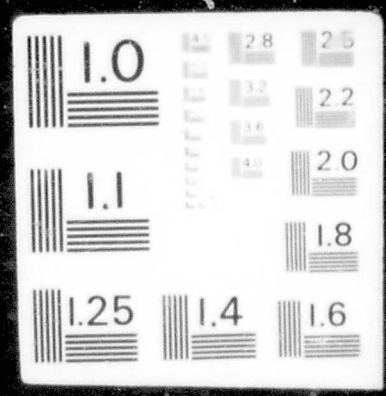
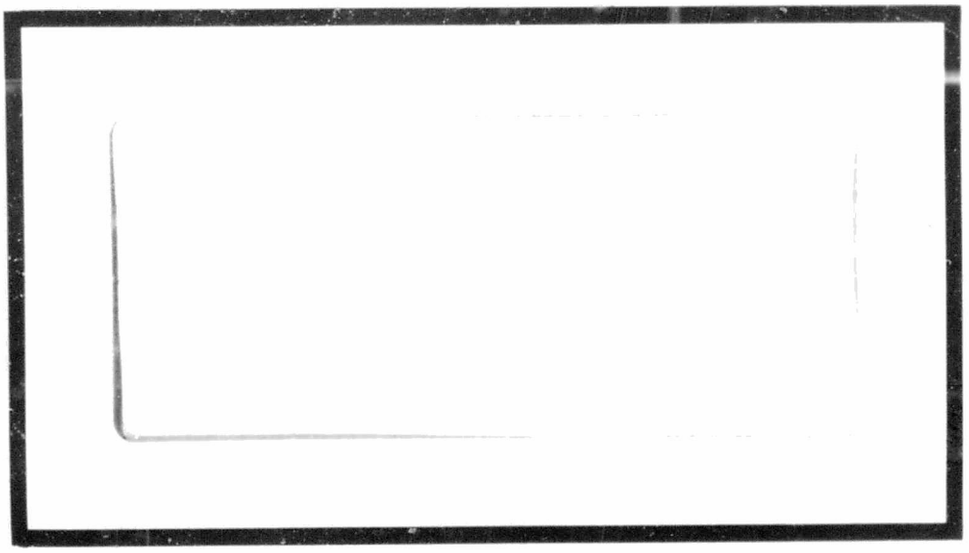




EVALUATION OF DOCUMENT RETRIEVAL SYSTEMS: LITERATURE PERSPECTIVE, MEASUREMENT, TECHNICAL PAPERS

Prepared by

WESTAT RESEARCH, INC.

Under Contract NSF-C 491 for

Office of Science Information Service
National Science Foundation
Washington, D. C. 20550

December 31, 1968

PREFACE

This report presents some extensions of the methodology of document retrieval system evaluation. Part I provides an overview of some of the milepost studies in evaluation. Part II presents in some detail a discussion of measurement from a nonconventional viewpoint. Part III is a collection of technical papers on various aspects of statistical evaluation.

A companion report of this same date, prepared under the same contract, "Procedural Guide for the Evaluation of Document Retrieval Systems," provides suggestions on the implementation of document retrieval evaluations.

Edward C Bryant

Edward C. Bryant
Project Director

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

PART I

THE LITERATURE PERSPECTIVE

by

R. R. V. Wiederkehr

1.1 Introduction

There is now a vast store of published reports and articles dealing with the evaluation of information systems - a store which is growing at a very rapid rate. The rate of growth is indicated by the number of citations in various articles and bibliographies. For example, Henderson's bibliography [1], which incorporates reports up to 1966, cites 324 references of which only 36 were published earlier than 1960, the earliest publication date being 1953. In the Annual Review of Information Science and Technology, the number of references concerned with the Design and Evaluation of Information Systems was 41 for 1965, 116 for 1966, and 201 for 1967.

Since an excellent book by Lancaster [2] and a number of good review articles by Bourne [3], Rees [4], Borko [5], King [6], Treu [7], and Wessel and Cohnssen [8], and bibliographies by Henderson [1], and Neeland [9] already exist, no attempt will be made in this section to cover exhaustively the literature concerned with evaluation of information retrieval systems. Instead, attention will be focused only on selected highlights related to this study.

In reviewing the development of evaluation of information retrieval, it will be convenient to consider two lines of development: the evaluation of existing information systems and the development of measures for evaluation. Whereas the first line of development tends to be experimental, the second tends to be theoretical. These two lines of development will be presented in Sections 2 and 3.

1.2 Evaluation of systems

1.2.1 Early efforts

In 1953, Taube [10] conducted a study to compare the performance of alternate indexing systems. He enumerated several factors which should be included in the evaluation of information retrieval systems: cost, size of equipment, time to organize and search information file, number of access points per item, rate of obsolescence, rate of growth, specificity, suggestiveness, etc.; but confessed that the relative importance of these factors was not known. Using these factors as a basis of comparison, Taube studied the Uniterm system, several classification systems, several subject heading systems, and several standard methods of indexing. He concluded that the Uniterm system was superior with respect to all factors except suggestiveness.

Other early workers concerned with designing and evaluating information retrieval systems were Taube and Heilprin [11], Thorne [12] and Gull [13]. These early efforts were characterized by:

(1) the development of a number of alternate systems, including computers and other mechanized systems, for storing and retrieving information.

(2) construction of mathematical models for measures such as cost, time, number of access points per item.

(3) small-scale tests of alternate systems - mostly alternate indexing systems.

1.2.2 Large scale experiments

These early developments uncovered a number of difficulties and problems which apparently could be resolved by large-scale experimentation. One of the first groups to undertake large-scale experimentation was ASLIB who, under the direction of C. W. Cleverdon, conducted a series of large-scale tests at the College of Aeronautics, Cranfield [14]. The primary objective of this project was to compare the

effectiveness of four indexing and classification systems for a single body of documents or file. The four methods tested were: the Universal Decimal Classification, alphabetical subject headings, Uniterms and a specially-prepared faceted classification. To calculate this effectiveness, two measures were used: the recall (the ratio of the number of relevant documents retrieved to the total number of relevant documents in the file) and the precision (the ratio of the number of relevant documents retrieved to the total number of documents retrieved). Recall was taken to be the primary measure and precision was taken to be the secondary measure. In fact, the concept of the precision ratio (relevance ratio) did not emerge until the project had been underway for some time. The method used to obtain values for these measures was to formulate a number of test questions, each designed to retrieve a single source document, and then to conduct searches based on each of these test questions. The fraction of searches yielding the source document was used to estimate the recall, and the ratio of the number of relevant documents to the total documents retrieved was used to estimate the precision. Other objectives of this project were to determine the effects of type of document indexed, length of time of indexing, qualification of indexer, and to conduct failure analyses to determine reasons for failing to recover source documents. The results indicated that there was surprisingly little difference in the performance of the four systems tested. Human errors in indexing and searching were more serious than errors due to the file organization. It was concluded that file organization is relatively unimportant in the performance of IR systems. Specificity of the vocabulary and exhaustivity of indexing are much more important factors.

A continuation of this line of work compared the performance of a manual index based on a faceted classification with the performance of the mechanized index of metallurgical literature developed by Western Reserve University. The approach used was similar to that of the Cranfield project [15]. This study was perhaps most significant for its

further development of techniques for the analysis of system failures. This failure analysis was later taken much further by Lancaster in the MEDLARS study.

The results of the Cranfield experiments stimulated much discussion and criticism. Swanson [16], for example, in an excellent review of this topic, praised the Cranfield project for collecting an immense quantity of valuable data and producing much well-written material discussing the issues and problems in evaluating retrieval systems, but warned against accepting the Cranfield results as being generally applicable because of a number of conditions under which the results were obtained. For example, the Cranfield results apply only to source documents; they may not apply to non-source documents. Also, there was a lack of control in experiments over the possible influence of human memory. Rees' [17] assessment of the Cranfield project, stated briefly, is that the great value of this work lies in the area of test methodology rather than in the experimental results. In particular, the techniques of failure analysis have proved most helpful in gaining insight on how an information retrieval system functions in practice.

As a result of the experimental work done at Cranfield, Western Reserve University and several other locations, and the ensuing discussions and criticisms, it was realized that a need existed for improved methodologies for testing and evaluation. In 1964 the National Science Foundation sponsored a conference [18] "to review the work on testing and evaluation of document searching systems and procedures and to consider promising directions for future work in this area". The major findings of the conference may be summarized as follows:

(1) A need exists to develop and experiment with measures of performance and criteria for evaluation.

(2) A need exists for emphasis on the experimental design of tests and experiments.

(3) It is desirable that the development of reporting standards facilitate the communication of the results of test and evaluation experiments.

(4) The nature of the notion of "relevance" is fundamental to the evaluation of system performance.

(5) It is feasible and fruitful to consider public or non-individual senses of search specifications based on written versions of them.

(6) The need exists for the development of test and evaluation methods which concentrate on selected features of the document retrieval system, rather than on total systems.

(7) More tests of features of operating systems, designed to determine the advantages and disadvantages of various measures and test methods used, should be conducted.

In response to the need to develop and experiment with measures of performance and criteria for evaluation (Finding 1 of the NSF Conference), two complementary research efforts have been undertaken under the sponsorship of NSF. One study, under the direction of Cuadra and Katter [19], at the Systems Development Corporation, has been concerned with determining the effect of numerous factors, such as judges, documents, etc. on the variability of relevance judgments by conducting a series of fixed-effects experiments. Some conclusions of this study are that relevance judgments are affected by: the skills and attitudes of the particular judges, the documents used, the information requirements statements, the instructions and settings in which the judgments were made, and the type of rating scale used to express the judgments. The other study, under the direction of Rees and Schultz [20], at Case Western Reserve University, has been concerned with determining the effect on relevance judgments of factors such as documents, judgmental groups, research stages, and document representations. A simulated field experimental approach was used. Their approach emphasizes the making

of relevance judgments in the natural setting of the user. Some conclusions of this study are (1) relevance ratings depend on personal characteristics (such as scientific orientation and involvement in research) and should be taken into account in the formulation of information requirements; (2) dichotomous relevance judgments are not very sensitive to the factors affecting these judgments, and (3) stable, meaningful judgments of relevance require that the relevance judges be relatively homogeneous.

In response to the need for emphasis on the experimental design of tests and experiments (Finding 2 of the NSF Conference), a study was undertaken by Snyder, et. al. [21] at Human Sciences Research, Inc., also under the sponsorship of NSF. The major purposes of this study were: to review critically the experimental design practices in previous tests and evaluation studies, to identify poor experimental design practices, and to suggest recommendations for improving experimental design practices. The approach was to consider each study with respect to 15 "review dimensions": study objective, research user perspective, system objective, system stage, research setting, subsystem studied, independent variables, criteria measures, design comparison, control variables, analysis and statistics, measurement sensitivity, sampling, research description and research interpretation, and conclusions. Errors in experimentation which occurred most frequently were identified to be: confounding of independent variables, poorly posed hypotheses concerning the factors that affect the criteria measures (such as recall and precision), unknown variability of the concept of relevance from judge to judge, unknown variability of the relevance rating from judge to judge, the choice of dichotomous relevance scale might be improved, lack of controls over possible sources of variation, not ensuring that samples are representative, selecting too small a sample, employing inadequate statistical analyses, unawareness of related work, and inadequate reporting.

Another product of this study was a list of areas wherein the state-of-the-art requires further development. These areas included:

"development of intermediate criteria, examination of information transmission through the system, investigation of how people make relevance judgments, and examination of concept identification and search strategy in the query chain." They also suggest "that a better understanding of system criteria--criteria needed to evaluate the various elements of the system and interrelationships among them - are prerequisite to advances in the technology of document retrieval systems." Snyder, et. al. outline procedures for employing the present state-of-the-art in experimental method and technique for avoiding the above errors.

Meanwhile, during 1965 and 1966, the ASLIB-Cranfield Project continued through a second phase, called Cranfield II, which is described in a final report [22], [23]. The major objective of this renewed effort was to investigate the effect of factors determining the performance of retrieval systems. Performance in this study is measured in terms of four parameters: recall ratio, precision ratio, fallout ratio, and generality number. Cranfield II was carried out in a laboratory-type setting with the environmental and operational conditions carefully controlled.

Based on the results of this study, Cleverdon concludes that there exists a basic inverse relationship between recall and precision. In conducting a search in a retrieval system, whatever one does to enlarge the search is likely to improve recall but degrade precision. Likewise, whatever one does to reduce the scope of a search strategy is likely to improve precision at the expense of recall. He also suggests that use of index languages involving single terms produced the best performance, index languages based on the EJC thesaurus yielded intermediate performance and index languages based on concepts gave the worst performance. In his review of the Cranfield II report, Rees [24], warns that "acceptance of the Cranfield findings must be tempered by a reasoned scrutiny of the assumptions underlying the work... the difficulty of replicating the Cranfield results impede the investigation of generalizability."

Salton and co-workers have been engaged in evaluating the SMART system, initially at Harvard [25] and more recently at Cornell [26]. The SMART system is a fully automatic document retrieval system which processes both documents and requests without prior manual analysis, i.e., the system automatically analyzes the content of both document and request, performs a match, and produces an ordered output starting with those documents most responsive to the request. Performance is measured in terms of four global parameters (rank recall, log precision, normalized recall, and normalized precision) and ten local parameters (the precision at 10 preselected values of recall). Additional composite measures based on these fourteen are also introduced and used as overall performance measures. For each processing option several searches are performed; and from the output of these searches the fourteen parameters and certain composite measures are evaluated, and used to assess the effectiveness of the particular processing option. The purpose of assessing the various processing options was to generate useful criteria for designing information systems. As a result of these assessments, the following conclusions were drawn:

(1) Weighted subject identifiers are always more effective than weights restricted to 0 and 1.

(2) Full document abstracts are far more effective as a source of content identification than titles alone.

(3) A thesaurus process performs more effectively than methods using original words only.

(4) Fully-automatic text analysis procedures are approximately equivalent in performance to methods based on manually assigned keywords.

(5) Search systems based on a large number of document groups (containing only a few documents) produce better results than systems based on fewer clusters of larger size.

(6) A system based on document transformations produces greater improvements than query transformations alone.

Another experimental effort was conducted by Giuliano and Jones [27] at Arthur D. Little, Inc. The objectives of this study were to evaluate an experimental prototype associative searching system relative to a conventional system, to explore the effectiveness of human mediation in the associative search process, and to test the feasibility of machine identification of content-bearing strings of words for indexing. Variables considered related to characteristics of the user population, the document (message) collection, the indexing scheme, and the search procedure. The major measure of performance of a retrieval system used in this study was the performance curve - a curve of the cumulative value of retrieved documents as a function of rank in the list of retrieved documents ordered by the measure of mismatch between document and search request. For each of a number of search options a small number of searches was performed and the resulting performance curves were constructed and used to evaluate each option. Major conclusions of this effort were: that associative searching is apparently more effective than coordinate searching, that a panel of judges is not significantly more effective than a single judge, and that it is feasible to employ machine identification of content-bearing pairs of words for indexing.

Two major evaluation programs in 1967 involved large operating systems. The first was a comprehensive evaluation of the Foreign Technology Division's Central Information Reference and Control (CIRC) system. Taulbee [28] describes this program in some detail, although results have yet to be released. The National Library of Medicine's system (MEDLARS) was evaluated in considerable depth by Lancaster [29]. The objectives of this study were to determine how effectively and efficiently MEDLARS is meeting the demand search requirements of MEDLARS users, to recognize factors adversely affecting the

system performance and to disclose ways in which user requirements could be satisfied more efficiently and/or economically. Results of exhaustive analysis of search failures led to the conclusions and recommendations concerning user-system interaction; the index language; searching strategies; the indexing process; computer processes; the relationship between indexing, searching and index language; use of foreign language material; and quality control of the MEDLARS operation.

A series of experiments was conducted at the U.S. Patent Office that evaluated systems during early stages of file development. The principal objective of these experiments was to provide some insurance that the completed systems would perform satisfactorily. The experiments involved preliminary search experiments and indexing experiments on samples of documents. Failure analyses were performed to suggest system modifications before they became too expensive.

1.3 Development of Measures for Evaluation of Document Retrieval Systems

In Section 1.2 the major experimental efforts directed towards evaluating existing information retrieval systems occurring over the past 15 years were reviewed. Each such effort employed one or more measures used as criteria for evaluating the existing information retrieval systems. Although the selection of such measures has a strong influence on the outcome of an evaluation, as yet no measure or set of measures has been found that is universally acceptable to information retrieval systems evaluators. Because the issue of selecting a proper measure is a vital one, it is fitting to trace the development of these measures.

In 1956 Perry, Kent and Berry [33] considered a number of quantitative measures to be used as criteria for evaluating and designing information retrieval systems. They concluded that the effectiveness and efficiency of an information retrieval system can be measured in terms of two factors: the recall factor (the ratio of the number of retrieved documents judged to be relevant to the search request to the total number of

relevant documents in the file) and the pertinency factor (the ratio of the number of retrieved documents judged to be relevant to the search request to the total number of documents retrieved). This pair of factors, often under different names, played a dominant role in evaluating document retrieval systems, and have been widely used by many investigators such as Cleverdon, Lancaster, Salton and others. The present generally accepted names for these terms are the recall ratio and the precision ratio.

In 1959 Mooers [34] proposed three measures which extend the notions of recall ratio and precision ratio to a trichotomy instead of a dichotomy. Two of these measures in effect are recall ratios for pertinent and crucially pertinent documents, and one is similar to the complement of the precision.

Composite measures have been proposed by a number of researchers. Swanson [35] relates the recall ratio R to the amount of irrelevant material retrieved, I, and proposes the following measure: $M = R - pI$ where p is a penalty. This measure is a type of net gain. Bornstein [36], Verhoeff et. al. [37] and Wyllys [38] have also proposed composite measures which are reviewed by Swets [39].

Bourne, et. al. [40] described two different criteria for evaluating information retrieval systems:

(1) Performance-requirement matching with weighting, and

(2) Performance evaluation with a time-cost model.

The first method involved identifying factors affecting the relative merit of the information retrieval system, quantifying the performance of the system with respect to each factor, quantifying the user requirements or target performance of each factor based on a sampling of user opinions, observing the deviation of the system from the target value, weighting each factor by its relative importance, and summing the resulting weighted deviations to yield an overall measure of performance. The first method is recommended only as an immediate and rough measure of performance.

The second method employs the measures of annual operating cost, and time required to satisfy users' requests to evaluate the performance of document retrieval systems. Time limitations in this study permitted the development only of a model for costs but not one for time. This second method, which apparently has not yet been highly developed, is considered to be more sound than the first and might be applied to future evaluations.

Meanwhile, another study was undertaken by David Hertz at Arthur Anderson & Co. [41] to develop criteria and measures of effectiveness for evaluating information retrieval systems. The major criteria selected in this study for the evaluation of information retrieval systems were: cost, time and volume; models were developed for each of these measures. Also, a performance simulation model was developed to test alternate types of systems, and to evaluate their performance. The major conclusions of this study were that:

(1) Cost, time and volume are necessary measures for any evaluation.

(2) Performance simulation is an important mechanism for learning about information retrieval systems.

(3) Subjective criteria should be integrated into the final evaluation of a system.

This excellent study apparently has been overlooked by many researchers in the field of evaluation of information retrieval systems.

In 1963 Swets [39] reviewed the previous measures for evaluating information retrieval systems, discussed their shortcomings, and proposed a new improved measure: the operating characteristic curve. The operating characteristic curve of an information retrieval system is a plot of the probability of retrieving the document given that it is relevant (in effect, the recall ratio) versus the probability of retrieving the document given that it is not relevant, as the acceptance criteria for relevance is varied from one extreme (low documents retrieved) to the other (high

documents retrieved). More recently, Swets [42] has constructed the operating characteristic curves from the data generated by: the project at Cranfield, England, under Cleverdon, the project at Harvard and Cornell Universities under Salton, and the project at Arthur D. Little Inc., under Giuliano and Jones. The operating characteristic curves were found to be a convenient means for representing the corresponding information retrieval systems, for when plotted on probability paper, operating characteristic curves are approximately represented by straight lines. Such a straight line can be completely characterized by two parameters: the slope and an appropriate intercept; these two parameters can be used to characterize the effectiveness of an information retrieval system. This characterization has the advantage that it is independent of variations in the acceptance criteria.

In 1968 Cooper [43] reviewed the previous measures including those of Swets, Salton, and Giuliano and Jones, and made the following observation concerning previous measures:

(1) Many previously proposed measures are not single measures (e.g., recall and precision are a pair of measures).

(2) Many previously proposed measures assume there are two sets of documents, a retrieved set and an unretrieved set, without accounting for the possibility of an order of retrieval involving more than two sets.

(3) "Most proposed measures have no built in capability for comparison of system performance with purely random retrieval."

(4) Most proposed measures do not account for how many relevant documents the user actually needs.

Cooper then goes on to introduce the concept of the "expected search length" in a "weak ordering." The following statements present the essence of this concept. If documents in the file are ordered by the retrieval system according to their expected degree of relevance to the search request, and the user quantifies the amount of relevant information desired, say six

relevant documents, then the expected search length is defined to be the number of nonrelevant documents in the ordered file which precede the sixth relevant document. By comparing the expected search length for the actual system with that of a hypothetical system which randomly orders its output documents, the fractional reduction in expected search length in going from the random system to the actual system can be obtained. This fractional reduction is called the mean expected search length reduction factor. Cooper claims that this factor overcomes the shortcomings of earlier factors as enumerated above.

****

An appropriate measure to be used as a criterion for evaluating an information retrieval system should account for both how effectively the objectives are being met as well as how efficiently resources are being used. Consequently, it is desirable to have measures of effectiveness, such as how many useful documents were retrieved, and measures of efficiency, such as the cost and time. Recall and precision only partly satisfy this desire.

In the research and development phase of any system, the primary objective is to demonstrate the technical feasibility of the system. Accordingly, effectiveness is of prime importance and efficiency is often ignored. Once the technical feasibility of the system has been proven, the objective shifts to demonstrating the economic feasibility of the system. In most operating systems economic feasibility is of prime importance, in which case both the effectiveness and the efficiency should be taken into account.

Since most efforts to date concerning the evaluation of information retrieval systems have treated systems in the research and development phase, most of the measures considered have been measures of effectiveness, such as recall and precision. However, as the systems become operational on a large scale, measures of efficiency and overall measures

which account for both effectiveness and efficiency are anticipated. Both Bourne and Hertz (author of the Arthur Anderson study) have recognized this point, but few others.

