightly superior to the stem dictionary for the Cranfield collection, probably because the more specialized aerodynamics vocabulary provides fewer opportunities for word reduction. The suffix dictionaries offer a convenient mark against which the effectiveness of thesaurus-type dictionaries can be measured.

Thesaurus dictionaries, phrase dictionaries and hierarchical arrangements of terms are described and evaluated for retrieval effectiveness in section VII. A thesaurus is generally used to assemble certain terms into common thesaurus groups according to specified similarity criteria. Terms within the same group can then be reduced to a unique class number, thus providing a certain amount of language normalization. The best thesaurus dictionaries produce an average retrieval performance superior to that provided by the stem dictionaries. For high-precision users, the thesaurus results are not, however, very different from the stem results. Thesaurus construction rules have been devised to insure that a thesaurus is obtained which will, in fact, operate satisfactorily in a retrieval environment, and produce the expected improvements for high recall users.

The results exhibited in section VII for the phrase dictionaries and hierarchical subject arrangements show that the effect of these devices is not as yet sufficiently reliable to warrant their inclusion in operational situations.

Suggestions are also made in section VII for additional retrieval experiments using stored dictionaries, and for the generation of additional language normalization tools.

An experiment in fully-automatic thesaurus construction is described in section VIII by R. T. Dattola and D. M. Murray. The procedure consists in breaking a document collection down into sub-collections, using document-document correlation methods. For each sub-collection, a thesaurus is then constructed using term-term correlation methods. Finally,

a "super-thesaurus" is generated for the whole collection by merging the individual term groupings obtained from the subcollections. Retrieval experiments show that such a fully-automatic super-thesaurus produces better retrieval results than manually constructed thesauruses.

Statistical word-word association procedures are examined and evaluated in section IX by M. E. Lesk. Associative procedures produce groups of terms (or documents) based on the co-occurrence of the terms within the documents of a collection, or within the sentences of a document. The effect is then similar to that of a thesaurus, except that the construction method is automatic. The data included in section IX show that the associative method furnishes results which are essentially independent from those obtained by a normal thesaurus procedure. The associative term groups are unrelated to the thesaurus groups, and there appears to be no basis for the conjecture that second order term associations are equivalent to synonym groupings.

Like the synonym groupings of a thesaurus, word-word associations do occasionally improve the recall of a retrieval system; they also improve the precision by promoting certain relevant documents to higher rank positions.

A detailed analysis of the search requests used with the ADI documentation collection is contained in section X by E. M. Keen. Various characteristics of the search requests are examined, including criteria for identifying unclear request statements, requests expressing multiple needs, requests with identifiable important words, requests with restrictive negative statements, and so on. Using these characterizations, certain

hand modifications are made to the ADI queries and the resulting improvements in retrieval effectiveness are exhibited. Comparisons are also made between certain automatic search procedures and manual searches using a KWIC type index.

A full summarization of all recall and precision tables for all collections and all search experiments is included in appendix A. Appendix B contains a listing of the 35 ADI requests for both original forms and hand-modified versions.

It is hoped that some of the evaluation results presented in this report may lead to the design of more effective fully-automatic retrieval systems in the foreseeable future.

G. Salton