References

[1] Henderson, Madeline M., (December 1967), "Evaluation of Information Systems: A Selected Bibliography with Informative Abstracts," U. S. Department of Commerce, National Bureau of Standards, Technical Note 297.

[2] Lancaster, F. W. (1968), Information Retrieval Systems: Characteristics, Testing and Evaluation, New York, John Wiley and Sons, Inc.

[3] Bourne, Charles P., (1967), "Evaluation of Indexing Systems," in American Documentation Institute Annual Review of Information Science and Technology, Volume 1, Carlos A. Cuadra, editor, New York, Interscience Publishers, pp. 35-61.

[4] Rees, Alan M., (1967), "Evaluation of Information Systems and Services" in American Documentation Institute Annual Review of Information Science and Technology, Volume 2, Carlos A. Cuadra, editor, New York, Interscience Publishers, pp. 63-86.

[5] Borko, Harold, (1967), "Design of Information Systems and Services," in American Documentation Institute Annual Review of Information Science and Technology, Volume 2, Carlos A. Cuadra, editor, New York, Interscience Publishers, pp. 35-61.

[6] King, Donald W., (1968), "Design and Evaluation of Information Systems," in American Society for Information Science Annual Review of Information Science and Technology, Volume 3, Carlos A. Cuadra, editor, Chicago, Encyclopedia Britannica, Inc. pp. 61-103.

[7] Treu, Siegfried, (1967), "Testing and Evaluation -- Literature Review," in Electronic Handling of Information: Testing and Evaluation, Allen Kent, Orrin E. Taulbee, Jack Belzer and Gordon D. Goldstein, editors, Washington, D. C., Thompson; London, Academic Press, pp. 71-88.

[8] Wessel, C. J. and B. A. Cohnssen, (February, 1967), Criteria for Evaluating the Effectiveness of Library Operations and Services Phase I: Literature Search and State of the Art, Washington, D. C., AD 649 468.

[9] Neeland, Frances, (1966), "A Bibliography on Information Science and Technology for 1966," Parts 1, 2, 3, 4, Santa Monica, California, System Development Corporation.

[10] Taube, Mortimer, (1953), "Evaluation of Information Systems for Report Utilization," in Studies in Coordinate Indexing, Volume I, Washington, D. C., Documentation, Inc., pp. 96-110.

[11] Taube, Mortimer and Laurence B. Heilprin, (August 1957), "The Relation of the Size of the Question to the Work Accomplished by a Storage and Retrieval System," Report No. AFOSR-TN-57-483, Washington, D. C., Documentation, Inc., AD 136 476.

[12] Thorne, R. G., (September 1955), "The Efficiency of Subject Catalogues and the Cost of Information Searches," Journal of Documentation, Volume XI, pp. 130-148.

[13] Gull, C. D., (October 1956), "Seven Years of Work on the Organization of Materials in the Special Library," American Documentation, Volume VII, pp. 320-329.

[14] Cleverdon, Cyril W., (October 1962), "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, PB 162 342.

[15] Aitchison, Jean and Cyril W. Cleverdon, (October 1963), "A Report on a Test of the Index of Metallurgical Literature of Western Reserve University," Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, AD 419 956.

[16] Swanson, Don R., (January 1965), "The Evidence Underlying the Cranfield Results," The Library Quarterly, Volume XXXV, pp. 1-20.

[17] Rees, Alan M., (October 1963), "Review of a Report of the ASLIB-Cranfield Test of the Index of Metallurgical Literature of Western Reserve University," Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research.

[18] National Science Foundation, (February 10, 1965), "Summary of Study Conference on Evaluation of Document Searching Systems and Procedures," Washington, D. C.

[19] Cuadra, Carlos A., Robert V. Katter, Emory H. Holmes and Everett M. Wallace, (June 30, 1967), "Experimental Studies of Relevance Judgments: Final Report," Santa Monica, California, System Development Corporation.

[20] Rees, Alan M. and Douglas G. Schultz, (June 30, 1967), "A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching: Final Report," Cleveland, Ohio, Case Western Reserve University, Center for Documentation and Communication Research.

[21] Snyder, Monroe B., Anne Schumacher, Steven E. Mayer and M. Dean Havron, (January 1966), "Methodology for Test and Evaluation of Document Retrieval Systems: A Critical Review and Recommendations," Report No. HSR-RR 66 16 SK to the National Science Foundation, McLean, Virginia, Human Sciences Research, Inc.

[22] Cleverdon, Cyril W., Jack Mills and Michael Keen, (1966), "Factors Determining the Performance of Indexing Systems," Volume 1, "Design," Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics.

[23] Cleverdon, Cyril W. and Michael Keen, (1966), "Factors Determining the Performance of Indexing Systems," Volume 2, "Test Results," Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics.

[24] Rees, Alan M., (October 1963), "Review of a Report on the ASLIB-Cranfield Test of the Index of Metallurgical Literature of Western Reserve University," Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research.

[25] Salton, Gerard, (December 1964), "The Evaluation of Automatic Retrieval Procedures -- Selected Test Results Using the SMART System," in "Information Storage and Retrieval" (Scientific Report No. ISR-8 to the National Science Foundation), Cambridge, Mass., Harvard University, Computation Laboratory, pp. IV-1-IV-36.

[26] Salton, Gerard, (June 1967), "The SMART Project--Status Report and Plans: Reports on Evaluation, Cluster and Feedback," in "Information Storage and Retrieval" (Scientific Report No. ISR-12 to the National Science Foundation), Ithaca, New York, Cornell University, pp. I-1-I-12.

[27] Giuliano, Vincent and Paul E. Jones, Jr., (August 1966), "Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems," Report No. ESD-TR-66-405, Cambridge, Mass., Arthur D. Little, Inc., AD 642 829

[28] Taulbee, Orrin E., (1967), "An Approach to Comprehensive Evaluation," in Electronic Handling of Information: Testing and Evaluation, Allen Kent, Orrin E. Taulbee, Jack Belzer and Gordon D. Goldstein, editors, Washington, D. C., Thompson; London, Academic Press, pp. 217-229.

[29] Lancaster, F. W., (January 1968), "Evaluation of the MEDLARS Demand Search Service," Bibliographic Services Division, National Library of Medicine, U. S. Department of Health, Education and Welfare, Public Health Service.

[30] King, Donald W., (May 1965), "Evaluation of Coordinate Index Systems During File Development," Journal of Chemical Documentation, Volume 5, No. 96, pp. 96-99.

[31] King, Donald W. and Patricia M. McDonnell, (November 1966). "Evaluation of Coordinate Index Systems During File Development, Part II: Application," Journal of Chemical Documentation, Volume 6, No. 4, pp. 235-239.

[32] King, Donald W. and P. Isakov, (September 1967), "Preliminary Evaluation of the Glass Technology Coordinate Index File," presented at the Seventh Annual Meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices (ICIREPAT), Stockholm, Sweden.

[33] Perry, James W., Allen Kent and Madeline M. Berry, (1956), "Operational Criteria for Designing Information Retrieval Systems," in Machine Literature Searching, New York, Interscience Publishers, Inc., pp. 41-48.

[34] Mooers, Calvin N., (August 1959), "The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval Systems," Report No. RADC-TN-59-160 to U. S. Air Force, Rome Air Development Center, Cambridge, Mass., Zator Company.

[35] Swanson, Don R., (October 21, 1960), "Searching Natural Language Text by Computer," Science, Volume 132, pp. 1099-1104.

[36] Bornstein, Harry, (October 1961), "A Paradigm for a Retrieval Effectiveness Experiment, "American Documentation, Volume XII, pp. 254-259.

[37] Verhoeff, Jacobus, William Goffman and Jack Belzer, (December 1961), "Inefficiency of the Use of Boolean Functions for Information Retrieval," Communications of the ACM, Volume 4, pp. 557-558, 594.

[38] Wyllys, R. E., (1962), "Document Searches and Condensed Representations," paper presented at The First Congress on Information System Sciences, Hot Springs, Va., November 18-21, 1962.

[39] Swets, John A., (July 19, 1963), "Information Retrieval Systems," Science, Volume 141, pp. 245-250.

[40] Bourne, Charles P., G. D. Peterson, B. Lefkowitz and D. Ford, (December 1961), "Requirements, Criteria and Measures of Performance of Information Storage and Retrieval Systems," Final Report to the National Science Foundation on SRI Project 3741, Menlo Park, California, Stanford Research Institute, AD 270 942.

[41] Hertz, David B., (March 1962), "Research Study of Criteria and Procedures for Evaluating Scientific Information Retrieval Systems," Final Report to the National Science Foundation, Contract NSF C218, New York, Arthur Andersen and Co.

[42] Swets, John A., (June 1967), "Effectiveness of Information Retrieval Methods," Report for Air Force Cambridge Research Laboratories, U. S. Air Force, No. AF 19 (628)-5065, Cambridge, Mass., Bolt, Beranek and Newman, AD 656 340.

[43] Cooper, William S., (January 1968), "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," American Documentation, Volume XIX, pp. 30-41.

PART II. MEASUREMENT

by

D. W. King

E. C. Bryant

2.1 Introduction

This report is generally concerned with the evaluation of systems that in some way involve the process by which knowledge is recorded and transferred to the use of others. While the scope of such systems can be very broad, specific examples are frequently called information retrieval systems, document retrieval systems, or library systems. Strictly as a matter of convenience, such systems are referred to in this part of the report as document retrieval systems. There is no intention to limit attention to a specific kind of system by this language. The word "document" may be interpreted very broadly to include published material, research notes, magnetic tapes, microfilms, or any other vehicle for the storage of knowledge (information) that may be transmitted from one individual to another or from an individual in one time frame to the same individual in another.

Document retrieval systems require funding, planning, and management, and these activities require decisions. If evaluation has anything to offer in the field of document retrieval, it must make its contribution to the decision-making process. Such evaluation can range from the purely subjective to the highly objective. It is obvious that measurement is at the heart of objective evaluation, and it constitutes the principal focus of attention in this part of the report. Measurement implies quantification, but in the field of document retrieval there has been no consensus as to what to quantify or how to quantify it. We do not propose here to provide any ultimate answers to questions that to some extent will always remain unanswerable, but we do hope to set forth a general framework for measurement and to make certain recommendations

with respect to the use of measures for specific applications. We do not expect unanimity of agreement concerning these recommendations, but we hope that this report may stimulate further penetrating analysis of some of the problems.

We assume that certain questions must be answered in a discussion of measures of evaluation. Among them are:

1. Why measure at all?
2. What should one measure?
3. What measures are appropriate to specific requirements?
4. How does one use measures?

The remainder of Part II of this report is devoted to a discussion of these questions.

2.2 Why measure at all?

"Measure" is a highly flexible word having, in a typical English dictionary, well over a dozen meanings when used as a noun and perhaps half that many when used as a verb. For our purposes, it seems desirable to take a simple meaning of the verb, i.e., "to take or mark the limits or the dimensions of," and for the noun, "the extent or dimensions of anything." We specifically wish to avoid the connotation of a standard scale, since this is a luxury not permitted us by the systems being measured. In other words, measurement is simply quantification.

We measure (quantify) as an aid to evaluation, but it must be realized that not all evaluation requires quantification. A wine taster can evaluate a wine without knowing its exact chemical composition. However, in the universe of document retrieval systems it is presumed that quantification leads to more consistent, and hence more meaningful, evaluation. To the best of our knowledge this presumption has never been tested, but is seems reasonable on intuitive grounds.

2.3 What should one measure?

Every document retrieval system has virtually an endless list of characteristics that describe the system, the environment within which it operates, and the ways in which it responds to stimuli from that environment. Some of these characteristics are worth measuring and some are not.

Which of the characteristics one decides to measure depends upon the objectives of the evaluation, that is, upon the kinds of decisions which might be affected by the evaluation results. These decisions may be broadly categorized as funding decisions and management decisions.

Funding decisions include the allocation of resources to the system, as well as decisions to establish or to discontinue systems. Management decisions are those that govern the use of funds to acquire documents, to index documents, to provide search and reference services, and so on. These decisions are highly dependent upon funds available, of course, and decisions to provide funds are dependent upon management's use of funds previously provided. The distinction between funding decisions and management decisions is useful in determining whether to focus upon economic measures or performance measures. These ideas will be developed subsequently.

At the risk of oversimplification, we can say that one would like to be able to measure cost, performance, and benefits. There is little controversy over the unit of measurement for cost -- the dollar provides a convenient unit for such measurement. There is some controversy over whether or not one should use discounting techniques, but this is a minor problem. There is a great deal of controversy over what constitutes cost.

All of the problems associated with measurement of cost are magnified when one tries to measure benefits. Some benefits can only artificially be expressed in monetary units and some not at all.

More attention has been paid to measures of performance than to measures of cost and benefits. Most of the difficulties arise because performance is not a uniquely defined property -- it is many things to many people. No single measure can be expected to satisfy all groups of persons who wish to quantify performance, although some attempts to do so appear in the literature.

Even within the management of a system, one person will be interested in acquisition, another in indexing, another in hardware, and so on. Thus, it seems wise to establish as a first principle that one must provide multiple measures of performance. We discuss some of the practical limitations on measurement in the following section.

2.4 What measures are appropriate to specific requirements?

A document retrieval system is composed of (1) functions, (2) organizational structure, (3) processes, and (4) things (documents, buildings, computers, files, and so on). For the kinds of evaluation which concern us,we are most interested in functions, organizational structures, and processes.

2.4.1 System functions

Document information transfer systems involve flow of information via a message unit from a source (author) to a destination (user). For the purpose of this report, a message unit is a collection of words or symbols such as the full text of a report, a research finding, or a journal article that is transmitted from an author to users. Some abstracts of full texts may be considered message units

if their purpose is to inform rather than merely to provide access to the full texts. Also, it is conceivable that some day scientific and technical research reports may be formatted so that portions of full texts might serve as message units. However, full texts are the primary message units of consideration in this chapter. For convenience they are called "documents," although a broader meaning is intended.

The document information transfer process usually consists of a series of six basic system functions as given in Figure 2.1.

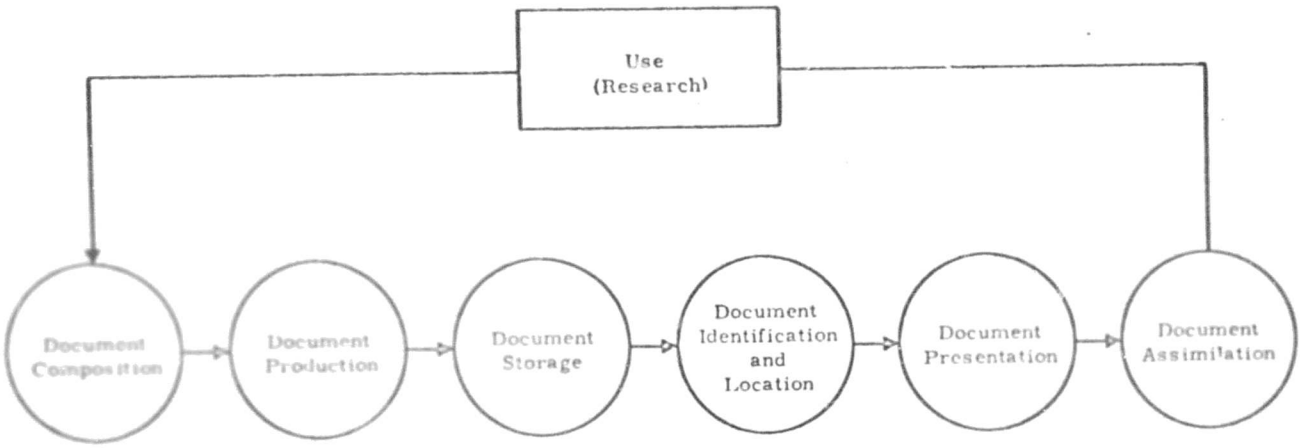


Figure 2.1. Basic document information transfer functions

The arrows in the schema represent flow only to the extent that the document information transfer functions often occur in the sequence shown. A short description of these functions is given to clarify terminology used subsequently.

(1) Composition -- preparation of a report or publication, orally or in writing.

(2) Production -- typing, printing, or taping a document.

(3) Storage -- maintenance and preservation of copies of documents at identified locations.

(4) Identification -- determination of the identity and location of documents to be distributed in response to retrospective searches or by selective dissemination.

(5) Presentation -- physically turning over a copy of a document to a user.

(6) Assimilation -- perception by a user of the information (if any) in a document.

(7) Use -- the use to which the information is put, which in turn may result in new composition.

It seems clear that most information transfer processes involving flow of documents from authors to users incorporate all of the six basic functions to some degree.

Actual transmission of a document from author to reader can take place by a variety of channels described by the functions above. A few examples of document information transfer channels are given below:

a. Transmission of scientific and technical reports --

Composition	Production	Storage	Identification	Presentation	Assimilation
Staff	→ CFSTI (microfiche)	→ Library	→ Retrospective search	→ Viewer	→ Read
Researcher	→ Author (type)	→ Author	→ Agency directive	→ Agency distr.	→ Read
Technical Writer	→ CFSTI	→ CFSTI	→ Selective Dissemination	→ Mail	→ Read

b. Transmission of books --

Composition	Production	Storage	Identification	Presentation	Assimilation
Staff	→ Publisher	→ Library	→ Browsing	→ Hand	→ Read
Author	→ Publisher	→ Author	→ Complimentary copy	→ Hand	→ Read
Annual Review	→ Publisher	→ Publisher	→ Advertisement	→ Mail	→ Read

o. Transmission of journal articles --

Compo-sition		Production		Storage		Identifi-cation		Presen-tation		Assimi-lation
Staff	→	Publisher	→	Library	→	Bibliog-raphy	→	Viewer	→	Read
Author	→	Publisher	→	Personal library	→	Memory	→	Hand	→	Read

Only three of the above-described functions are of interest to us in the kind of evaluation discussed in this report. They are (1) storage, (2) identification, and (3) presentation of documents. Thus, we would like to be able to measure costs incurred by performance of and benefits accruing from these three system functions.

2.4.2 System organizational structure

Document information transfer systems are managed by organizational entities that provide services to store, identify, and present documents. The functions may be performed at several levels of organizational complexity including:

(1) By an individual.

(2) By a "local" library such as a departmental government library, company library, or university library.

(3) By a central reservoir service such as the Library of Congress, the Clearinghouse for Federal Scientific and Technical Information, or the National Library of Medicine.

We define a "proprietary system" to be that part of the information transfer system which is under the management control of a given organizational entity. Evaluation almost always takes place within a proprietary system, and the proprietary entity may be government-wide such as the Library of Congress, or it may be part of a larger entity such as a government agency, a company, or a university library. For evaluation purposes, it is important to

distinguish the class of systems in which the user belongs to (is a member of) the proprietary system. This particularly is true in establishing measures of system benefits which are described in a subsequent section.

The basic document information transfer functions may be performed at any of the levels of centralization. Furthermore, it is clear that a system under evaluation may perform transfer functions within the proprietary entity, outside the proprietary entity, or both, depending on the particular system. For example, if the library of a government agency is being evaluated, it is clear that evaluation should include not only the specific services provided by the library but also the external services that the library utilizes, such as the Clearinghouse for Federal Scientific and Technical Information and inter-library loans.

The organizational unit for which the evaluation is being done exerts a great influence on the definitions of costs, performance, and benefits, as well as on the selection of specific measures. Aggregate costs to users may be substantially different from aggregate costs to society if the Federal Government is providing a major portion of the support from tax funds. Also, benefits to society are almost certain to be different from benefits to members of the organizational unit being examined. Finally, a different set of performance measures is necessary for diagnosis of a system than is necessary for broad management policy decision.

2.4.3 System processes

This section discusses ways in which three basic information transfer functions (storage, identification, and presentation) can be accomplished.

Storage may be characterized by degree of centralization and by the organizational entity that has administrative control over the storage. At one end of the spectrum is a system that performs

bibliographic service -- it has no storage of its own but provides access to documents stored almost anywhere. At the other extreme is the National Lending Library for Science and Technology which has a large centralized collection but which provides essentially no bibliographic service. One of the variables in the storage process, then, is the extent to which the system under evaluation attempts to store the documents that its users may require. Thus, the objectives of the system with respect to storage policies affect the nature of measurements that must be taken. Other storage variables are input processing and handling, method of physical storage, purging practices, and request processing. Each of these affects the selection of performance measures.

The identification function may operate in one of two system modes as follows:

(1) Passive mode in which the document information transfer process is initiated by the user; e.g., a retrospective search

(2) Active mode in which the document information transfer process is initiated by the system; e.g., selective dissemination of information.

Passive systems may further be classified by the processes actually used to perform the identification function. In the passive mode, document identification occurs (1) by the user having prior knowledge of a document's identity (by title, author, or subject matter), (2) by the user requesting references by subject matter without prior knowledge of specific documents, and (3) by the user "browsing" to generate new ideas or to investigate new fields without prior knowledge as to what documents are actually sought. In the first instance, the document identification process is minimal since the user already knows of the document's identity. In the second case, however, the document identification process involves retrospective searching capabilities that interact with a number of sub-processes such as acquisition, cataloging, indexing, searching, screening, and search output. Browsing requires further unique document identification processes since a user needs to peruse full-text or at least informative document representations from a fairly substantial but ill-defined file.

Systems operating in an active mode also involve the basic function of identification where the system (not the user) assumes the initiative for the final document information transfer process. In the active mode, document identification often occurs by prior agreement between the system and user, in which case the system disseminates document identification (and location) information (title, author, keywords, or abstracts) to a user or group of users on the chance that the user will need and subsequently retrieve and use the document. The document identification function in the active mode also interacts with sub-processes such as acquisition, indexing, abstracting, reproduction, dissemination, and user matching techniques.

The document presentation function involves physically placing a copy of the document into the possession of the user. This function may also take place at all levels of centralization, but in this section presentation is defined as the terminal function in document transfer from author to user. The processes associated with this function include ordering, transmission (e.g., mail, teletype), viewing, handling, and so on.

So far we have seen that measures are needed for costs, performance and benefits for storage, identification, and presentation of documents for two modes of system operation (active and passive), and that each of the functions is a combination of many tasks or processing procedures. Clearly, then, a large number of measures may be required.

2.4.4 A framework for the selection of measures

Table 2.1 lists the three principal functions of storage, identification, and presentation with some of their principal components that require evaluation. It also shows some of the things one might consider measuring in order to form evaluative judgements about those functions and components. The listing is not intended to be comprehensive, but for illustrative purposes it covers the principal items of interest to the typical evaluator.

2.5 How does one use measures

One of the problems in evaluation of document retrieval systems is that one frequently cannot measure directly the thing he would like to quantify.

Table 2.1 Partial list of measurement candidates for the evaluation of document retrieval systems

Functions and Objects to be Evaluated	Candidates for Measurement
a. Storage	
a.a. Acquisition	1. Responsiveness to Requests from Users 2. Selection of High-demand Documents 3. Ordering Mechanics
a.b. Files	1. Composition 2. Location 3. Organization 4. Storage Form
b. Identification	
b.a. Indexes	1. Indexing Procedures 2. Index Structure -- Depth Hierarchical Structure Correlation of Terms 3. Index Languages
b.b. Retrospective Searching	1. Entry Vocabularies 2. Assistance 3. Procedures 4. Equipment 5. Screening (use of document representations) 6. Interaction with System
b.c. Selective Dissemination	1. Screening 2. Current Awareness 3. Use of Document Representations
c. Presentation	1. Form 2. Timeliness

For example, it would be convenient to have a single measure of performance so that, in comparing two systems, one could select the one with better performance. We have seen, however, that performance is a composite of many things, some easily quantifiable and some almost impossible to quantify. Also, we frequently want to measure one thing but must measure a substitute (a proxy). For example, user satisfaction is a concept that we would like to quantify. It is not measurable directly, however, so we measure the proportion of the user's literature citations obtained through the system, the proportion of search output examined by the user, his qualitative assessment of satisfaction with particular services, and so on.

The measures presented in Section 2.4 seem, on intuitive grounds, to contribute something to one's knowledge of a system and how well it works. What is needed is a framework for tying together these measures so that one can see their implications with respect to the overall system viewed as a unit. Such a framework is called a "model". The literature contains many attempts to construct document retrieval models -- some quite primitive and some utilizing sophisticated mathematical concepts. A great deal more must be learned before global document retrieval models can be constructed that faithfully picture a generalized document retrieval system operating in a real world environment. This is not to discourage model building -- far from it -- but rather to point out that any models presented here must be considered stages in an evolutionary process in which perfection is still far in the future.

It seems abundantly clear to us that an understanding of how a system works, what its environmental constraints are, and what will happen to it as a result of certain operational patterns is a necessary prerequisite to the use of measures in the formation of evaluative judgments. With this in mind, we discuss some macro-models in the following section which rely heavily on the concepts of costs, benefits, and performance.

2.5.1 Digression on costs, performance, and benefits

Cost describes the input of resources to a system in terms of monetary units. Measures of system performance describe attributes that can be controlled by system management, such as speed, accuracy, and quality -- all subject, of course, to budgetary restraints. Benefits describe consequences of system performance in terms of value, return on investment, effect on behavior of the user, effect on other systems, and non-quantifiable consequences that may be a direct result of the system or a result of interactions with other systems.

It seems sensible to base system decisions on a comparison between the cost necessary to attain a particular degree of performance and the benefits that are derived from this performance. Thus, system performance variables might be chosen with two purposes in mind. The first purpose might be to relate performance variables to costs and to benefits. The second purpose might be to diagnose the system by means of the performance variables so that evaluation can lead to improvement when performance is not satisfactory from the standpoint of the cost/benefits comparison.

The schema in Figure 2.2 gives a general relationship among cost/performance/benefits measures and four basic functions of document information transfer. As shown in the schema, system performance variables relate the performance of the system processes to cost and to benefits. For example, in retrospective searching (one mode of document identification), system cost is partially determined by average total number of documents identified, the number of transactions completed, and average time per transaction. System benefits are partially derived from such performance measures as search accuracy and response time.

The schema in Figure 2.2 above is purposely oversimplified.

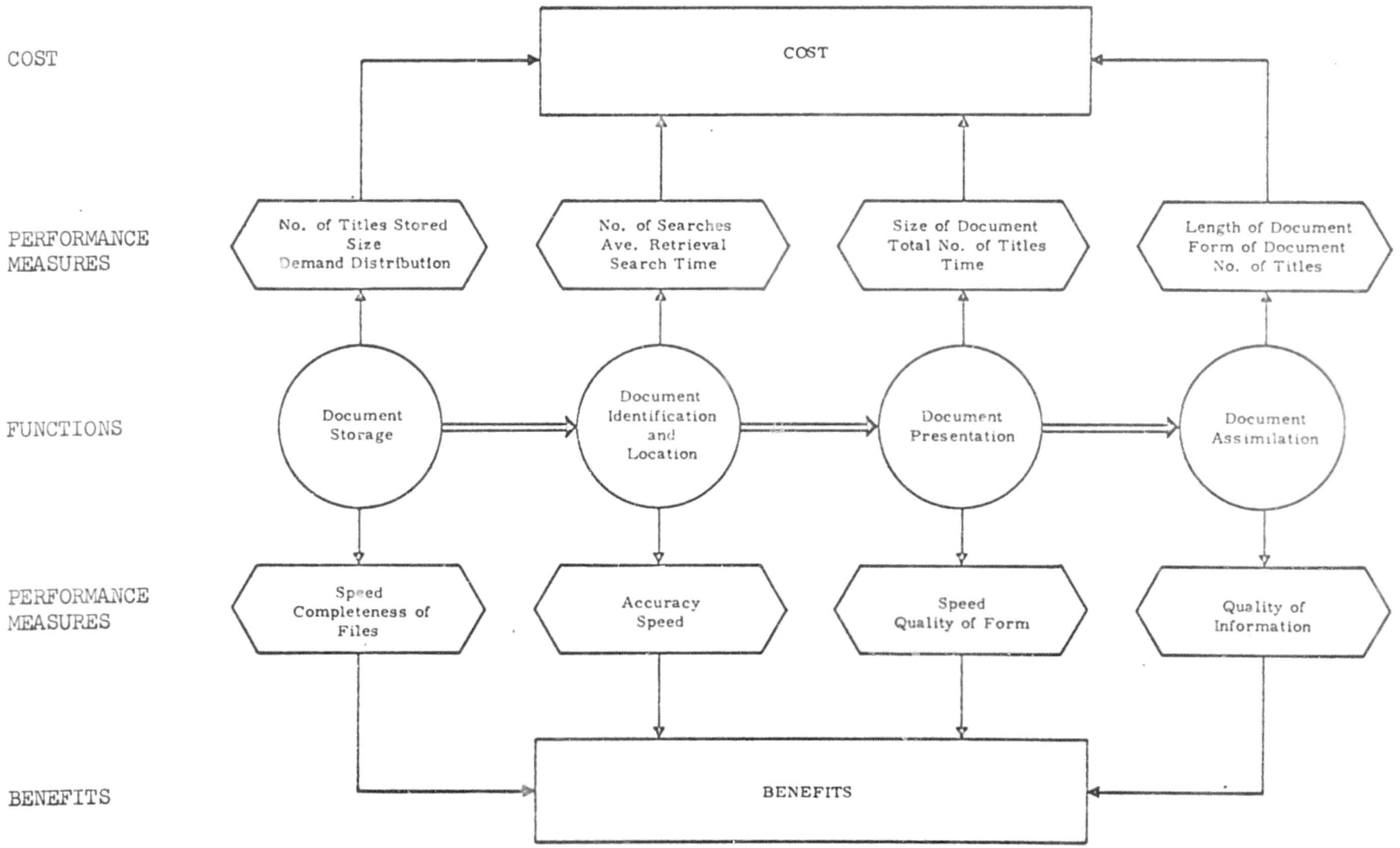


Figure 2.2. Schema for cost, performance, and benefits measures for storage, identification, presentation and assimilation in document information transfer systems

System cost, for example, must be subdivided into fixed costs and two kinds of variable costs. The first kind of variable cost is a function of the number of transactions (document information transfers by one process or another) performed over a specified planning period, such as five years. The second kind of variable cost is that directly related to alternative system processes under evaluative investigation. Fixed costs include system equipment and development costs which are depreciated over a specified planning period and operational costs (such as staff salaries, rental, etc.) that do not vary by number of searches or by alternative system processes.

An example may help explain the general philosophy of choosing performance measures to determine costs. Suppose a government agency's information center is considering indexing at two distinct levels of depth (that is, terms per document) for input to a retrospective search system. The variable costs attributable to the two alternative index processes are fairly well defined. Greater indexing depth represents higher indexing costs (per document). However, the variable costs attributable to transactions (retrospective searches) not only affect the cost of identification but also reflect the cost of other transfer functions. Greater indexing depth (1) yields a larger number of documents identified by the system (for a given query), (2) affects search time because of the increased number of terms to be matched, (3) requires more documents to be screened and located, and (4) potentially yields more documents retrieved and used. All of these variables directly yield measurable increased costs. Of course, the question of indexing depth is not entirely resolved by estimating its effect on system costs. Greater indexing depth should also produce more relevant documents for a given query, thus reflecting greater accuracy.

One of the most important considerations in determining performance variables and benefits measures is whether or not a user is

a part of the system being evaluated. If he is a part of that system, document retrieval performance directly affects the goals and objectives of that entire organizational unit. Therefore, the management of the proprietary entity, as system funders, will want to know whether or not document information transfer system expenditures yield sufficient benefits to make them worthwhile. Furthermore, if a user is part of the system under evaluation, his participation may also be investigated.

If the user is not part of the proprietary entity, the entity has no administrative control over what the user does with the information or with the consequences of the use of the information. However, the system still has an important interest in the consequence of the service in terms of whether or not a user continues to employ the system. Thus, when users are not a part of the proprietary system, the system resides in a market-like environment, and all the economic, marketing, and competitive ramifications of the environment must be considered in evaluation.

The schema in Figure 2. 3 shows relationships between performance and benefits when users are not part of the proprietary system being evaluated. The user assimilation function no longer assumes a prominent role. However, processes necessary to accomplish storage, identification, and reception all have performance variables such as accuracy, speed, and quality. These performance attributes determine the user's degree of satisfaction with service. User satisfaction and price of the service, along with promotion and advertising, provide motivation to begin or to continue using the system, which in turn creates overall demand for system use. The price for the service is determined in part by cost and in part by income per transaction. The price per transaction times the total number of transactions determines income produced by demand.

When users are not a part of the proprietary system under evaluative consideration, the system is usually one of two kinds: either a local facility, such as a public technical library, in which all three functions are performed either in-house or by request from another service facility, or a central reservoir, such as the Clearinghouse for Scientific and Technical Information and the American Institute of Physics, that provide one or more system functions.

A technical library may identify a document in-house but may send away to obtain it. The performance variables cited in Figure 2.3 still remain the same, as do their relationships to user satisfaction, motivation to use the system, and income. Cost of document delivery in this case is the price paid plus handling costs.

In some instances, the Federal Government or some other external source partially supports development or operation of a system in order to ensure that the system is available to the professional communities. In these instances, the sponsoring agencies should have an interest in the use of the system and the value derived from the system in order to determine if their own resources are properly allocated.

A schema relating performance variables and benefits measures when users are part of the proprietary system is given in Figure 2.4. Document assimilation becomes an integral part of the system, and its accomplishment yields performance variables that partially determine user satisfaction. Satisfaction motivates the user to continue using the system, which creates demand for more system transactions. Document assimilation also determines the use made of the information, which in turn affects the behavior of the user and of the system of which he is a part. These changes in behavior contribute to the value of the proprietary system.

The arrows in Figure 2.2 through 2.4 may be interpreted roughly

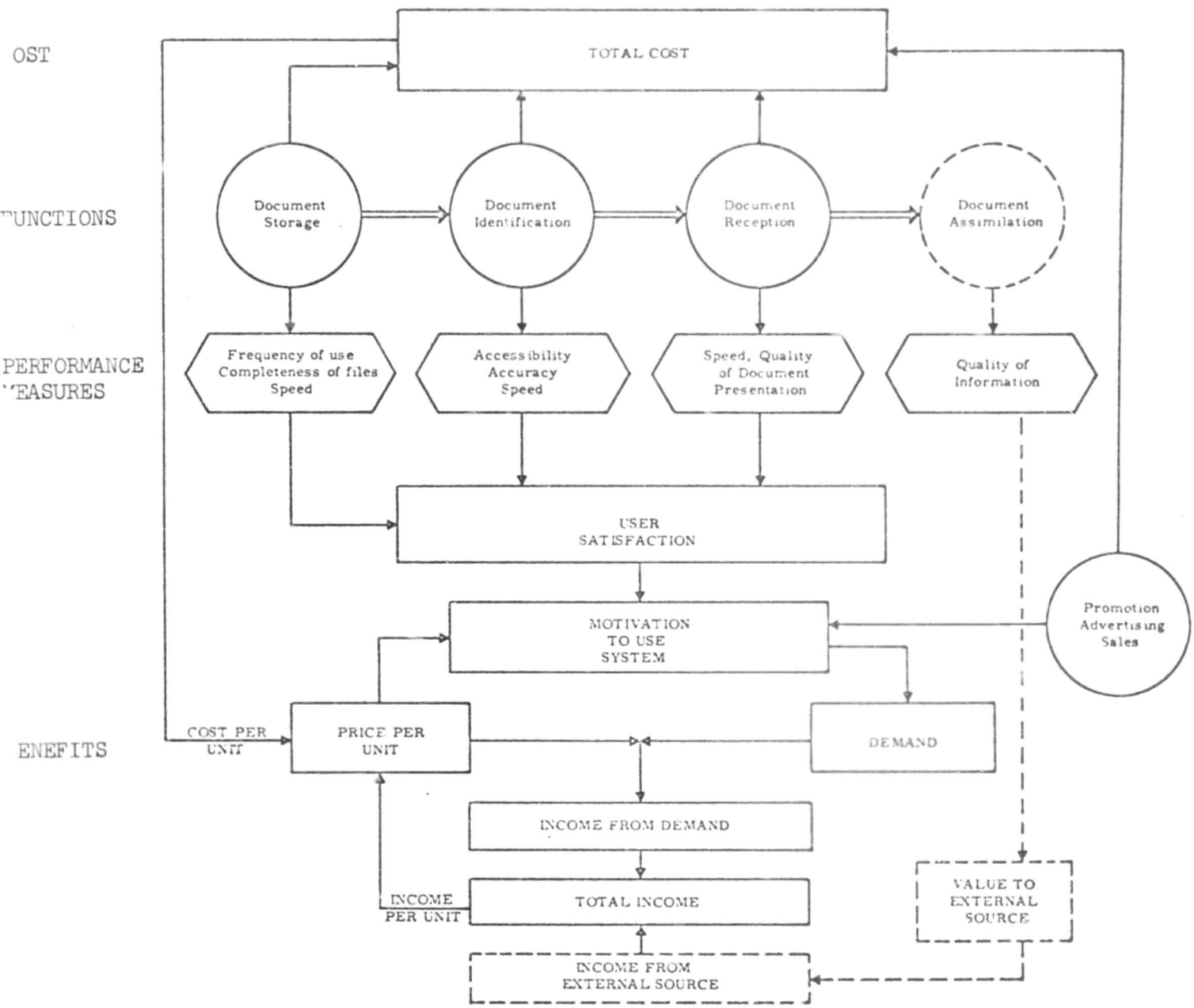


Figure 2.3. Relationships among functions and measures when users are not a part of the system being evaluated

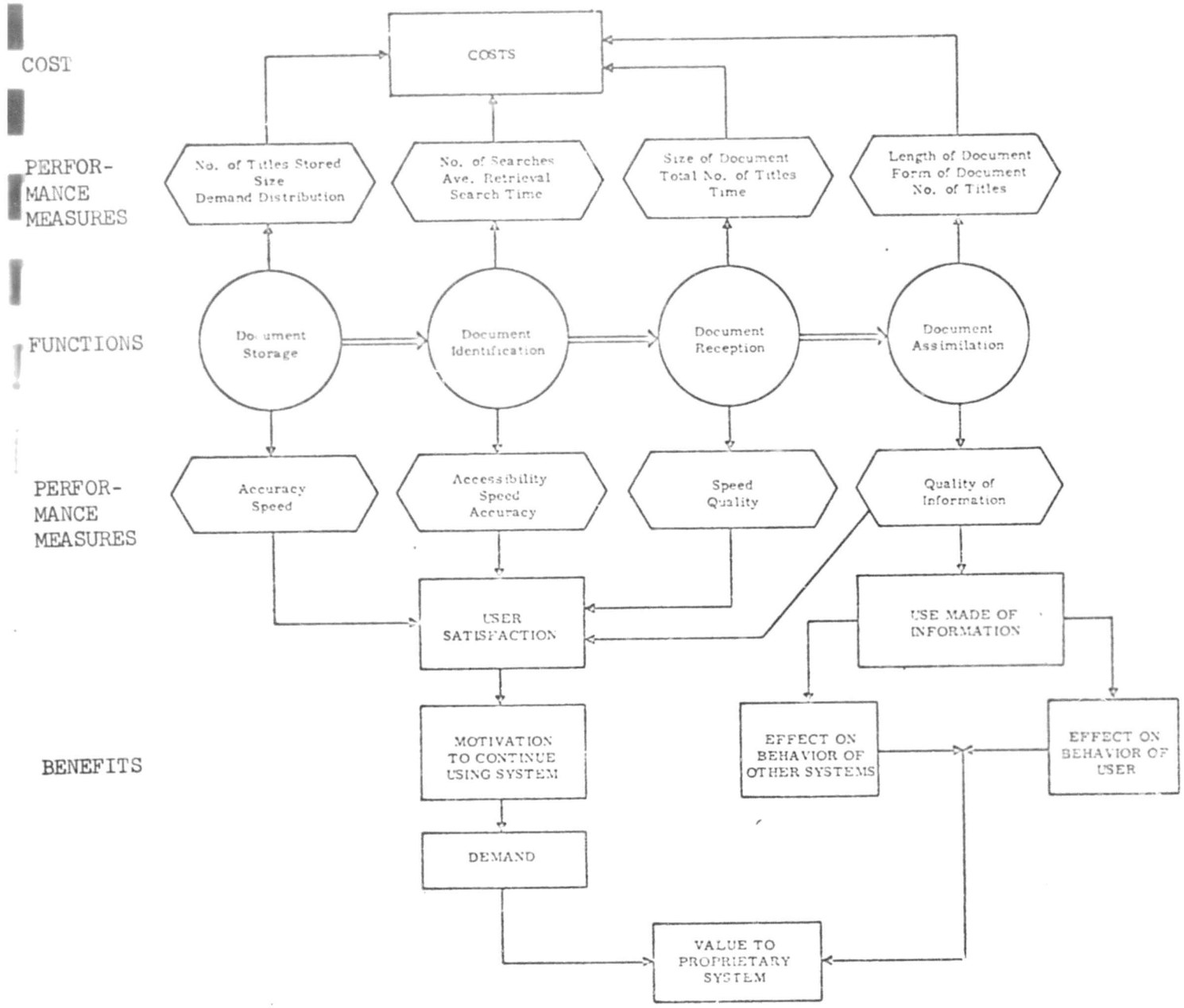


Figure 2.4. Schema for performance and benefits measures when users are part of the proprietary system

as meaning "may be measured by" when the arrows terminate at a performance measure, and "has (or have) an effect on" when the arrows terminate at other boxes. The functional nature of the effects has not been specified, so the modeling is incomplete. We would like to be able to provide such functional relationships, but understanding of document retrieval systems has not advanced to that point. What we hope to have accomplished in this section is to show which measures are related to costs and which to benefits, and something about the interpretations that can be placed on "benefits" in document retrieval systems.

2.5.2 Some specific uses for measures

In the previous section we emphasize the use of performance measures in macro-evaluation, that is, at the funding and policy-making level. In this section we examine how selected measures may be used for specific diagnostic and management purposes.

The user's requests for documents are based upon his prior knowledge of what he wants.

(1) He may have prior knowledge of a document's identity (by title, author, and publication).

(2) He may request references by subject matter without prior knowledge of specific documents.

(3) He may "browse" a file of documents without prior knowledge as to what is sought in the way of documents or even the specific problem he wishes to solve.

When a user has prior knowledge of a document's identity by bibliographic reference, the only remaining identification process may be to locate a copy of the document. The principal performance variable for deriving system benefits is response time, measured from request to receipt of a copy of the document. System accuracy should not be a factor in this case. The response time can be used for diagnostic purposes by portioning the time into various retrieval activities to determine which activities are taking an unsatisfactory amount of time.

Referring to Figure 2. 3, where users are not part of the proprietary system, the total response time may consist of locating the document, as well as physically retrieving it from storage. User satisfaction can be related to the entire response time by recording user's degree of satisfaction for a number of searches. Simple regression analysis can determine user satisfaction (as a dependent variable) from response time (as an independent variable). However, it may make more sense to observe occurrence and nonoccurrence of repeated system use and to measure system demand directly as a function of response time in a simple regression model.

Referring to Figure 2. 4, where users are part of the proprietary system, user satisfaction and repeated use of the system can be determined in the same way as mentioned above. However, value may be difficult to determine. If the cost of locating and presenting documents can be estimated, one approach may be to provide the user with estimated cost and probable response time prior to retrieval and to let the user decide at that time whether or not to obtain the item in view of this information.

Cost varies with the number of transactions performed in the manner described above. If a desired document resides within the system, average cost per search is a function of the number of searches and the frequency of use of a particular document since the cost should be allocated over all uses of that document. If a desired document resides outside the proprietary system, the cost is the sum of the price paid to obtain a copy of the document and appropriate locating and handling costs. Cost also varies with alternative system processes, and these alternatives should be investigated with regard to both costs and response time.

When users request references by subject matter without prior knowledge of specific documents, the system is subject to error in

identifying the correct documents. Therefore, identification accuracy must be considered in determining the performance of a system.

The most difficult system performance measure to obtain is quantification of system accuracy, since it is not abundantly clear what accuracy is. The system, in effect, makes a relevance assessment on every document in the file when it responds to a search query. This relevance assessment may be translated into a zero-one variable when the system retrieves some documents and does not retrieve others. It is a relevance score or ranking when the system responds in that manner. We assume that a knowledgeable judge can provide a relevance assessment on every document with respect to the verbalized request presented by the user, and it seems reasonable to suppose that the user can provide his own assessment of relevance.

Accuracy has a different interpretation when the user is a part of the system than when he is not. We suggest that accuracy may be interpreted as follows:

a. The relationship between system relevance assessment and user relevance assessment <u>if the user is a part of the system</u>.

b. The relationship between system relevance assessment and relevance to the verbalized request <u>if the user is not a part of the system</u>.

In other words, the system cannot bear the responsibility for correct formulation of the request unless the user is, in fact, part of the system.

Researchers engaged in two recent major research projects to investigate relevance agree in principle, at least, that relevance should be defined as a relationship between a user's information question (or information requirement statement) and a document [1, 2]. This relationship is called <u>user relevance judgment</u>. A user relevance judgment can be either a dichotomy (i.e., a document does or does not answer the user's question) or a multi-valued scale, such as a "degree" of relevance

given by scale values (e.g., 0 to 10).

User questions are processed through the system, resulting in a response that hopefully predicts (or resembles) the user relevance judgment. The system relevance response is defined as the system's assessment of the relationship between the user's question and a document. The system relevance response may also be a dichotomy with value one or zero (i.e, a document is identified or not identified) or may be multi-valued such as is the case with weighted responses from associative retrieval systems. It is emphasized that system relevance responses are independent of user relevance judgments in the sense that the relevance numbers are assigned by different entities.

The multi-valued measures of user relevance judgment and system relevance response can be plotted against one another as in Figure 2.5. User relevance judgment can be estimated from system relevance response mathematically, by measures of correlation or by conditional expectation. Thus, in some instances it may be possible to relate performance to system benefits if the relationship of user relevance and benefits is known.

Similarly, system relevance response can be estimated from user relevance response. For diagnostic purposes, the variation about the conditional estimate can be analyzed by residual analysis to determine the extent to which various processes contribute to the variation.

Someone must judge the relevance relationship between a question and a document to establish the user relevance judgment. Also, an information system intermediary may judge relevance between a question and a document to produce the system relevance response. Two major studies [1, 2] on human judgments of relevance indicate that these judgments are subject to considerable variation due to differences in documents, judgment conditions, questions, judges, and even different judgments made by the same judge over time. Even though these

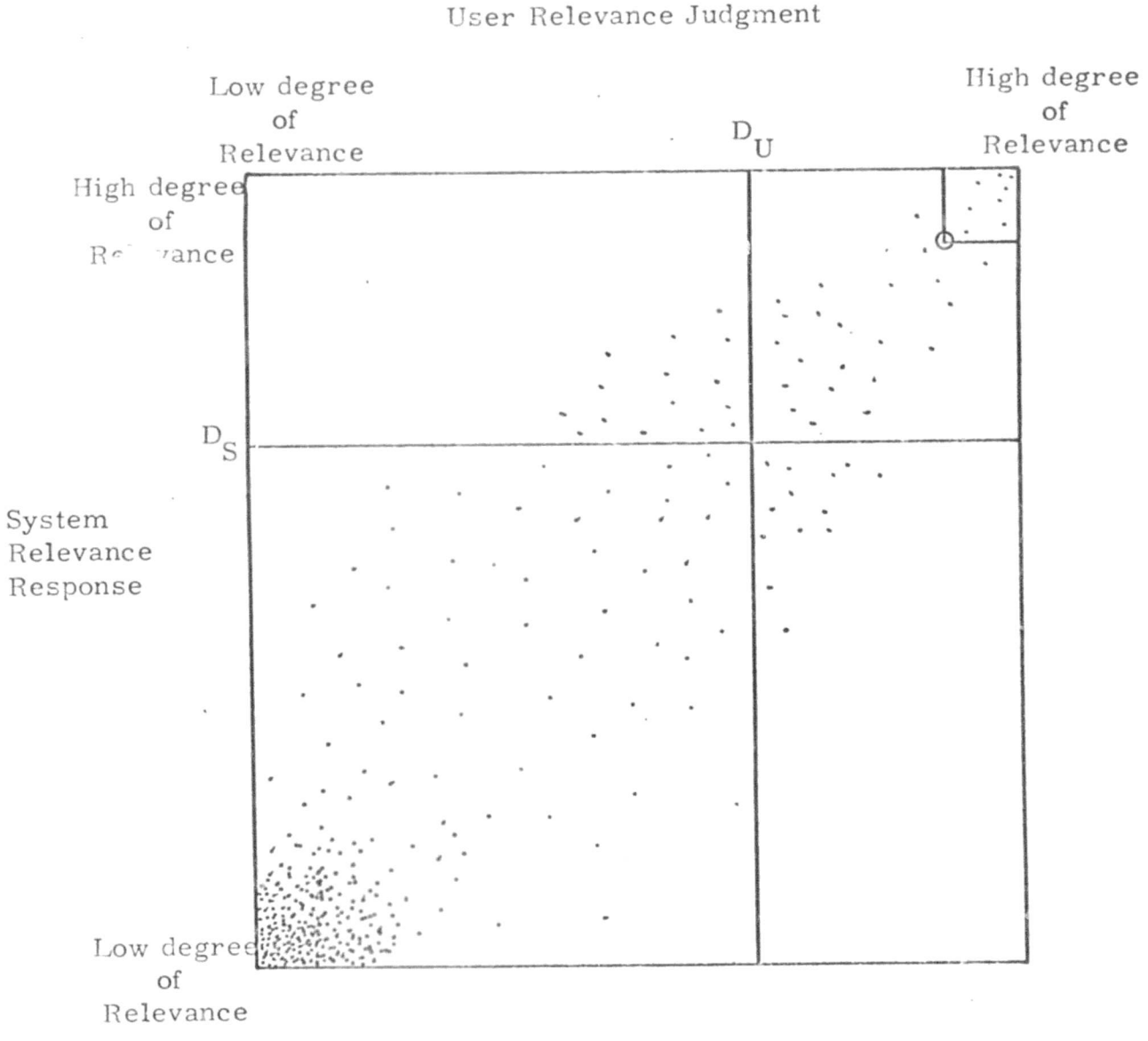


Figure 2.5. User relevance judgment plotted against system relevance response for individual documents and a given search question

factors are not directly related to cost or to benefits, they should be isolated if possible to permit better estimates of user relevance judgments (hence, benefits) and also to enable a researcher to diagnose system processes more effectively.

The continuous values of user relevance judgment and system relevance response plotted in Figure 2.5 provide some advantages over dichotomous judgments in that additional discrimination power is possible, and, secondly, a more flexible search strategy can be employed in which the decision point D_S is lowered to identify new documents to where the user is satisfied with the search results.

Swets [3] has considered continuous values of system relevance responses and dichotomous values of user relevance judgment. He suggests that the frequency distribution of documents found to the right of the user decision point D_U in Figure 2.5 (i.e., those judged relevant) and the frequency distribution of documents found to the left of D_U (i.e., those judged not relevant) be plotted against values on the system relevance response scale as shown in Figure 2.6.

He also defines D_S as a critical point above which the system selects items for examination. The proportion of Distribution B to the right of D_S in Figure 2.6 can be identified with the probability of retrieving a document, given that it is relevant. The proportion of Distribution A to the right of D_S can be identified with the probability of retrieving a document, given that it is not relevant. Swets plots the probabilities against one another for all values of D_S and refers to this relationship as the operating characteristic curve which he recommends as a measure of performance.

Despite some advantages inherent in the relationships indicated above, we feel that dichotomous assessments of relevance are more meaningful for evaluative purposes since the decision processes of both the system relevance responses and of the user relevance judgments are

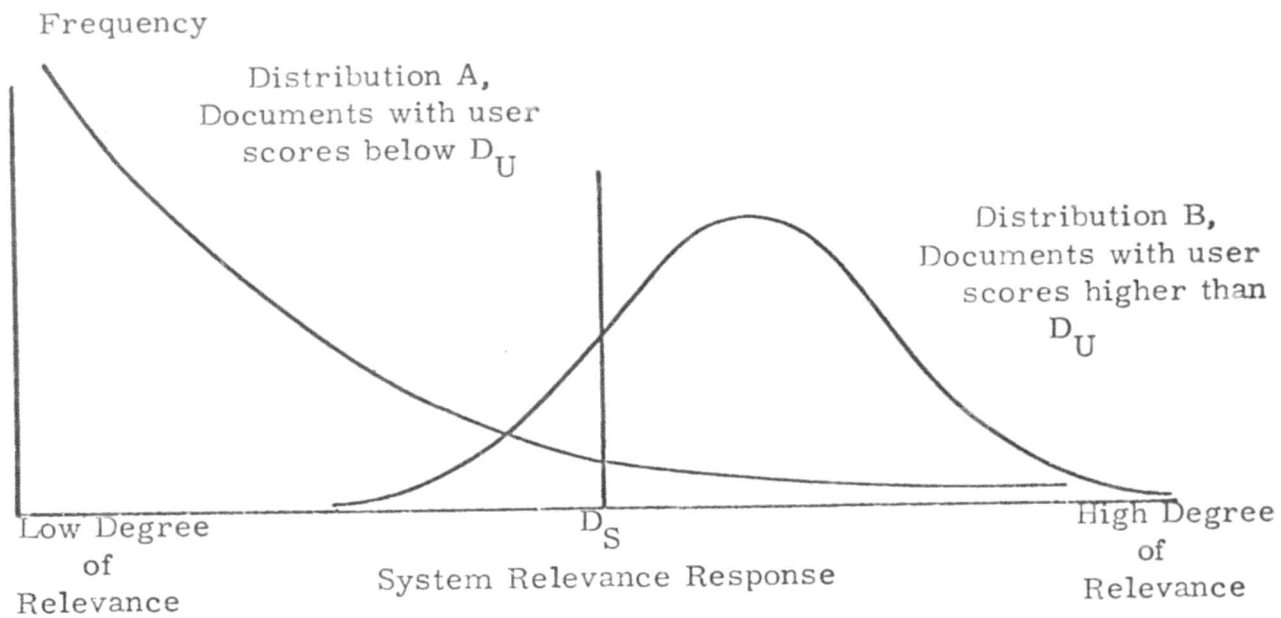


Figure 2.6. Frequency plots of system relevance scores for documents judged relevant or nonrelevant by the user

typically dichotomous in nature. Even though user relevance judgments may be multi-valued, the user may feel that a relevance decision depends on whether or not his relevance judgment exceeds a critical value designated by D_U in Figure 2.5. Similarly, a decision may be made to send or not to send documents to a user, depending on a critical system relevance response value, labelled D_S in Figure 2.5. Thus, dichotomous values may be necessary either because of decision requirements or because the relevance assessments are dichotomous in nature. In either case, the scores in Figure 2.5 can be transformed into the retrieval categories of Figure 2.7. The cell values X_{11}, X_{12}, X_{21}, and X_{22} in Figure 2.7 correspond to the number of documents observed in the four quadrants in Figure 2.5.

Performance measures for estimating value from system output can also be determined from the entries in the table in Figure 2.7.

		User Relevance Judgment: Not Relevant $V_{\bar{r}}$ D_U	Relevant V_r	Total
System Relevance Response: Relevant R_r D_S		X_{11}	X_{12}	$X_{1.}$
Not Relevant $R_{\bar{r}}$		X_{21}	X_{22}	$X_{2.}$
Total		$X_{.1}$	$X_{.2}$	$X_{..}$

Figure 2.7. Retrieval categorization with dichotomous values of user relevance judgments and system relevance response.

For example, the probability of user relevance judgment, given that the system assesses a document to be relevant, is $P(X_{12}|X_{.1})$ or $P(V_r|R_r)$ which corresponds to the well-known precision estimate. Thus, if the value of a relevant document is known (an unlikely event), the value to the system can be determined from the system output.

When a user is not part of the proprietary system, it is not as essential to determine the value of the output as it is to determine how well the system satisfies the user. Referring to Figure 2.3, the best gross measure to observe to determine the effectiveness of the system is the demand placed on the system by users. However, this measure is not timely enough or discriminating enough for sensitive management control. Therefore, it may be a better procedure to estimate user satisfaction from accuracy, speed, quality, accessibility, and similar measures. User satisfaction can be estimated by regression modeling where satisfaction scores are assigned by the user and estimated from the performance measures. It is important also to obtain a statement of intention to continue using the system and then to observe whether or not the user actually does use the system. The group of users in the test can

be compared against users not in the test to determine their relative use of the system. Then, at least in theory, system demand (hence, income) can be estimated (utilizing the regression model mentioned above) from system accuracy as measured by the precision ratio and by total retrieval.

Given a user relevance judgment, the expected system relevance response is estimated by $P(X_{12}|X_{.2})$ or $P(R_r|V_r)$ which corresponds to the well-known recall ratio. Diagnosis of the system can be performed by analysis of sources of system failures.

Deterioration in system accuracy stems from four principal processes:

(1) Interpretations of a user's question by an intermediary

(2) Translation of a user's question into terms available to the system, i.e., formulation of a search query

(3) Document indexing and coding

(4) Screening of documents that have been identified by the search process

Even though system relevance response and user relevance judgment are determined independently, they are usually highly correlated (in a statistical sense). This is because all of the usual retrieval processes involve a form of the user's question. For example, an intermediary interprets a user's question (initially stated in natural language), he translates that question into a system query (stated in terms available to the system), the system matches the query and indexed documents in terms available to the system, and an intermediary screens documents from the system output based on his interpretation of the user's question. All of these forms of the user's question are related to documents in the file in a manner similar to user relevance judgment and should hopefully be highly correlated to user relevance judgment. The relationships are given in the schema of Figure 2.8.

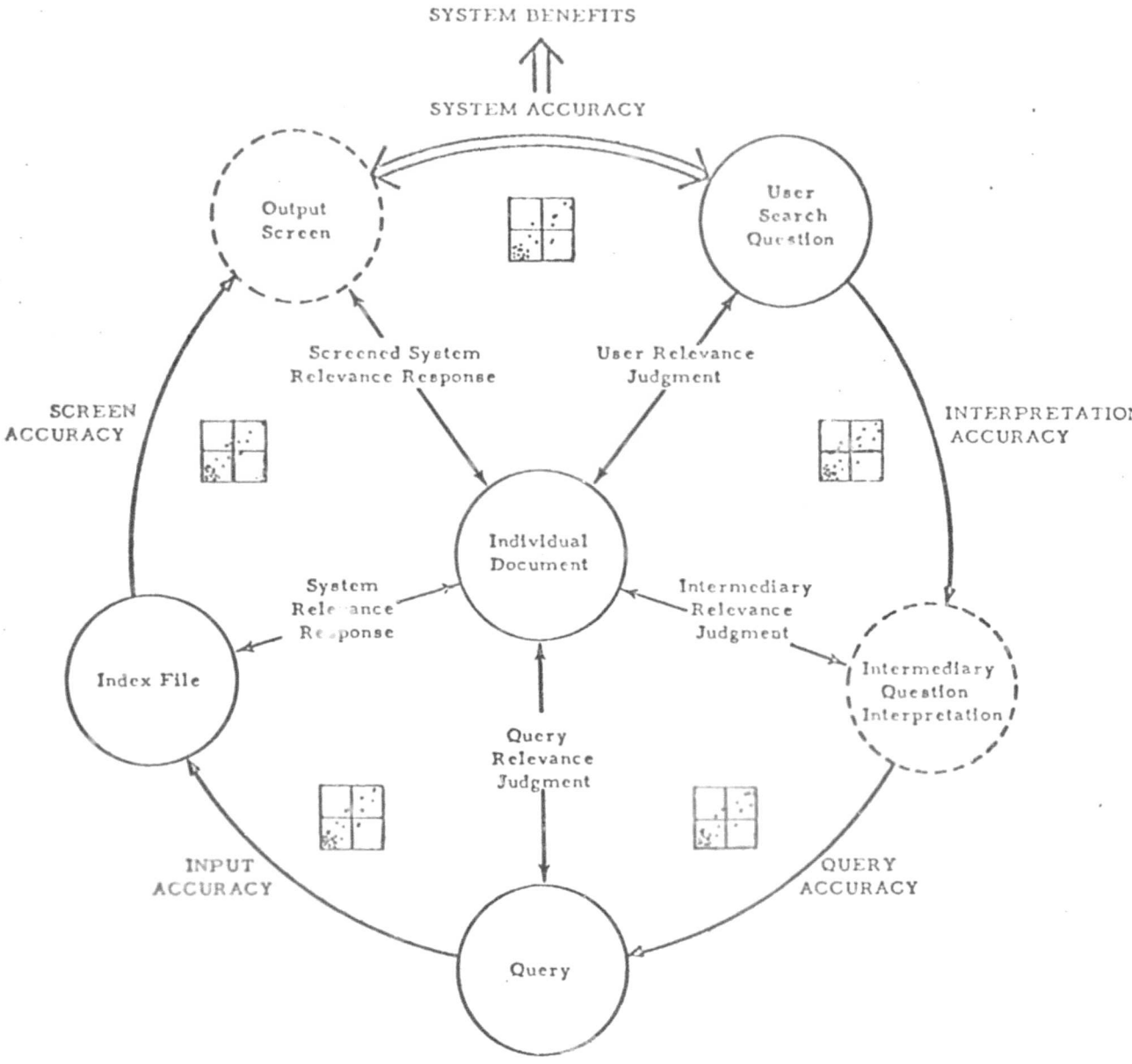


Figure 2.8. Schema presenting system accuracy and relationships that contribute to it

The user relevance judgment, which is the relationship between a user search question and an individual document, is shown in the upper right-hand corner. Continuing clockwise around the schema, are the following:

(1) Intermediary relevance judgment, which is an intermediary's judgment of the degree of relationship between the user's question and a document.

(2) Query relevance judgment, which is an assessment of degree of relationship between a query and an individual document (i. e., does the document actually contain the concepts required by the query?).

(3) System relevance response, which is the system's assessment of the relationship between a query and a document. The system response is often a computer print-out that implies a value of 1 for those documents identified and zero for others.

(4) Screened system relevance response, which corresponds to an intermediary's judgment of the relation between the user's question and a document.

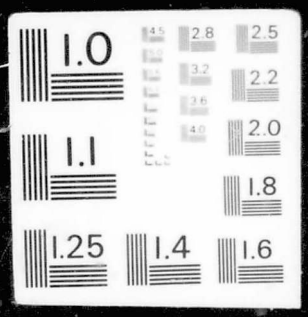
All of the relationships stated above can be multi-valued or dichotomous. Furthermore, all of the corresponding scores can be tted against one another to form relationships that are analogous to shown in Figures 2. 5 and 2. 7. The corresponding relationships uestion forms and documents provide measures for interpretation query accuracy, input accuracy, and screen accuracy in the schema. A lack of accuracy in each case contributes on of system accuracy, and the relative contribution can ng each link in the chain. Thus, system accuracy, when d, serves as a mechanism for diagnosing a system and .

cts user relevance judgment against intermediary ionship similar to that shown in Figure 2. 5 or

Figure 2.7 will exist. Clearly, there will be some deviation between the two judgments of relevance. Interpretation accuracy is the correlation between the two judgments. Some researchers, including Lancaster [4] and O'Connor [5] feel that question interpretation by the intermediary contributes substantially to deterioration of system accuracy. O'Connor points out that the deviations may be attributed to disagreements concerning the questions as well as judgments of the documents. The important thing, however, is to isolate the degree to which interpretation contributes to system accuracy and to investigate further if results indicate that faulty interpretation yields unsatisfactory system accuracy.

Query accuracy can also be found from the relationship between intermediary relevance judgment and query relevance judgment in the manner shown in Figures 2.5 and 2.7. If one plots query relevance judgment against user relevance judgment, the resulting relationship confounds the effects of question interpretation (by the intermediary) and query formulation. The two effects may need to be isolated to pinpoint ways of improving the system if accuracy is not satisfactory. There is ample evidence that query formulation can be a considerable source of difficulty [4, 6].

As mentioned previously, query relevance judgment is the score of a document against the query formulated in system language. System relevance response is the system's assessment of the relationship between a user's question (now in the form of a query) and a document. It is produced by matching the index file and the document for terms (or associated terms) stated in the search query. Deviations from a perfect correspondence in plots analogous to those in Figures 2.5 and 2.7 can be attributed to: (1) lack of agreement in interpretation of terms between indexers and searchers, (2) inadequate term list and structure, (3) indexing errors, and (4) depth of indexing. A number of



Relevance with respect to verbalized request	Relevance with respect to coder's interpretation	
	$C_{\bar{r}}$	C_r
V_r	$P(C_{\bar{r}} \mid V_r)$	$P(C_r \mid V_r)$
$V_{\bar{r}}$	$P(C_{\bar{r}} \mid V_{\bar{r}})$	$P(C_r \mid V_{\bar{r}})$

Relevance with respect to coder's interpretation	Relevance with respect to encoded request	
	$E_{\bar{r}}$	E_r
C_r	$P(E_{\bar{r}} \mid C_r)$	$P(E_r \mid C_r)$
$C_{\bar{r}}$	$P(E_{\bar{r}} \mid C_{\bar{r}})$	$P(E_r \mid C_{\bar{r}})$

Relevance with respect to encoded request	Relevance with respect to response by system	
	$R_{\bar{r}}$	R_r
E_r	$P(R_{\bar{r}} \mid E_r)$	$P(R_r \mid E_r)$
$E_{\bar{r}}$	$P(R_{\bar{r}} \mid E_{\bar{r}})$	$P(R_r \mid E_{\bar{r}})$

Relevance with respect to verbalized request	Relevance with respect to screener's interpretation	
	$S_{\bar{r}}$	S_r
V_r	$P(S_{\bar{r}} \mid V_r)$	$P(S_r \mid V_r)$
$V_{\bar{r}}$	$P(S_{\bar{r}} \mid V_{\bar{r}})$	$P(S_r \mid V_{\bar{r}})$

Figure 2.9. Conditional probabilities used in retrospective search models

Conditional probabilities will be designated by the standard notation $P(A|B)$ to be read "the probability of A, given B". Thus, $P(C_r|V_r)$ means "the probability that a document is relevant to the coder's interpretation, given that it is not relevant to the verbalized request".

Whether or not one can express relationships among the components of a retrospective searching system as probabilities, and the context within which such probabilities have meaning, requires some elaboration. Let us consider a probability such as $P(R_r|V_r)$, that is, the probability that a document relevant to the verbalized request (V_r) will be retrieved (R_r). If one chooses a request at random from the stream of requests entering the system, presumably it would be possible to determine whether a specified document in the system is relevant to that request or nonrelevant to that request. Also, one can observe whether such a document is retrieved or is not retrieved by the system. The relative frequency with which relevant documents* are retrieved by the system should approach stability as the number of observations is increased. Since an observation is identifiable with a document, such stability should occur either if many documents are matched against a single request or if a few documents in each of many searches are matched against their separate search requests. If the ratio generated by the latter method does, in fact, approach stability as the number of requests increases, the value approach as a limit will be referred to as "the probability of retrieval by the system, given relevance to th · ·balized request," that is, $P(R_r|V_r)$. In practice one is always working with relative frequencies since the limiting values are unknown. It is convenient in model construction, however, to work with the conceptual limits and to call them probabilities.

One can construct a model of the retrospective search system that has the following features:

---

* Relevant to the verbalized request.

(1) It shows the following summary figures:

   a. The probability that a relevant document will be retrieved.

   b. The probability that a nonrelevant document will not be retrieved.

(2) It shows the components (functions) that are the principal sources of error. This identification is provided by the display of Figure 2.9. Ideally, of course, all entries in these tables should be zeros and ones, with the ones in the lower left-hand and upper right-hand corners. The amount of departure from this idealization indicates the extent of departure from perfection.

(3) The effect of error-prone components on the total output of the system can be obtained. For example, it is possible to show what effect errors in interpretation by the coder have on system performance.

(4) The model will also show how specified improvement in any component will affect system output.

The model constitutes a simple application of the rules of probability in order to determine the following probabilities:

$$(1)\quad P(E_r|V_r) = P(E_r|C_r)\,P(C_r|V_r) + P(E_r|C_{\bar{r}})\,P(C_{\bar{r}}|V_r)$$

$$(2)\quad P(E_r|V_{\bar{r}}) = P(E_r|C_r)\,P(C_r|V_{\bar{r}}) + P(E_r|C_{\bar{r}})\,P(C_{\bar{r}}|V_{\bar{r}})$$

$$(3)\quad P(R_r|V_r) = P(R_r|E_r)\,P(E_r|V_{\bar{r}}) + P(R_r|E_{\bar{r}})\,P(E_{\bar{r}}|V_{\bar{r}})$$

$$(4)\quad P(R_r|V_{\bar{r}}) = P(R_r|E_r)\,P(E_r|V_{\bar{r}}) + P(R_r|E_{\bar{r}})\,P(E_{\bar{r}}|V_{\bar{r}})$$

$$(5)\quad P(S_r, R_r|V_r) = P(S_r|V_r)\,P(R_r|V_r)$$

$$(6)\quad P(S_r, R_r|V_{\bar{r}}) = P(S_r|V_{\bar{r}})\,P(R_r|V_{\bar{r}})$$

The notation $P(S_r, R_r|V_r)$ indicates the probability that the system has classified the document as relevant and that the screener has also, given that the document is, in fact, relevant with respect to the verbalized request. The conditional probabilities listed above can be summarized

in two-by-two tables as shown in Figure 2. 10. Since each row adds to unity, one can easily fill in the additional items.

Note that the last two-by-two table in Figure 2. 10 is simply a distribution of the last column of the next to last table. We have

$$P(R_{\bar{r}} | V_r) + P(S_{\bar{r}}, R_r | V_{\bar{r}}) + P(S_r, R_r | V_{\bar{r}}) = 1$$

$$P(R_{\bar{r}} | V_{\bar{r}}) + P(S_{\bar{r}}, R_r | V_{\bar{r}}) + P(S_r, R_r | V_{\bar{r}}) = 1$$

The expected number of documents in each of the cells is simply the probability in the cell times the number of documents in the file which are relevant (or nonrelevant) with respect to the verbalized request.

Some important features of the model are the following:

(1) It provides a means for tying together the influences of the various components of the searching system so that measures of performance of the searching subsystem are derived from measures of performance of the components.

(2) It permits one to determine the effect of a change in performance of a component on the performance of the subsystem.

(3) The conditional probabilities used as performance measures can be identified with the customary measures of recall and precision as follows:

   a. $P(S_r, R_r | V_r)$ is the theoretical recall ratio for the system.

   b. Let N_r be the number of documents relevant to a verbalized request and $N_{\bar{r}}$ the number nonrelevant in the file. Then, the theoretical precision ratio is

$$\frac{N_r \cdot P(S_r, R_r | V_r)}{N_r \cdot P(S_r, R_r | V_r) + N_{\bar{r}} \cdot P(S_r, R_{\bar{r}} | V_{\bar{r}})}$$

Relevance with respect to verbalized request	Cumulative through encoding stage	
	$E_{\bar{r}}$	E_r
V_r	$P(E_{\bar{r}} \mid V_r)$	$P(E_r \mid V_r)$
$V_{\bar{r}}$	$P(E_{\bar{r}} \mid V_{\bar{r}})$	$P(E_r \mid V_{\bar{r}})$
	Cumulative through system output	
	$R_{\bar{r}}$	R_r
V_r	$P(R_{\bar{r}} \mid V_r)$	$P(R_r \mid V_r)$
$V_{\bar{r}}$	$P(R_{\bar{r}} \mid V_{\bar{r}})$	$P(R_r \mid V_{\bar{r}})$
	Cumulative through screening	
	$S_{\bar{r}}$	S_r
V_r	$P(S_{\bar{r}}, R_r \mid V_r)$	$P(S_r, R_r \mid V_r)$
$V_{\bar{r}}$	$P(S_{\bar{r}}, R_r \mid V_{\bar{r}})$	$P(S_r, R_r \mid V_{\bar{r}})$

Figure 2. 10. Cumulative probabilities through various stages of the retrospective search

The model can be described as a finite Markov chain with absorbing states. The fact that the cell probabilities (except for the screening process) are conditional only upon the previous step makes it possible to structure the model as a matrix of transition probabilities, as in Figure 2.11. The cell entries are for illustration only. Absorbing states are $R_{\bar{r}}$ and R_r. In canonical form, the matrix of Figure 2.11 can be written

$$\left[\begin{array}{c|c} Q & B \\ \hline 0 & I \end{array}\right]$$

where Q represents the matrix of transitions from nonabsorbing to nonabsorbing states, B represents the transitions from nonabsorbing to absorbing states, I is the identity matrix, and 0 is the zero matrix.

	$V_{\bar{r}}$	V_r	$C_{\bar{r}}$	C_r	$E_{\bar{r}}$	E_r	$R_{\bar{r}}$	R_r
$V_{\bar{r}}$	0	0	0.90	0.10	0	0	0	0
V_r	0	0	0.20	0.80	0	0	0	0
$C_{\bar{r}}$	0	0	0	0	0.70	0.30	0	0
C_r	0	0	0	0	0.20	0.80	0	0
$E_{\bar{r}}$	0	0	0	0	0	0	0.90	0.10
E_r	0	0	0	0	0	0	0.05	0.95
$R_{\bar{r}}$	0	0	0	0	0	0	1	0
R_r	0	0	0	0	0	0	0	1

Figure 2.11. Matrix of transition probabilities for the retrospective searching model.

Then, by the theory of finite Markov chains [11], the fundamental matrix is given by $(I - Q)^{-1}$. The results of most interest are the probabilities of $R_{\bar{r}}$ and R_r, given relevance or nonrelevance with respect to the verbalized requests. The first row of $(I - Q)^{-1} B$

yields $P(R_{\bar{r}} | V_r)$ and $P(R_r | V_r)$, while the second row yields $P(R_{\bar{r}} | V_{\bar{r}})$ and $P(R_r | V_{\bar{r}})$. The screening probabilities can be applied to $P(R_r | V_{\bar{r}})$ and $P(R_r | V_{\bar{r}})$ to complete the quantification of performance. This model appears in more descriptive form with a worked example in the Procedural Guide [12].

2.6 Some examples of specific measures

We approach the task of making recommendations concerning use of specific measures in specific situations with some apprehension. Not only is there a great deal of controversy over the usefulness of evaluation measures, but the criteria for the selection of a measure have not even been agreed upon. The recommendations that we make must therefore be considered tentative, hopefully subject to substantial improvement as additional research is done and experience gained.

The recommendations that follow have been selected on the basis of their ability to contribute to decision-making processes and on the basis of their operational feasibility. Again, the list is not intended to be comprehensive. Most measures reflect performance, some reflect cost, and a few reflect benefits -- under the present state of the art we simply do not have feasible ways of characterizing benefits. Reference codes, such as a.b. 1, refer to the identifications in Table 2. 2.

It is clear that the above selected measures do not all measure directly the characteristic of interest. In many cases they measure something that, hopefully, is related to that characteristic. Also, it is clear that detailed costs could be found for each function and that user opinion could be inserted as a measure at nearly all of the levels.

Table 2.2.

a. Storage

a.a Acquisition

Summary measures

- Number of titles acquired during period
- Total cost of acquisition activity
  - Fixed costs - space, equipment, administrative, etc.
  - Variable costs - wages, materials, etc.
- Distribution of purchase price per document

a.a.1 Responsiveness to requests from users

- Number of purchase requests received from users
  - Number of these ordered
  - Number rejected - with reasons
- Backlogs of orders, beginning and end of year
- Distribution of times, per order, from receipt of acquisition request to presentation to requester
- Number of items requested from the system which were not in the files
  - Number of these subsequently ordered

a.a.2 Selection of high demand documents

- For a convenient period, say three years, the distribution of demand for 12 months after acquisition
- For an identified user population, percentage of its literature uses (references cited) which are in the file
- Distribution of these percentages by age of document

a.a.3 Ordering mechanics

- Distribution of elapsed time from receipt of request to purchase order
- Distribution of elapsed time between receipt of document and release to user
- Distribution of elapsed time between placing orders and receiving ordered items
- Cost per order filled (with subdivisions, as necessary, to reflect high cost tasks)

Table 2.2 (Continued)

a.b Files

Summary measures

- Number of items in storage
- Distribution of demand per item in storage
- Average storage cost per item stored (subdivided, as needed)

a.b.1 Composition of files

- Percent of user's needs met by items in the file
- Percent of demanded items "out of file" at time demanded
- Distribution of demand by title and age of document

a.b.2 Location of files

- Distance from principal user groups
- Average delay in receipt of requests from the file
- Average delay in receipt of requests from principal competing sources

a.b.3 Organization of files

(subjective judgments concerning assessibility, filing by author, subject matter, chronology, and so on)

b. Identification

b.a Indexes

Summary measures

- Total number of documents indexed during planning period
- Average total retrieval
- Average proportion of relevant titles not retrieved
- Average proportion of nonrelevant titles retrieved

b.a.1 Indexing procedures

- Proportion of terms chosen, given they should be
- Proportion of terms chosen, given they should not be
- Consistency measured by randomly paired indexing

Table 2.2 (Continued)

a.b.2 Index structure

- Number of terms in term list
- Number of hierarchies permitted
- Average number of terms selected per document
- Average number of facets selected per document
- Correlations among selected terms

b.b.1 Entry vocabularies

- User time per search
- Proportion of terms chosen correctly
- Number of terms chosen per search query by categories of logical structure
- Number of failures to retrieve as a result of use of incorrect terms

b.b.2 Assistance

- Cost of intermediary per search
- Proportion of terms chosen correctly by intermediary
- Number of terms chosen by intermediary per search query by categories of logical structure
- Average intermediary query processing time per search
- Opinions of users with respect to use of intermediaries

b.b.3 Procedures

- Proportion of search errors due to improper use
  - Of term list
  - Of equipment
  - Of intermediaries
- Average delay time per search
- Average cost per search

b.b.4 Equipment

- Cost of equipment depreciated over planning period
- Processing accuracy
- Processing time

Table 2.2 (Continued)

b.b.5 Screening

- Cost of document representation (preparation, storage, processing, output)
- Cost of intermediary screening
- Cost of user screening

b.b.6 Interaction with system

- Number of queries per search
- Average searching time per search
- Proportion of used documents retrieved from system
- Proportion of used documents not retrieved from system

b.c Selective dissemination

Summary measures

- Number of titles disseminated
- Number of full text disseminations

b.c.1 Screening

- Proportion of relevant documents correctly chosen
- Proportion of nonrelevant documents chosen
- Cost of selective dissemination function, suitably subdivided
- Average age of documents disseminated

b.c.2 Current awareness

- Average time from composition to identification
- Cost to prepare file
- Cost of alternative dissemination forms
- Cost per dissemination per user
- Number of items sent per dissemination per user
- Number of users
- Proportion of titles disseminated that are used (immediately, subsequently)
- Proportion of titles not disseminated that are used (immediately, subsequently)

Table 2.2 (Continued)

b.c.3 Use of document representations

- Number of titles disseminated over planning period
- Number of other disseminations over planning period
- Cost to prepare and process document representations
- Cost to store alternative document representations
- Proportion of relevant documents correctly chosen from representations
- Proportion of nonrelevant documents chosen from representations

c. Presentation

Summary measures

- Number of documents presented during planning period
- Cost (or price) per item presented

c.a Form

- Distribution of size per document
- Rating of presented documents

c.b Timeliness

- Average age of documents presented
- Distribution of time between identification and presentation

---

2.7 Some examples of the use of measures

This section contains two examples of the use of measures in evaluation.

2.7.1 An example in the evaluation of acquisition

Generally speaking, the principal reasons for storing documents at a particular location are to reduce document transmission time and to provide a means of identifying documents on the shelf (to provide browsing). Here we are concerned with the former reason. Acquisition implies a decision whether to acquire a copy of a title (or set of titles) in anticipation of use or to wait and order a copy upon demand.

The main benefit of acquiring a copy ahead of actual demand is that subsequent requests can be satisfied with little delay time. On the other hand, additional costs are necessary to provide the increased service. These costs generally are attributed to acquisition, processing and storage of all documents, whether or not they are actually requested.

If users are part of the proprietary system, the main benefit is directly determined from reduction in transmission time. However, if users are not part of the proprietary system, the main benefit is measurable by income if a price is charged for purchasing (or borrowing a copy), or total demand if a price is not charged. Price is partially dependent on cost per request and partially on income per request. Increased demand should reduce the cost per request and increase income, thereby reducing the price, which in turn should increase demand. Thus, an important synergistic effect must be considered. Motivation to use the storage system is dependent on price and user satisfaction*, and fulfillment time is probably the principal ingredient of user satisfaction.

It is clear that prior acquisition of high demand documents will improve a system by decreasing acquisition and storage costs per use or by decreasing average access time, where the average is taken over all requests. An example is given for a document selection policy in which either a document is acquired prior to demand in anticipation of demand or a document is acquired on demand. The hypothetical policy assumes that demand can be estimated for each document under question by regression analysis** or some other suitable technique.

---

* Promotion, advertising, and sales techniques may also affect motivation to use the system.

** An example of a regression analysis utilized to estimate demand for individual documents from document characteristics (subject, sources, age, and so on) is given by King et. al.[13] although that analysis was applied for a different purpose.

Hypothetical values for such a distribution for a single document are given in Table 2.3, and it is presumed that one wishes to decide whether or not to acquire that document. For the purpose of the example, assume that the acquisition cost is $6.00 per document and the request processi cost is $1.00 per document requested. Thus, acquisition prior to demand requires one to invest $6.00 in order to have the document available for anticipated demand, that is, to save request fulfillment time. Also, assume that it takes 3 days to process a request from storage and 14 days to process a request for an item that must be purchased.

Table 2.3 Hypothetical example of costs and time delays under two document acquisition decisions

Estimated number of requests per year	Probability estimated from regression	Cost		Time delay in days	
		To acquire prior to demand	To acquire on demand	To acquire prior to demand	To acquire on demand
0	0.50	$ 6.00	0	0	0
1	0.20	7.00	7.00	3	14
2	0.15	8.00	8.00	6	17
3	0.10	9.00	9.00	9	20
4	0.05	10.00	10.00	12	23
Weighted averages		$ 7.00	$ 4.00	3.0	8.5

On the average (over all similar documents), it is expected to cost $7.00 per request to acquire ahead of time and $4.00 to acquire on demand. Average delay time is expected to be 3 days if acquired ahead of demand and 8.5 days if acquired on demand. Therefore, the policy of acquiring before demand costs the system $3.00 per request to save 5.5 days of delay time, a cost of about $0.55 per day saved. Whether or not this is worthwhile is a value judgment to be provided by management.

In order to evaluate the overall decision policy, one can compute total costs and time by summing expected costs and expected time over all documents. Alternative policies can be evaluated by changing the cost and delay times to correspond with those of the alternatives being tested.

2.7.2 An example in evaluation of indexing procedures

It is assumed that one can make judgments concerning the validity of the assignment of terms to documents and that a "standard indexing" can be established. Relative frequencies of indexing consequences are symbolized in Table 2.4.

Table 2.4 Categorization of indexing errors

	Standard Indexing	
Actual Indexing	Should not be indexed	Should be indexed
Not indexed	$\hat{P}_0$	$\hat{P}_1$
Indexed	$\hat{P}_2$	$\hat{P}_3$
	$\hat{P}_0 + \hat{P}_2 = 1$	$\hat{P}_1 + \hat{P}_3 = 1$

The $\hat{P}_i$'s are relative frequencies of errors (or correct indexings) aggregated over collections of terms, indexers, and documents, under the assumption that these aggregations yield meaningful and interpretable results. If the relative frequencies approach limits as the number of observations becomes large, it is meaningful to interpret these limits as conditional probabilities.

Clearly, it is desirable to be able to translate the relative frequencies in Table 2.4 into the conditional probabilities given in

Figure 2.9, which in turn can be incorporated into a model to predict search accuracy. Consider a k-term search query requiring that documents be indexed by all k terms in order to be retrieved. Let P_2 and P_3 be "true values" of $\hat{P}_2$ and $\hat{P}_3$, averaged over all indexers, terms, and documents in the file. Let Q_j denote the portion of the entire file ($X_{j.}$ in the notation of Figure 2.7) which <u>should</u> contain j of the k terms in the search query. If one assumes independence of indexing errors from term to term, the following values of Figure 2.7 may be estimated:

$$\text{Total retrieved documents } (\overline{X}_{1.}) = X_{..} \sum_{\substack{\text{all values of } n_2,\ n_3 \text{ such that} \\ n_2 + n_3 = k}} P_2^{n_2} P_3^{n_3} Q_{n_3}$$

$$\text{Number of relevant missed documents } (\overline{X}_{22}) = X_{..} [Q_k\ 1-P_3^{k}]$$

$$\text{Number of non-relevant retrieved documents } (\overline{X}_{11}) = X_{..} \sum_{n_3 < k} P_2^{n_2} P_3^{n_3} Q_{n_3}$$

Other values of cell entries of Figure 2.7 can be obtained arithmetically.

An example illustrates the use of the models above, as well as demonstrating how the indexing information can be applied in the comprehensive retrospective searching model discussed previously. Suppose that one is interested in whether or not to have indexing reviewed as an indexing practice. One would expect greater indexer accuracy as a result of indexing review but would expect cost of indexing to be increased. An indexing experiment was performed at the U. S. Patent Office to answer this question [14, 15]. Results of this experiment are given in Table 2.5.

Table 2.5 Relative frequencies $\hat{P}_3$ and $\hat{P}_2$ and indexing time for two indexing procedures

Indexing procedure	$\hat{P}_3$	$\hat{P}_2$	Average indexing time per document
Single indexer	0.69	0.0014	64.3 min.
Single indexer reviewed	0.95	0.0002	111.6 min.

Applying these values to the equations shown, we obtain the following estimates (that were validated by means of a search experiment) for the cell entries in Figure 2.7.

Table 2.6 Observed cell entries (Figure 2.7) for search results for two indexing procedures

Retrieval category	Single indexer	Single indexer reviewed
Non-relevant retrieved ($\overline{X}_{\cdot\cdot}$)	4.5	5.6
Relevant retrieved ($\overline{X}_{12}$)	18.2	21.1
Non-relevant not retrieved ($\overline{X}_{21}$)	3592.7	3591.6
Relevant not retrieved ($\overline{X}_{22}$)	9.6	6.7

Following are the estimates of the conditional probabilities used in the retrospective searching models:

	Single indexer	Single indexer reviewed
$P(R_{\bar{r}} \mid E_r)$	0.345	0.241
$P(R_r \mid E_r)$	0.655	0.759
$P(R_r \mid E_{\bar{r}})$	0.0013	0.0016
$P(R_{\bar{r}} \mid E_{\bar{r}})$	0.9987	0.9984

Assume that the remaining hypothetical transition probabilities in Figure 2.11 are appropriate for the other searching processes and that hypothetical screening probabilities are:

$$P(S_{\bar{r}} \mid V_r) = 0.05 \qquad P(S_{\bar{r}} \mid V_{\bar{r}}) = 0.80$$
$$P(S_r \mid V_r) = 0.95 \qquad P(S_r \mid V_{\bar{r}}) = 0.20$$

Search performance measures that were computed from the retrospective searching model are shown in Table 2.7.

Table 2.7 Computed search performance measures for two indexing procedures*

Performance measure	Single indexer	Single indexer reviewed
Average recall ratio	0.46	0.53
Average number of relevant documents retrieved	12.2	14.1
Average precision ratio	0.069	0.068
Average total retrieval	165	192

The cost of indexing review is nearly 75% higher than indexing without review (112 minutes vs. 64 minutes). On the other hand, approximately two more relevant documents are retrieved when indexing is reviewed than when it is not reviewed (14.1 vs. 12.2). The retrieval cost per relevant document retrieved is about the same for the two indexing processes, as indicated by nearly equal average precision ratios. Therefore, management decisions regarding the two indexing processes would probably depend largely on (1) the total number of documents indexed, which determines

* Computational methodology for the retrospective searching model is discussed in detail in the Procedural Guide [12].

the total indexing costs of the two processes, and (2) the total number of searches conducted per year. For instance, if the total difference in indexing cost was $2000 and the number of searches conducted per year was 500, the indexing review process would cost $4 per search plus additional screening and handling costs which might not be worth the additional two relevant documents retrieved.

The advantage of having models for the interpretation of alternative indexing procedures is clearly evident. Similarly, other procedures can be investigated, such as specifying the time that should be expended in indexing a document or specifying a change in the professional level of indexers. In all instances, one should be able to compare cost and search performance for all of the alternative indexing procedures by the use of the evaluative models.

Another important question is how many index terms should be used on the average, i. e. the depth of indexing. The effects of the depth of indexing can also be determined by the models stated above. Thus, one should be able to compare, say, two levels of depth of indexing with regard to cost and search performance.

It is expensive to develop a sophisticated index language for vocabulary control in information retrieval. The more sophisticated the index language, on the whole, the more expensive it will be to apply and maintain. One important economic consideration is the size of the vocabulary. The larger the number of index terms in the vocabulary (i. e., the greater the number of document classes that can be uniquely defined), the greater its specificity and the greater the precision capabilities of the system. However, a large vocabulary is costly to develop, costly to apply, and costly to update. The specificity of the vocabulary must be related directly to the specificity of the requests made to the system. This implies the strong economic necessity for conducting a careful analysis of representative requests during system design.

On the other hand, the development of a rich, readily accessible entry vocabulary is likely to reduce the costs of the indexing operation by reducing the amount of intellectual effort involved in the indexing process. In the early stages of indexing documents into a new retrieval system, many intellectual decisions have to be made. As these decisions are recorded in the entry vocabulary, the intellectual burden on subsequent indexers is reduced.

Another factor to be considered is that of the number and variety of index language devices included in the vocabulary to improve its search performance. In particular, various devices designed to improve search performance are expensive to apply in information retrieval systems. Examples are links and roles, subheadings, and term weighting. To be economically justifiable, it must be shown that the added input and manipulation costs involved in the use of these devices is offset by appreciable savings in screening time at output.

Of course, in the operation of a document retrieval system, one can use a carefully controlled index language carefully applied at the time of indexing, thus minimizing time and costs at the searching stage. Alternatively, one can adopt a rather free indexing, with little vocabulary control, and expend additional effort (by the use of sophisticated searching aids) at the output stage. This matter of relative weight given to input effort as opposed to output effort is an extremely important one to consider in the economic evaluation of a document retrieval system because it affects indexing policy and practice, the system vocabulary, and searching strategies and procedures. One factor to be considered is the volume of documents input in relation to the volume of requests handled. If many documents are indexed but comparatively few requests are handled, it is usually sensible to economize at the input stage and to expend more time at output. In the opposite conditions (few documents, many requests), the reverse could be true. However, another important factor to be considered is the need to

save time at output, that is, to determine the response time requirements of system users. All of the options available to the various input and searching processes can be evaluated by means of the retrospective searching model, and insight can be gained by applying the model.

References

[1] Cuadra, Carlos, Robert V. Katter, Emory H. Holmes and Everett M. Wallace, (June 30, 1967), "Experimental Studies of Relevance Judgments: Final Report," Santa Monica, California, System Development Corporation.

[2] Rees, Alan M. and Douglas G. Schultz, (June 30, 1967), "A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching: Final Report," Cleveland, Ohio, Case Western Reserve University, Center for Documentation and Communication Research.

[3] Swets, John A., (June 15, 1967), "Effectiveness of Information Retrieval Methods," Cambridge, Mass., Bolt Beranek and Newman, Inc., AFCRL-67-0412.

[4] Lancaster, F. W., (1968), "Interaction Between Requesters and a Large Mechanized Retrieval System," Information Storage and Retrieval, Vol. 4, pp. 239-252.

[5] O'Connor, John, (July 1967), "Relevance Disagreements and Unclear Request Forms," American Documentation, Volume XVIII, pp. 165-177.

[6] Keen, E. M., (January 1968), "Search Strategy Evaluation in Manual and Automated Systems," ASLIB Proceedings, Vol. 20, No. 1, pp. 65-81.

[7] St. Laurent, Mary Cuddy, (February 1967), "A Review of the Literature of Indexer Consistency," University of Chicago, Graduate Library School, Grant NSF-GN-380.

[8] Katter, Robert V., (August 31, 1967), "Study of Document-Representations; Multidimensional Scaling of Indexing Terms: Final Report," Santa Monica, California, System Development Corporation, TM-3627.

[9] Dym, Eleanor D., (1967), "Relevance Predictability. I: Investigation Background and Procedures," in Electronic Handling of Information: Testing and Evaluation, Allen Kent, Orrin E. Taulbee, Jack Belzer, and Gordon D. Goldstein, editors, Washington, D. C., Thompson; London, Academic Press, pp. 175-185.

[10] Kent, Allen, J. Belzer, M. Kurfeerst, E. D. Dym, D. L. Shirey, and A. Bose, (April 1967), "Relevance Predictability in Information Retrieval Systems", Methods of Information in Medicine, pp. 45-51.

[11] Kemeny, John G. and J. Laurie Snell, (1960), Finite Markov Chains, D. Van Nostrand Company, Inc., Princeton.

[12] Westat Research, Inc., (December 31, 1968), "Procedural Guide for the Evaluation of Document Retrieval Systems," Report to the National Science Foundation, Washington, D. C.

[13] King, D. W., E. C. Bryant, J. M. Daley and D. T. Searls, (February 1965), "A Decision Model for Determining Document Printing Quantities," Bethesda, Md.: Westat Research, Inc., PB 167 175.

[14] King, D. W., (May 1965), "Evaluation of Coordinate Index Systems During File Development," Journal of Chemical Documentation, Volume V, No. 2, pp. 96-99.

[15] Bryant, E. C., D. W. King and P. J. Terragno, (August 1963), "Analysis of an Indexing and Retrieval Experiment for the Organometallic File of the U. S. Patent Office," WRA PO 10, Bethesda, Md.: Westat Research, Inc., PB 166 488.

PART III

TECHNICAL PAPERS

Technical papers in this part provide theoretical background for some of the methodology presented in Part II and in the Procedural Guide.

The first paper is concerned with estimation of recall, but is generally applicable to the estimation of any of the various conditional probabilities one encounters in evaluation of document retrieval systems. (See particularly Chapter 3 of the Procedural Guide.) The problem of reliability of estimates has also been addressed. A somewhat simpler approach to variance estimation has been advocated in the Procedural Guide which should yield satisfactory approximations in applied cases.

Also, in the same paper, operating characteristic curves have been advocated which rely on the principles of probit analysis from the field of bioassay. The methodology appears to have considerable merit, but to our knowledge has not been applied in actual cases.

The second paper discusses search characteristic curves which relate the number of wanted documents retrieved to the retrieval effort. The applicability of the generalized Beta distribution to this problem is discussed and comparisons are made with the work of Swets and the recall and precision concepts of Cleverdon.

The third paper provides theoretical support for the second paper with respect to the Beta distribution.

The fourth paper looks at the classification of search results from the information theoretic viewpoint and provides a methodology which parallels that of analysis of variance in its flexibility.

The fifth paper discusses economic evaluation from the standpoint of costs and benefits. Some mathematicl models are presented and some

examples of their use with hypothetical data are given. It is generally recognized that the principal weakness in the benefits-costs approach to document retrieval evaluation lies in the subjective nature of the estimates of benefits. There may be some advantage, however, in performing such analyses to demonstrate the magnitudes of benefits which are necessary to produce a net benefit over cost. Analysis of the problem in this context may be sufficient to provide management with the decision framework which it needs.

3.1 SOME ESTIMATION PROBLEMS ASSOCIATED WITH EVALUATING INFORMATION RETRIEVAL SYSTEMS

by

R. H. Shumway

3.1.1 Introduction

In the evaluation of large scale information retrieval systems, the collection of summary statistics describing the performance of the system is fairly standard. A number of such studies have been carried out and are reported by Cleverdon et. al. [2], Giuliano and Jones [5], Salton [9], and Lancaster [7]. Most of these evaluations have made use of the recall and precision ratios as indexes of performance for an operating system where the recall (ratio) is defined as the proportion of the relevant documents retrieved and the precision (ratio) is the proportion of retrieved documents which are relevant.

Many investigators have pointed out that while the recall indicates the coverage of the relevant literature achieved by the search results, the precision serves as an indicator of the "richness" of the retrieved documents. For example, by retrieving 100 percent of the document collection, one may easily guarantee 100 percent recall at the expense of extremely low precision. A usual assumption is that the richness of the retrieved documents should decrease as one looks through a ranked set of documents. Cleverdon et. al. [2] have described this pattern as an inevitable relationship between precision and recall, and the plots relating increasing recall to decreasing precision for various systems and search strategies form a set of useful evaluation measures. Furthermore, if the recall is a heuristic measure of the value of a retrieval scheme, then the precision is a rough measure of the cost since it gives the ratio of the relevant documents obtained to the total retrieval that one must examine. Hence, both measures are usually deemed necessary.

Decision theory can also provide another simultaneous measure of cost and benefit through the following general model. Let the conditional probabilities of retrieving or not retrieving relevant or nonrelevant documents be specified by the two-way table

		Retrieved r	Not Retrieved $\bar{r}$	
Relevant	R	$P(r \mid R)$	$P(\bar{r} \mid R)$	1
Not Relevant	$\bar{R}$	$P(r \mid \bar{R})$	$P(\bar{r} \mid \bar{R})$	1

where, for example, $P(\bar{r} \mid \bar{R})$ denotes the probability that a document is not retrieved, given that it is not relevant. Associated with the above table of probabilities is a corresponding table of costs denoted by

		Retrieved r	Not Retrieved $\bar{r}$
Relevant	R	$C(R, r)$	$C(R, \bar{r})$
Not Relevant	$\bar{R}$	$C(\bar{R}, r)$	$C(\bar{R}, \bar{r})$

so that $C(R, \bar{r})$, for example, is the cost of not retrieving a relevant document. The cost $C(R, r)$ is to be interpreted as a negative cost, or equivalently as a benefit, accruing from retrieving a relevant document. Then, if the prior probabilities of relevance and nonrelevance are specified as $P(R)$ and $P(\bar{R})$, the expected Bayes' cost is given by

$$C = P(R)\, P(r \mid R)\, C(R, r) + P(R)\, P(\bar{r} \mid R)\, C(R, \bar{r}) + P(\bar{R})\, P(r \mid \bar{R})\, C(\bar{R}, r) + P(\bar{R})\, P(\bar{r} \mid \bar{R})\, C(\bar{R}, \bar{r}) \qquad (1)$$

The difficulty with this approach is the necessarily arbitrary nature of any cost figures which might be assigned. The cost of retrieving a nonrelevant document can probably be taken to be proportional to the number of documents retrieved since the cost of getting the relevant documents is incurred by searching the retrieved set. However, the cost figures appearing in the

other cells could vary from search to search and from investigator to investigator depending upon the nature or purpose of the search.

Some thought, therefore, has been devoted to the construction of a single measure reflecting both the value of the relevant documents retrieved and the cost (expressed in simple terms) of retrieving those relevant documents. The retrieval profile of Giuliano and Jones [5] plots the recall against the total retrieval. This enables one either (1) to estimate the recall achieved for a fixed cost in documents retrieved, or (2) to specify a recall and estimate the cost in documents retrieved necessary to achieve that recall. In this way, a "retrieval profile" or "search characteristic" curve is a single measure of performance and represents a convenient way of comparing competing methods or systems. In addition, its relative simplicity recommends it as a reference tool for searching strategy.

In this discussion we shall concentrate on some sampling and estimation problems associated with the classical recall measure appearing in the search characteristic curve. We develop the mean and variance of an estimate for recall and specify a confidence interval for the estimate. Also, in Section 3.1.2, methods for combining single estimates for recall into overall recall estimates are derived. Section 3.1.3 uses the combined recall estimates to develop a nonlinear model for the classical search characteristic curve. Then, borrowing a general technique from probit analysis, confidence intervals for search characteristic curves for different systems are developed. These intervals enable a user to specify a certain retrieval cost which he will tolerate in order to achieve a given recall. One obtains as a residual benefit a procedure either for making a statistical comparison between two systems or for determining the search strategy which produces the optimum yield. We will proceed initially with the estimation procedure for recall.

3.1.2 Estimation of the recall ratio

The recall ratio is defined as the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection.

Thus, if we suppose that there are N_R relevant documents in the collection and n relevant documents are retrieved on a particular search the value of the recall ratio for that search is

$$R = \frac{n}{N_R} \qquad (2)$$

While any search experiment yields a value for n, it is difficult to assign a number to N_R without exhaustively searching the file.

A possible technique has been used by D. W. King (see Atherton [5]) and involves identifying externally a subset of relevant documents within the system before the search is performed. For a variant on the above technique refer to Lancaster [7]. An example of an externally identified set of documents is the set obtained from a group or bibliographic references. After the initial set is specified, the search is performed and the relevant documents which appear in both the externally identified and the retrieved set are counted. Then, if we suppose that n_R of the relevant documents can be identified externally and that k of these n_R documents appeared in a search of the system, an estimate of the recall ratio is given by

$$\hat{R} = \frac{k}{n_R} \qquad (3)$$

This is simply the recall ratio of the externally identified set. In the example above it would be the proportion of documents in the bibliography which appeared in the retrieval. The estimate can be justified by appealing to techniques used in wildlife marking where an initial capture of a species is marked and freed later to be captured as members of a new sample which yield an estimate of the total species population (Feller [3]). To summarize the conditions, suppose that of the N_R relevant documents in the file n_R can be identified in advance. The retrieval of n documents is a sub-sample of the original N_R documents where the original N_R documents are divided into two groups: those initially identified and those not initially identified. Figure 3.1 represents schematically the sampling procedure.

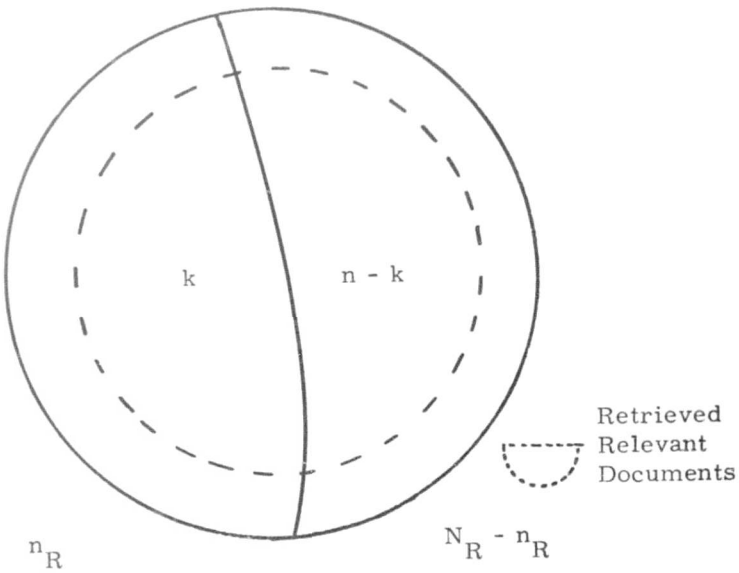


Figure 3.1. A sample of n relevant documents retrieved from a population of n_R initially identified relevant documents and $N_R - n_R$ not initially identified relevant documents.

Now, a sample of n relevant documents is retrieved of which k are in the group originally designated as relevant and n-k are in the group not previously identified. The probability of this event is

$$P_k(N_R) = \frac{\binom{n_R}{k}\binom{N_R - n_R}{n-k}}{\binom{N_R}{n}} \tag{4}$$

In a typical retrieval experiment, we know n_R, n and k with the problem being to determine N_R, the total number of relevant documents, and the recall ratio (2). We determine N_R (Feller [3]) by finding the value which maximizes the likelihood $P_k(N_R)$ given by Equation (4). This gives

$$\hat{N}_R = \frac{n_R\, n}{k} \tag{5}$$

Therefore, the maximum likelihood estimate of the recall is

$$\hat{R} = \frac{n}{\hat{N}_R} \tag{6}$$

or from Equation (5) we obtain the King-Lancaster estimate given in Equation (3).

Confidence limits on R may be derived through the following method. We may calculate, using Equation (4) for given k, n, n_R, the values of N_R, say ℓ_1 and ℓ_2, such that

$$\sum_{m=0}^{k} P_m(\ell_1) < \frac{\alpha}{2} \tag{7}$$

$$\sum_{m=k}^{k} P_m(\ell_2) < \frac{\alpha}{2} \tag{8}$$

so that the probability of observing k or less overlaps when $N_R = \ell_1$ is $< \frac{\alpha}{2}$ while the probability of observing k or more overlaps when $N_R = \ell_2$ is $< \frac{\alpha}{2}$ so that

$$P_F(\ell_1 < N_R < \ell_2) = 1 - \alpha$$

is a $1 - \alpha$ confidence interval for N_R. Then by Equation (2) we may write

$$P_F\left(\frac{n}{\ell_2} < R < \frac{n}{\ell_1}\right) = 1 - \alpha \qquad (9)$$

To illustrate the procedure, suppose that a preliminary scheme yielded $n_R = 4$ identified relevant documents while the retrieval contained $n = 3$ relevant documents of which $k = 2$ were already specified in the preliminary procedure. Then, using Equation (7) and (8) in conjunction with (4) we have

$$\sum_{m=0}^{2} P_m(4) = 0.00$$

$$\sum_{m=2}^{3} P_m(27) = 0.05$$

Since $\sum_{m=0}^{2} P_m(5) = 0.6$, which is greater than 0.05 one must use $\ell_1 = 4$ rather than $\ell_1 = 5$. The confidence interval can be written, using Equation (9):

$$P\left(\frac{3}{27} < R < \frac{3}{4}\right) = 0.90$$

and the 90% confidence interval ranges from .11 to .75.

The estimate of recall is given by

$$\hat{R} = \frac{k}{n_R} = \frac{2}{4} = 0.50$$

This indicates that, for a small number of relevant documents, the confidence limits for the recall in a single search are rather broad.

We may develop the sampling properties of the estimator $\hat{R}$ by defining a random variable X_i such that

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ retrieved relevant document was initially identified as relevant} \\ 0 & \text{otherwise} \end{cases}$$

Then the estimator $\hat{R}$ may be written

$$\hat{R} = \frac{k}{n_R} = \frac{1}{n_R} \sum_{i=1}^{n} X_i \tag{10}$$

where we have

$$P(X_i = 1) = \frac{n_R}{N_R}$$

$$P(X_i = 1, X_j = 1) = \frac{n_R(n_R - 1)}{N_R(N_R - 1)} \tag{11}$$

Now, the estimator for R is unbiased since

$$E(\hat{R}) = \frac{1}{n_R} \sum_{i=1}^{n} EX_i = \frac{1}{n_R} \cdot n \cdot \frac{n_R}{N_R} = \frac{n}{N_R} = R$$

by Equation (2). Similarly, the estimate for the variance of the estimator depends upon applying (11) which yields

$$\text{var} \left(\sum_{i=1}^{n} X_i\right) = R \cdot n_R \left(1 - \frac{n_R}{n} R\right) \frac{N_R - n}{N_R - 1} \tag{12}$$

so that

$$\text{var } \hat{R} = \frac{R\left(1 - \frac{n_R}{n} R\right)(N_R - n)}{n_R (N_R - 1)}$$

$$\approx \frac{R(1 - R)\left(1 - \frac{n_R}{n} R\right)}{n_R} \tag{13}$$

if N_R is fairly large so that $N_R \approx N_R - 1$. An estimate for the standard deviation is

$$\hat{\sigma}_{\hat{R}} \sqrt{\frac{\hat{R}(1 - \hat{R})\left(1 - \frac{n_R}{n}\hat{R}\right)}{n_R}} \qquad (14)$$

Since the estimator R is a sum of random variables converging to the normal distribution, an estimated 95% confidence interval for R becomes

$$\hat{R} \pm 1.96\, \hat{\sigma}_{\hat{R}} \qquad (15)$$

As an illustration, with 100 relevant documents identified (n_R) and 200 relevant documents retrieved (n) of which k = 50 are in the identified group we have

$$\frac{50}{100} \pm 1.96 \sqrt{\frac{15\,(0.5)\,(1 - (0.5)^2)}{100}}$$

which yields an interval running from 0.41 to 0.59. A method for reducing the single search variance is to extend the above argument to cases where data from more than one search is available. This involves defining a random variable X_{ih} which is 1 or 0 according to whether the i^{th} document in the h^{th} retrieval appeared in the initially identified relevant set or not. We let n_R^h, n_h and N_R^h be the same parameters as before with the subscript (or superscript) h pertaining to the h^{th} search. The recall ratio for the h^{th} search is

$$R_h = \frac{n_h}{N_R^h}, \quad h = 1, 2, \ldots, L \qquad (16)$$

If recall estimates from L searches are to be combined into an overall estimate of recall we may use

$$\hat{R} = \frac{1}{n_R} \sum_{h=1}^{L} \sum_{i=1}^{n_h} X_{ih} = \frac{1}{n_R} \sum_{h=1}^{L} \hat{R}_h n_R^h \qquad (17)$$

where

$$n_R = \sum_{h=1}^{L} n_R^h$$

so that the estimate is simply the weighted average of the single search recall ratios. Hence, taking the expectation

$$E(\hat{R}) = \frac{1}{n_R} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{n_R^h}{N_R} = \frac{1}{n_R} \sum_{h=1}^{L} n_R^h \frac{n_h}{N_R^h}$$

$$= \sum_{h=1}^{L} \frac{n_R^h}{n_R} R_h \qquad (18)$$

where R_h is the recall ratio for the h^{th} search. The expected value of the estimate is a weighted average of the recalls for the separate searches. If the recall ratios are equal then $R_1 = R_2 = \ldots = R_L = R$ and $E(\hat{R}) = R$. The variance of the estimator R is given by applying Equation (12) to

$$\sigma^2_{\hat{R}} = \text{var } \hat{R} = \frac{1}{n_R^2} \sum_{h=1}^{L} \text{var} \left\{ \sum_{i=1}^{n_h} X_{ih} \right\}$$

$$\approx \frac{1}{n_R^2} \sum_{h=1}^{L} n_R^h R_h(1 - R_h)\left(1 - \frac{n_R^h}{n_h} R_h\right) \qquad (19)$$

with the confidence interval given as usual by $R \pm 1.96\ \sigma_R$ where σ^2_R is obtained by substituting the estimates R_h for R_h.

The calculations in the first part of this section show that estimates for recall using the results of a single search can be ma[illegible] but tend to be quite variable. If data summarizing a number of searches is available then the pooled estimators of this section are appropriate. If either the number of searches or the number of retrieved relevant documents is large then the normal approximation can be used to calculate 95% confidence intervals.

3.1.3 Search characteristic curves

In the introduction we argued for the adoption of the retrieval profile or search characteristic curve as a single measure of system

so that the estimate is simply the weighted average of the single search recall ratios. Hence, taking the expectation

$$E(\hat{R}) = \frac{1}{n_R} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{n_R^h}{N_R} = \frac{1}{n_R} \sum_{h=1}^{L} n_R^h \frac{n_h}{N_R^h}$$

$$= \sum_{h=1}^{L} \frac{n_R^h}{n_R} R_h \qquad (18)$$

where R_h is the recall ratio for the h^{th} search. The expected value of the estimate is a weighted average of the recalls for the separate searches. If the recall ratios are equal then $R_1 = R_2 = \ldots = R_L = R$ and $E(\hat{R}) = R$. The variance of the estimator R is given by applying Equation (12) to

$$\sigma_{\hat{R}}^2 = \text{var } \hat{R} = \frac{1}{n_R^2} \sum_{h=1}^{L} \text{var} \left\{ \sum_{i=1}^{n_h} X_{ih} \right\}$$

$$\approx \frac{1}{n_R^2} \sum_{h=1}^{L} n_R^h R_h(1 - R_h)\left(1 - \frac{n_R^h}{n_h} R_h\right) \qquad (19)$$

with the confidence interval given as usual by $R \pm 1.96\, \sigma_R$ where σ_R^2 is obtained by substituting the estimates R_h for R_h.

The calculations in the first part of this section show that estimates for recall using the results of a single search can be made but tend to be quite variable. If data summarizing a number of searches is available then the pooled estimators of this section are appropriate. If either the number of searches or the number of retrieved relevant documents is large then the normal approximation can be used to calculate 95% confidence intervals.

3.1.3 Search characteristic curves

In the introduction we argued for the adoption of the retrieval profile or search characteristic curve as a single measure of system

effectiveness. In this section we shall apply some statistical techniques which are of use in making confidence interval statements about the recall ratio and the number of documents one must retrieve to realize a specified recall.

Search characteristic curves are available for a number of kinds of information retrieval systems (Figures 3.1 to 3.3) as plots which represent the number of documents retrieved on the vertical axis with the recall ratio as the horizontal scale. The method of plotting the data here with both axes in transformed units implies a definite non-linear functional relationship between recall and the number of retrieved documents. The number of documents retrieved, n, is on a logarithmic scale with the recall expressed in units of the integrated normal distribution. This implies the relation

$$R(n;\ \alpha,\ \beta) = \int_{-\infty}^{\alpha + \beta \log n} e^{\frac{-u^2}{2}} \frac{du}{\sqrt{2\pi}} \qquad (20)$$

where n denotes the number of retrieved documents and $R(n;\ \alpha,\ \beta)$ denotes the recall achieved when n documents are retrieved. The parameters α and β are the intercept and slope of the straight line which results if n is plotted against recall on log probability paper and relationship (11) holds. If, for a particular system, we can estimate the parameters α and β, the recall can be estimated for a given number of retrieved documents or the number of documents retrieved can be estimated subject to a fixed value for the recall ratio. Figures 3.2 to 3.4 show search characteristic curves from several retrieval experiments with the linearity of the data indicating that the model implied by Equation (10) does a reasonable job of representing the data.

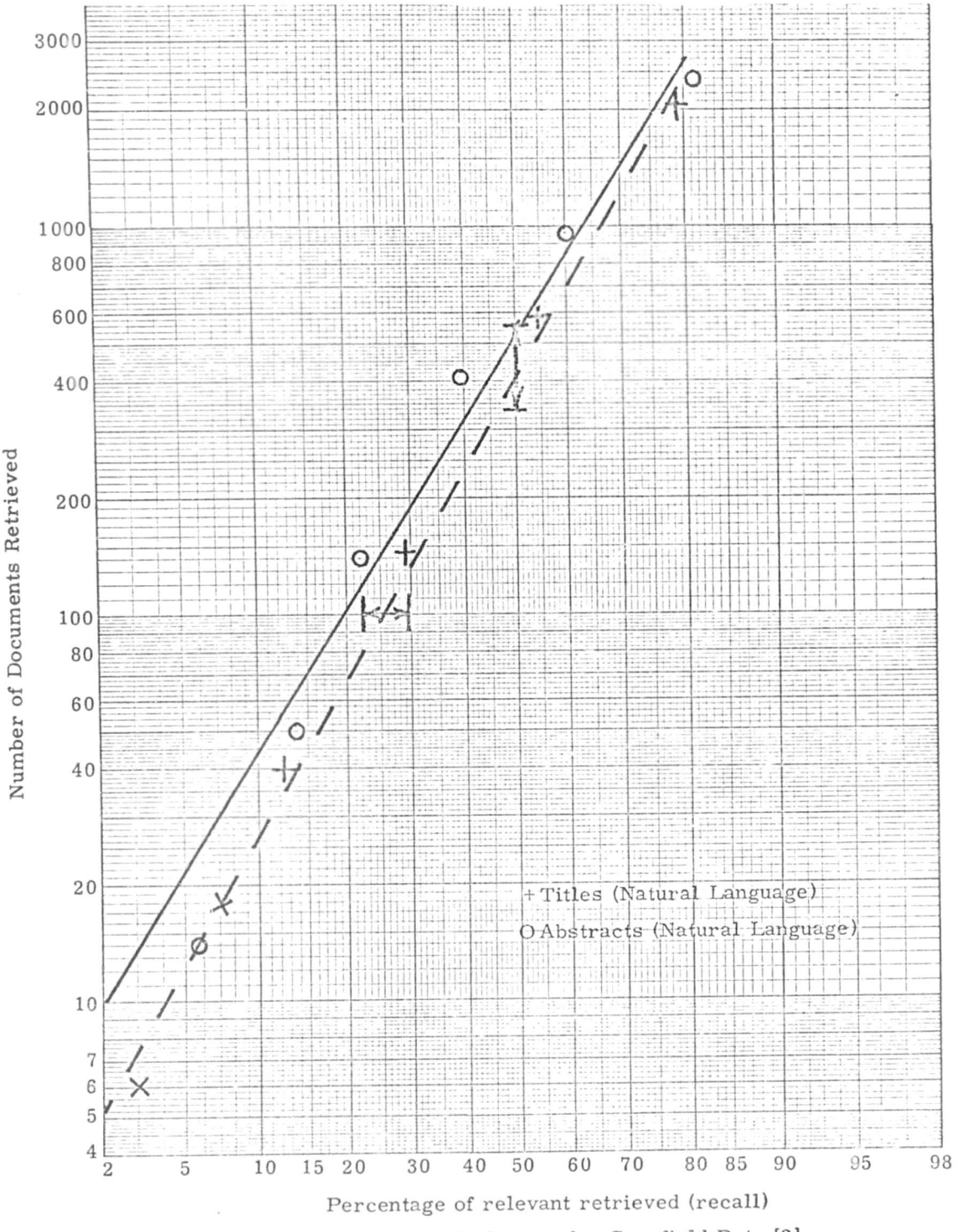


Figure 3.2. Performance Characteristic Curves for Cranfield Data [2]

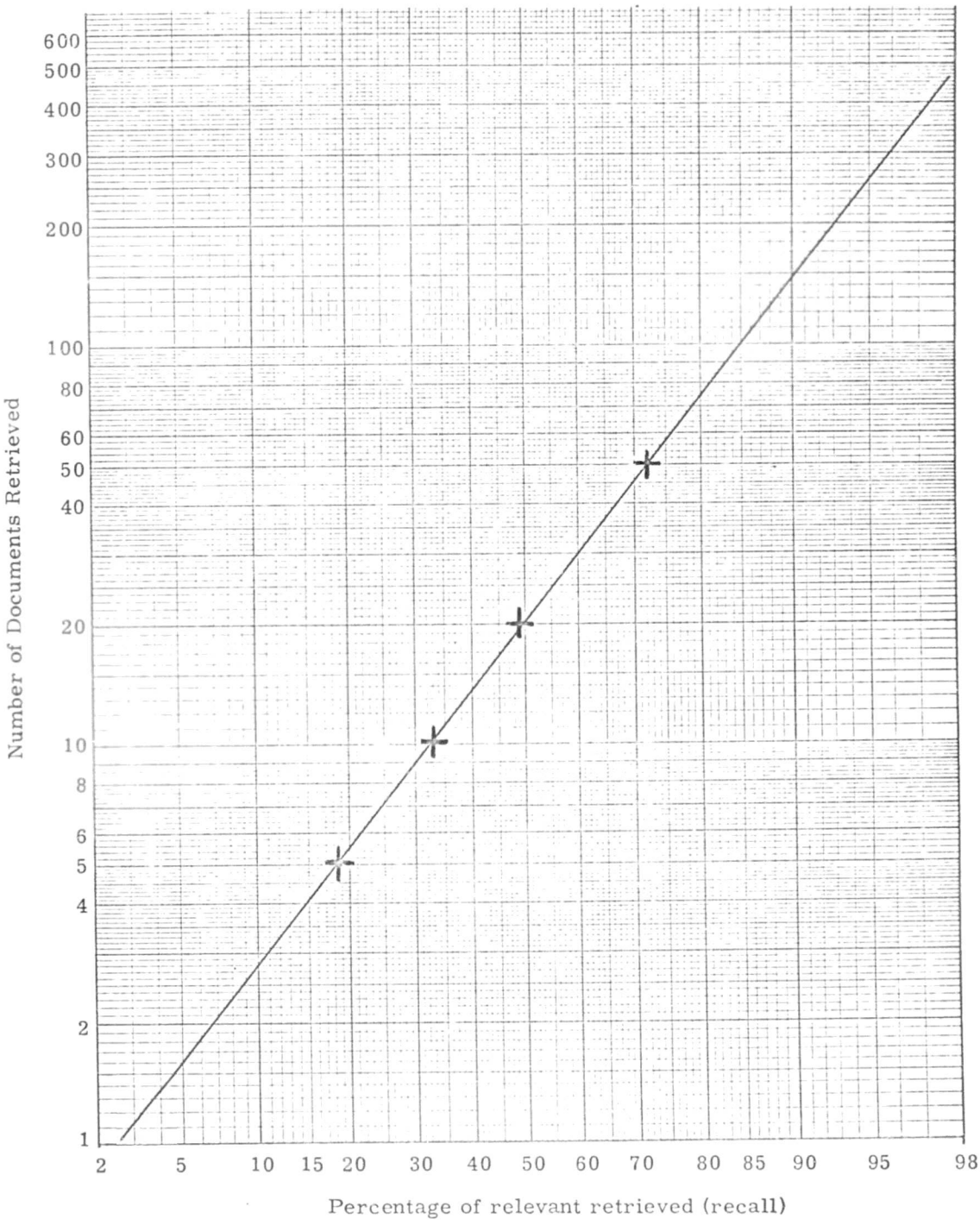


Figure 3.3. Performance Characteristic Curves for Patent Search, King [10]

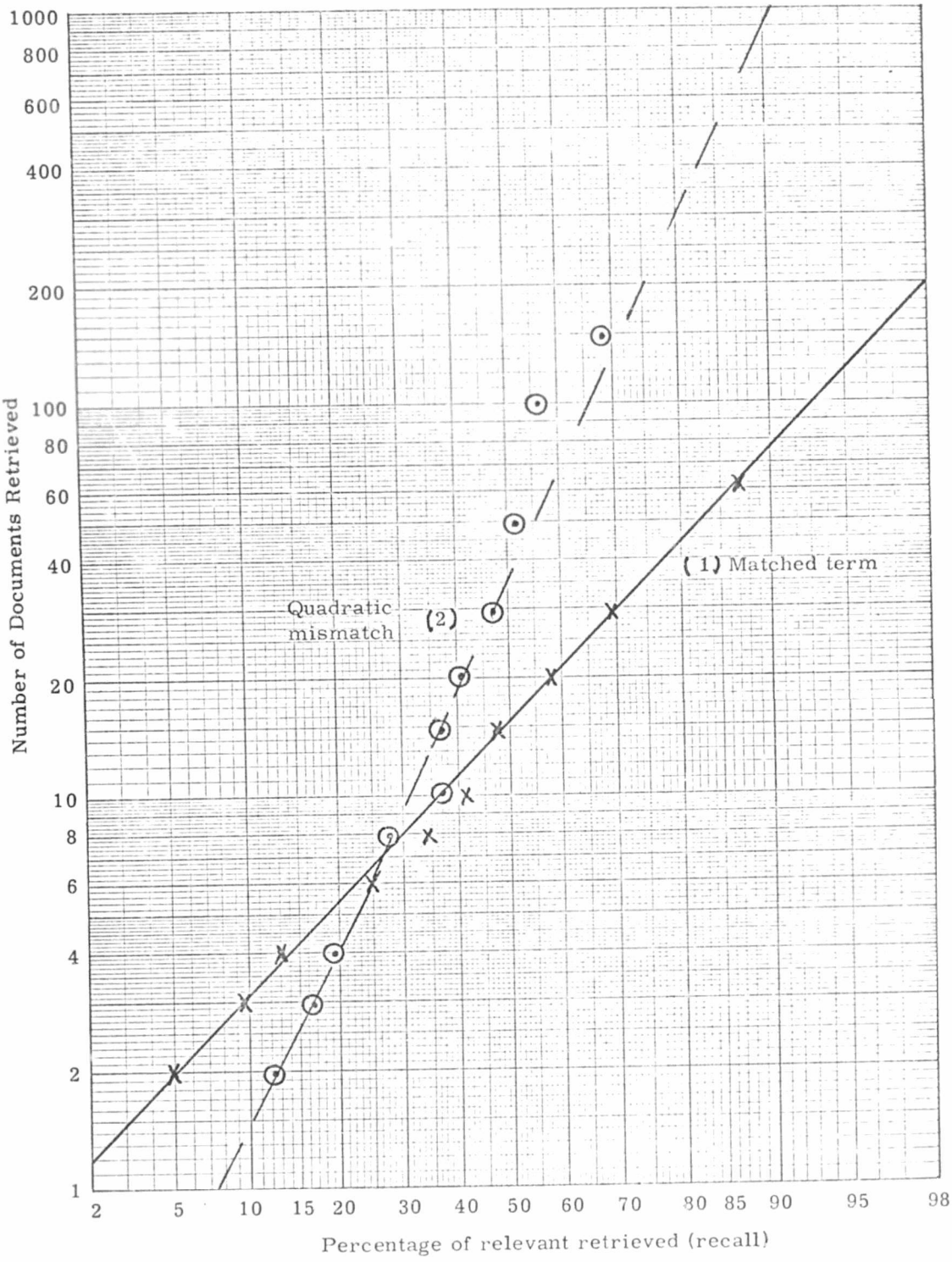


Figure 3.4. Performance Characteristic Curve for Associative Retrieval Experiment [1]

The general framework which leads to the m odel given in Equation (20) is most familiar to workers in the field of bio-assay who have made extensive use of probit analysis. In evaluating the response of an animal to dose administered at a given level one generally tests a group of animals at each of a number of doses and observes the proportion responding at each level [4]. The present discussion merely requires that one identify as "dose" the number of documents that one is willing to retrieve. Then the recall is identified as the proportion of relevant documents "responding" or retrieved by this particular dose. The analogy enables one to borrow directly a number of techniques used in probit analysis for direct application to the field of information retrieval.

We shall be concerned mainly with the estimation of the search characteristic curve and the associated uncertainty or statistical variability in the estimate. In particular, the questions which are most often asked deal with (i) Comparability of performance curves for different systems, (ii) Estimation of the recall achieved for a given retrieval expenditure and (iii) Estimation of the retrieval expenditure necessary to achieve a given recall.

As an example, the indexing-by title curve for the Cranfield retrieval experiments is given in Figure 3.1 (see Cleverdon et. al. [2]). The straight line fit to the data indicates the adequacy of the basic model and reference 3 provides the basic procedure for fitting the model by the probit method of analysis. The data for this particular search experiment are given in Table 3.1 on the following page.

The confidence limits are shown in Figure 3.1 and it can be seen that the Cranfield data yields fairly tight limits about the estimate. The data summarized in Table 3.2 in combination with search characteristic curves like Figures 3.1 to 3.3 can provide the user with simultaneous measures of cost and benefit for several competing methods or systems. In general, the recall measures roughly the benefit while the number of

Table 3.1. Search Characteristic Curve by Titles [2].

Number of Documents Retrieved	Number of Relevant Documents	Number of Relevant Documents Retrieved	Recall
2057	198	155	0.783
571	198	108	0.545
134	198	58	0.293
41	198	26	0.131
18	198	15	0.076
6	198	6	0.030
3	198	3	0.015

The basic procedure consists of fitting the line in Figure 3.1 by eye and then following the approximate probit analysis procedure given in reference 8. We have computed statistical measures associated with the search characteristic curve and presented them in Table 3.2.

Table 3.2. Summary of Some Performance Statistics for the Cranfield Data.

Characteristic	Estimate	95% Confidence Interval
* Number of Retrieved Documents Necessary to Achieve R = 0.50	420	341 - 518
** Recall Achieved by Retrieving 100 Documents	0.26	0.23 - 0.28
*** Recall Achieved by Retrieving 1,000 Documents	0.66	0.61 - 0.71

documents retrieved measures the approximate cost in user terms. The system cost will depend more upon the method utilized than on the number of documents retrieved. In general, search characteristic curves should be made available to the user who then specifies a cost in terms of the number of documents in a retrieved set which he would be willing to examine. If we suppose that he is willing to examine 100 documents the recall (from Table 3.2) should be between 23% and 28%. Then a search is performed with the retrieved relevant documents tagged by the user. These retrieved relevant documents are matched against a list of externally identified documents to provide estimates (by the method of Section 3.1.2) for the recall ratio and the number of missed relevant documents. Note that we also obtain confidence limits for the number of missed relevant documents and the recall ratio. If the user is satisfied with his retrieval the search is terminated, otherwise new estimates can be made for a subsequent increase in depth on the next search.

Search characteristic curves should be constructed using the pooled estimation procedure describe in the latter part of Section 3.1.2 if the complete examination of the unretrieved set is not practical. Otherwise, the estimate of recall made from a single search is subject to the high variability associated with the first part of Section 3.1.2.

3.1.4 Summary

We have considered here the rationale behind the use of the search characteristic curve as a simultaneous measure of cost and benefit in evaluating an information retrieval system. The measurement and modeling of the search characteristic curve were developed from techniques generally applicable in probit analysis. It was shown that estimates and confidence intervals for (1) recall given a restriction on retrieval or (2) total retrieval necessary to achieve a fixed recall could be developed. The estimation of the recall ratio using incomplete samples and a pre-specified set of relevant

documents was analyzed, and it was shown that this procedure has a high variability in the estimate for recall if single samples are used.

A procedure similar to the classical stratified estimation of proportions was used to develop a pooled estimate for recall with reasonable variance properties. Some examples were presented using the theoretical techniques on data which has appeared in the literature. References were given which present the probit computation in detail.

References

[1] Bryant, E. C., D. T. Searls, R. H. Shumway, and D. G. Weinman, (1967), "Associative Adjustments to Reduce Errors in Document Screening," AF49(638)-1484, Final Report.

[2] Cleverdon, Cyril and M. Keen, (1966), "Factors Determining the Performance of Indexing Systems," Vol. 2, Test Results, ASLIB Cranfield Research Project, Cranfield.

[3] Feller, W., (1950), Introduction to Probability Theory and Its Applications, John Wiley, p. 43.

[4] Finney, G. D. J., (1947), Probit Analysis, Cambridge University Press.

[5] Atherton, P., D. W. King, and R. Freeman, (1968), "Evaluation of the Retrieval of Nuclear Science Document References Using the Universal Decimal Classification as the Indexing Language for a Computer Based System," American Institute of Physics Report No. AIP/UDC-8.

[6] Giuliano, V. E., and Jones, P. E., (1966), "Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems," AD 642 829, CFSTI.

[7] Lancaster, F. W., (1968), Evaluation of the MEDLARS Demand Search Service, Bethesda, Md., National Library of Medicine.

[8] Natrella, M. G., (1962), Experimental Statistics, National Bureau of Standards Handbook 91, U. S. Government Printing Office, Washington, D. C.

[9] Salton, G., (1965), "The Evaluation of Automatic Retrieval Procedures - Selected Results Using the SMART System," American Documentation, Vol. 16, no. 3, pp. 209-222.

[10] King, D. W., "Evaluation During File Development of the Glass Technology Coordinate Index File." WR PO 19, For the U. S. Department of Commerce, Patent Office, November 1967.

3.2 SEARCH CHARACTERISTIC CURVES

by

Robert R. V. Wiederkehr

3.2.1 Introduction

Although a search for documents may proceed through a number of stages involving a user, a document retrieval system, and various types of document representations such as titles, abstracts, and full text, attention here will be focused on a one stage search. By a one stage search is meant a search where the user submits a request to the document retrieval system and the system responds by furnishing the user with a number of documents. Searches which involve a series of interactions between user and system are not one stage searches.

To evaluate the performance of an information retrieval system in executing a one stage search, it is helpful to quantify the relationship between the return from the search and the examining effort required to obtain the return. In this paper, the return will be measured by the number of relevant documents examined. If no documents are examined, then, obviously, no relevant documents are examined, and hence the return will be zero. On the other hand, if the entire file is examined i.e., the effort is a maximum, then the number of relevant documents examined will attain its maximum value. At intermediate value of search effort the search return will be assumed to increase monotonically.

The purpose of this paper is to develop models for a given document retrieval system and a given class of one stage searches which describe the relationship between the search return (expressed in the number of relevant documents retrieved) and the search effort

(expressed in the number of documents examined). This relationship is called the search characteristic curve. *

There are two ways in which the applicability of the notion of the search characteristic curve can be generalized. First of all, each stage of a multiple stage search can be described by a curve similar to a search characteristic curve. By appropriately combining these curves for each stage of the search an overall measure of return versus effort can be obtained.

Secondly, the number of categories or levels used to characterize the degree of relevance each examined document bears to the search request may be greater than two. (The usual treatment is to use only two categories: relevant and not relevant). The return from a search where there are k levels of relevance may be measured in the numbers of documents examined which fall into each of the k levels. In this case k - 1 search characteristic curves would result: one curve for each nonzero level of relevance. For example, if examined documents were categorized as being nonrelevant, relevant, or crucially relevant, as was proposed by Mooers [6], then k = 3 and two search characteristic curves would result: one for the relevant documents, and one for the crucially relevant documents.

To simplify the following presentation, neither of these generalizations of the search characteristic curve will be considered further. In other words, it will be assumed that all relevance judgments will be assumed to be dichotomous, i.e., examined documents are judged to be either relevant or not relevant. It will also be assumed that all searches are one stage searches.

---

* The expression, "search characteristic curve", has been used to describe various concepts in the literature. Therefore, the meaning attributed to it here must be restricted to its use in this paper.

3.2.2 Properties of a search characteristic curve

Two important types of one stage searches are the following:

a. A user's search request is submitted to the document retrieval system which responds by furnishing the user with a reasonable number of documents assessed by the system as being relevant to the search request. In this case all of the documents submitted to the user are examined by him and he judges a fraction of these as being relevant to his search request.

b. A user's search request is submitted to the system which responds by furnishing the user with a large number of documents ordered by the system according to the estimated degree of match between document and search request. In this case only a fraction of the documents are examined in order by the user who judges a fraction of these as being relevant to his search request.

For such searches, how can the search effort be varied? For a Type b search the effort may be readily increased by examining more documents in the ordered output from the system. For a Type a search it at first appears that the output is fixed because the user examines all of the documents in the output from the system. However, if the acceptance criteria used by the system in assessing whether or not a document is relevant to the search request is varied, then the size of the output from the systems can be varied and hence the number of documents examined can be varied. In either case, the search effort may be varied by varying the number of documents examined by the user.

To describe quantitatively how the search return (measured in number of relevant documents) varies with the search effort (measured in number of documents examined) it is convenient to define the following terms:

N = the total number of documents in the file

M = the number of relevant documents in the file

n = the total number of documents examined

m = the number of relevant documents retrieved when the number of documents inspected is n

f = n/N, the fraction of the file examined

r = m/M, the fraction of relevant documents which are examined (also the recall)

The search characteristic curve for the search is the relationship between m and n, and is illustrated in Figure 3.2.1. The normalized search characteristic curve, where both the abscissa and ordinate are normalized to a range between zero and one by the relationship between r and f, may also be shown in the same way. The search characteristic curve r versus n has been found to be particularly useful by some investigators.

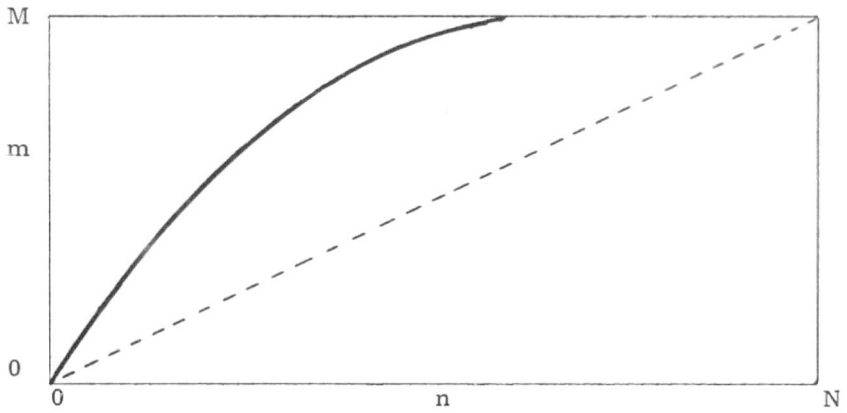


Figure 3.2.1. A Search Characteristic Curve.

There is an advantage to considering the normalized version of the search characteristic curve, viz. r versus f, because the same scales for abscissa and ordinate can be used for all values of N and M and, therefore, a great economy of expression will result. On the other hand, if searches with different values of M and N have substantially different shapes, then employing the normalized curves would mask these differences and possibly lead to invalid conclusions. However, it will be assumed here that searches have been appropriately categorized so that substantial differences in shapes of normalized search characteristic curves do not exist. Because of the economy of expression afforded by the normalized search characteristic curve, and because one can always use the normalized search characteristic curve and values of M and N to recapture the unnormalized search characteristic curve, the remainder of this paper will consider only the normalized form.

The properties which a search characteristic curve should have are the following:

i. It should pass through the points (f = 0, r = 0) and (f = 1, r = 1)

ii. r should be a monotonically increasing function of f

The second property follows from the fact that inspecting a greater fraction of the file should not decrease the number of wanted documents retrieved.

3.2.3. Models for search characteristic curves

In this section three models will be developed for the normalized search characteristic curves: one based on the Beta Distribution, one based on the equivalent number of random searches, and one based on a generalized Beta Distribution.

3.2.3.1 Search characteristic curves based on the beta distribution

Property (ii) suggests that r should be proportional to an appropriate cumulative distribution function and Property (i) suggests that

the appropriate distribution function is the Beta Distribution and that the proportionality constant is unity. Hence we may write:*

$$r = I_f(a,b) = \frac{1}{B(a,b)} \int_0^f t^{a-1} (1-t)^{b-1} dt, \quad 0 \leq f \leq 1 \qquad (1)$$

where a and b are the (positive) parameters of the Beta Distribution.

The rate of increase in the search return with search effort from (1) is

$$\frac{dr}{df} = \frac{1}{B(a,b)} f^{a-1}(1-f)^{b-1}, \qquad 0 \leq f \leq 1 \qquad (2)$$

and is illustrated in Figure 3.2.2 for several values of a and b.

The values of a and b which represent a particular search indicate how good the search is. A random search is characterized by an equal rate of increase of search return with search effort for all values of f. This corresponds to a = 1 and b = 1. Since the object of a search is to obtain as many of the relevant documents as possible with as little search effort as possible, a search which yields high values of dr/df for small values of f is a good one. This corresponds to large values of b and small values of a. Furthermore, a good search would be expected to have a decreasing rate of return. For this condition to hold for all values of f, a should be less than or equal to unity.

3.2.3.2 Search characteristic curves based on equivalent number

The Beta Distribution can be related to other well known distributions, notably the F-distribution and the Binomial distribution. The relationship between the Beta Distribution and the Binomial distribution is

---

* The complete Beta function is defined to be:

$$B(a,b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

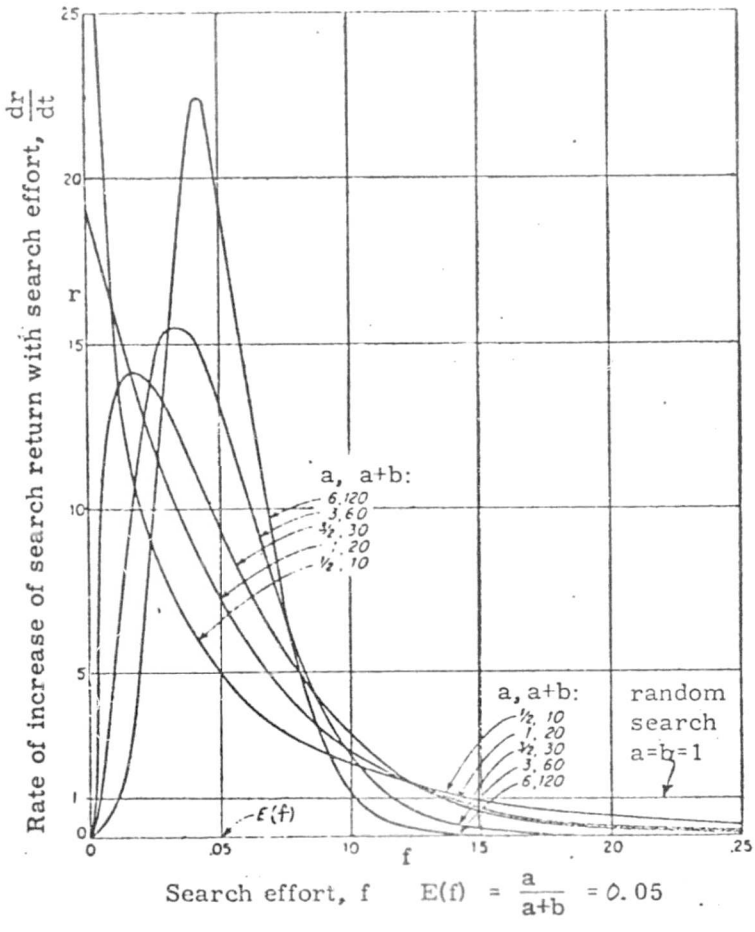


Figure 3.2.2. Variation of rate of return with search effort - beta distribution with mean value, a/(a+b) = 0.05

very instructive, for not only is it possible to use binomial tables to evaluate the Beta distribution but one can attach a definite meaning to the otherwise purely empirical parameters a and b.

The relationship between the Beta and Binomial distribution is (see Abramowitz [1]:

$$I_f(a,\ n - a + 1) = \sum_{j=a}^{n} \binom{n}{j} f^j (1 - f)^{n - j} \tag{3}$$

The right-hand side of (3) can be associated with a series of random searches for the wanted documents in the file. This association can be used to interpret the parameters a and b in terms of these random searches as will now be demonstrated.

Consider a single random search against the file where each random search consists of examining a fraction f of the total file. Because the search is random the proportion of the wanted documents which are examined and therefore retrieved, equals f, i.e., r = f. Furthermore, if n such searches are made against the file, then the proportion of wanted documents which are inspected in at least a of the n searches is:

$$\sum_{j=a}^{n} \binom{n}{j} f^j (1 - f)^{n - j} \tag{4}$$

which is the right-hand side of (3). This proportion can be made large by making a small and n large.

For a given value of f an actual search produces a fraction of the wanted documents r which is equivalent to the fraction of wanted documents found in at least a of n random searches. Therefore let n be called the equivalent number of random searches. Comparison of (1) and (3) reveals that for a search whose characteristic curve is $I_f(a, b)$ the equivalent number of random searches is given by:

$$n = a+b-1 \tag{5}$$

and r for this search equals the fraction of wanted documents found in at least a of these n equivalent random searches.

A particularly interesting class of searches are those for which a = 1. For these searches the number of equivalent random searches is b, as can be seen from (5). Furthermore, r equals the fraction of wanted documents found in at least one of these b random searches. In this case the search characteristic curve assumes the following simple form which can be obtained readily either by integrating (1) or by evaluating 1 minus the complement of the right-hand side of (3):

$$r = 1 - (1-f)^b \tag{6}$$

3.2.3.3 Characteristic curves based on a generalized beta distribution search

Equation (6) is considerably simpler than (1) and can be readily used to compute the value of r from a knowledge of f. This is a highly desirable property. On the other hand (6) contains only one parameter and therefore is not as flexible as (1) for purposes of describing search characteristic curves of operating systems. In this subsection a model for r is developed which has the two desirable properties mentioned. That is, the model permits r to be readily computed and yet is more flexible than Equation (1) in that it contains two rather than one parameter.

Recall that (6) was obtained from (1) by letting a equal unity and is of the form

$$I_x(1,b) = 1 - (1-x)^b \tag{7}$$

If we let

$$x = f^{1/k} \tag{8}$$

and substitute x into (7) we obtain the following function of f with two parameters, k and b, which may be used as a model for r

$$r = 1 - (1-f^{1/k})^b \tag{9}$$

The form of (9) is almost as simple as (6) and yet is more flexible because it has two parameters rather than just one. Also Property i is satisfied by (9) for if f is zero the r is zero, and if f is unity the r is unity.* Notice that (9) reduces to (6) when k is unity.

An effective search usually should not require that a large fraction of the file be inspected to obtain a reasonable number of relevant documents. Therefore, the more interesting region of a search characteristic curve is the region where f assumes small values. In this region the search characteristic curve described by (9) can be approximated by a simpler form which is easy to plot.

A Taylor's expansion of (9) about the point f = 0 up to first order terms yields:

$$r \simeq (b/k)f^{1/k} \tag{10}$$

Taking logarithms of (10) gives

$$\log r \simeq \log (b/k) + 1/k \log f \tag{11}$$

From (11) it is clear that a log-log plot of r versus f for small values of f should be linear with a slope equal to 1/k and an intercept equal to log (b/k). This information can be used for graphically estimating the values of b and k as will be shown in Section 3.2.6.

3.2.4 The relationship between the search characteristic curve and the operating characteristic curve proposed by Swets

Swets [8] has proposed using the operating characteristic curve of an information retrieval system as a measure of its performance. The operating characteristic curve is a plot of the probability of retrieving a document given that the document is relevant to the search request - denoted

* The right-hand side of (9) is a valid cumulative distribution factor whose statistical properties are developed in a companion paper by the author [10].

by $P(R|r')$ - against the probability of retrieving a document given that it is not relevant - denoted by $P(R|\bar{r}')$. In this notation R denotes the event that a document is retrieved, r' denotes the event that the document is relevant, and $\bar{r}'$ denotes that the document is not relevant. Furthermore, Swets has shown that the operating characteristic curves for several information retrieval systems are well represented by straight lines when the scales for $P(R|r')$ and $P(R|\bar{r}')$ are normal probability scales, i.e., the scales are linear with respect to the normal standard deviate.

The purpose of this section is to relate the operating characteristic curve proposed by Swets to the search characteristic curves developed in this paper. Toward this end it is convenient first to relate the probabilities of the events r' and R to quantities defined earlier.

With respect to a given search request the probability that a document selected from the file at random is relevant is equal to the fraction of documents in the file that are relevant, i.e.,

$$P(r') = \frac{M}{N} \tag{12}$$

The probability that a document selected from the file at random is retrieved (or examined) is the fraction of documents in the file that are inspected:

$$P(R) = \frac{n}{N} \tag{13}$$

By definition of f it follows that

$$P(R) = f \tag{14}$$

Also, the probability that a document is retrieved given that it is relevant is the fraction of relevant documents in the file that are retrieved.

$$P(R|r') = \frac{m}{M} \tag{15}$$

By definition of r it follows that

$$P(R|r') = r \tag{16}$$

In the development of Swets and most other investigators a document is either relevant, r', or it is not relevant, $\bar{r}'$, and these events are mutually exclusive. Consequently, we may write*

$$P(R) = P(R|r') \cdot P(r') + P(P|\bar{r}')\, P(\bar{r}') \tag{17}$$

This well known relationship follows from the additivity of probabilities and the definition of conditional probability, e.g., see Feller [4].

Because r' and $\bar{r}'$ are complementary events it follows that

$$p(\bar{r}') = 1 - p(r') \tag{18}$$

Substituting (13), (14) and (18) into (17) yields

$$f = P(R|r')\frac{M}{N} + P(R|\bar{r}') \cdot \left[1 - \frac{M}{N}\right] \tag{19}$$

Consequently, if an operating characteristic curve is given, i.e., if pairs of values of $P(R|r')$ and $P(R|\bar{r}')$ are given, and the fraction of documents in the file that are relevant is known, then one can construct a search characteristic curve using (16) and (19).

If the fraction of documents in the file which are relevant is small compared to unity, then (19) can be simplified considerably, for then

$$\frac{M}{N} \cong 0 \tag{20}$$

and

$$1 - \frac{M}{N} \cong 1 \tag{21}$$

so that (19) reduces to

$$f \cong P(R|\bar{r}') \tag{22}$$

---

* In the more general case where k rather than two categories of relevance are considered, the right-hand side would contain k terms instead of two.

Notice that this approximation is valid only if

$$P(R|\bar{r}') >> P(R|r') \frac{M}{N} \tag{23}$$

and therefore may deteriorate when large values of recall, $P(R|r')$, are reached.

In the event that (22) holds then Swets' operating characteristic curve and the search characteristic curve are identical. This follows from (16) and (22).

3.2.5 The relationship between search characteristic curves and recall-precision

A pair of measures which has been used by many investigators, notably Perry, Kent and Berry [7], Cleverdon [3], and Lancaster [5], for the purpose of evaluating information retrieval systems are the recall and the precision. The recall, r, has already been defined. The precision, p, is defined by

$$p = \frac{m}{n} \tag{24}$$

From the definition of r and f it follows that

$$p = \frac{M}{N} \cdot \frac{r}{f} \tag{25}$$

Let p be called the file precision ratio and be defined to be M/N. Then clearly

$$p = P \cdot \frac{r}{f} \tag{26}$$

Since r/f is the slope of a line passing though the origin and a point (r,f) of the search characteristic curve, one can readily generate the values of p using a search characteristic curve and the file precision ratio using (26).

3.2.6 Some observed search characteristic curves

Bryant, Searls, Shumway, and Weinman [2] have compared four different search strategies. The observations consisted of pairs of values for m, n averaged over a set of 13 queries, each processed according to

the four search strategies. The results are presented in Figure 3.2.3 which is a plot of log r versus log f. Also included in Figure 3.2.3 is a typical observed operating characteristic curve presented by Swets [8] - a curve with unit slope (S = 1), separation parameter E equal to 2, and $\frac{M}{N}$ equal to 0.03.

The search characteristic curves for search strategies 1, 2 and 4 appear to be reasonably well represented by (6) or equivalently by (9) with k = 1. These searches, therefore, can be described by the equivalent number of random searches. From Figure 3.2.3 it can be seen that search strategies 1, 2 and 4 have equivalent number of random searches approximately equal to 20, 50, and 10, respectively.

It follows that for k = 1 the value of b, the equivalent number of random searches, is the ratio of r to f along the r axis. For example, b for search strategy 2 is 0.05/0.001 or 50.

The search characteristic curves for strategy 3 and for the typical operating characteristic curve presented by Swets are not very well represented by values of k = 1. A value of k = 2, however, does appear to describe these curves reasonably well.

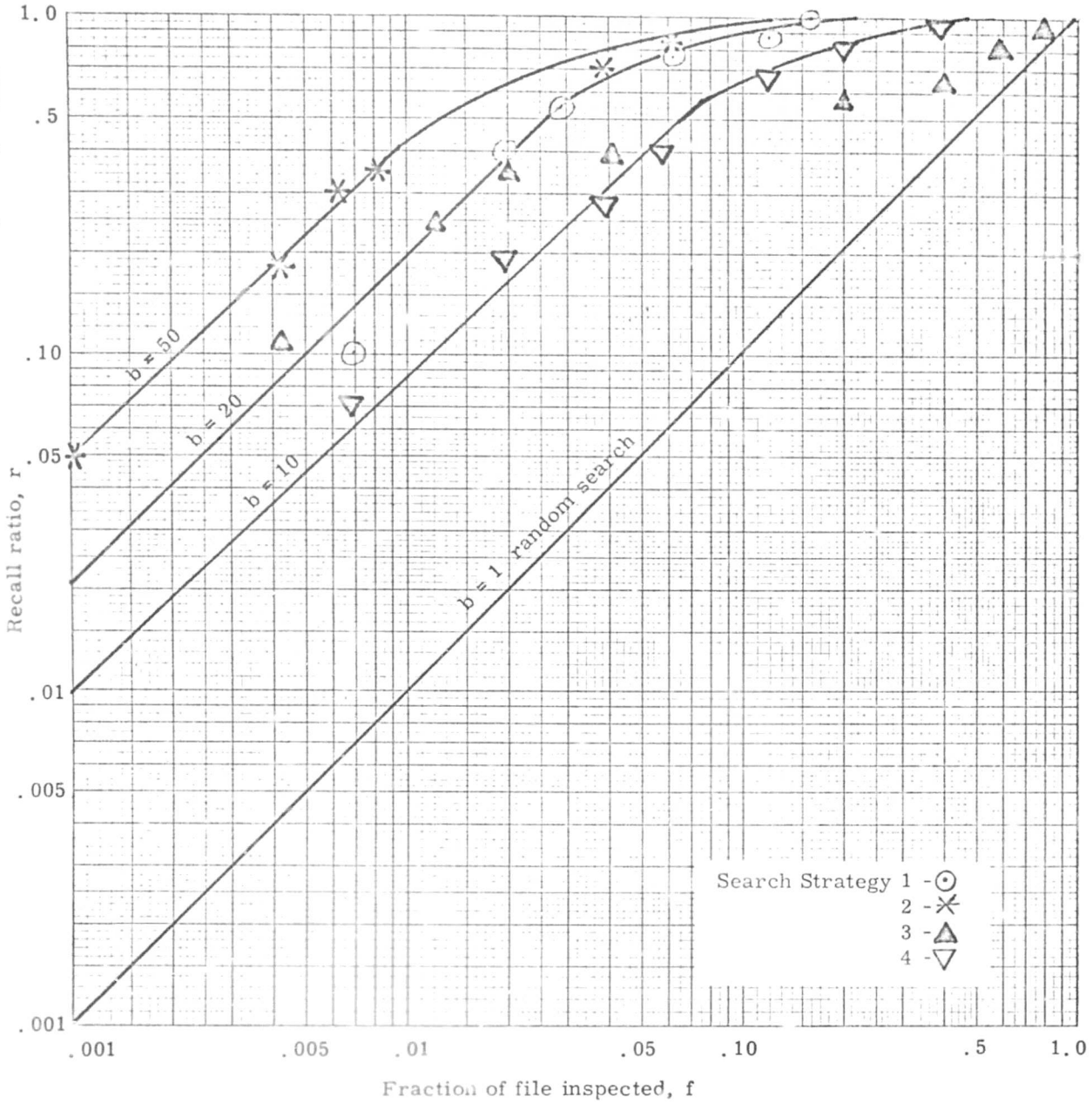


Figure 3.2.3. Observed search characteristic curves

References

[1] Abramowitz, M. and Stegun, I. A., (1964), Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematics Series No. 55, U. S. Government Printing Office, Washington, D. C., p. 945, Equation 26.5.24.

[2] Bryant, E. C., and D. T. Searls, R. H. Shumway, D. G. Weinman, (March 31, 1967), "Associate Adjustments to Reduce Errors in Document Screening," Contract AF 49(638)-1484, Westat Research, Inc., p. 26.

[3] Cleverdon, Cyril W., (October 1962), "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, 305 p. plus figures, PB 162 342.

[4] Feller, W., (1957), An Introduction to Probability Theory and Application, Vol. I, Second Edition, John Wiley and Sons, New York, p. 106.

[5] Lancaster, F. W., (August 1966), "Evaluating the Small Information Retrieval System," Journal of Chemical Documentation, 6: 158-160.

[6] Mooers, Calvin N., (August 1959), "The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval Systems," Report No. RADC-TN-59-160, Report to U. S. Air Force, Rome Air Development Center, Cambridge, Mass., Zator Co., 20 p.

[7] Perry, James W., Allen Kent and Madeline M. Berry, (1956) "Operational Criteria for Designing Information Retrieval Systems," in Machine Literature Searching, New York, Interscience Publishers, Inc., p. 41-48.

[8] Swets, John A., (July 19, 1963), "Information Retrieval Systems," Science 141: 245-250.

[9] Swets, John A., (June 1967), "Effectiveness of information Retrieval Methods," Report for Air Force Cambridge Research Laboratories, United States Air Force, No. AF19(628)-5065, Bolt Beranek and Newman, Inc., Cambridge, Mass., 47 p. AD 656 340.

[10] Wiederkehr, [illegible] V., (October 1968), "The Family of Modified Beta Probability Distribution," Contract NSF-C 491, Westat Research, Inc.

3.3 THE FAMILY OF MODIFIED BETA PROBABILITY DISTRIBUTIONS

by

Robert R. V. Wiederkehr

3.3.1 Introduction

In the process of exploring mathematical expressions suitable for describing search characteristic curves of information retrieval systems, a probability distribution related to the beta probability distribution was developed by the author in another paper [1]. In this paper some of the basic properties of this distribution, called the modified beta distribution, will be reviewed and extended.

A random variable defined over a finite interval (a, b) can be transformed via translation and scale change to a new random variable defined over the interval (0, 1). A convenient way of representing the distributions of such random variables is by selecting an appropriate member of the family of beta distributions. Although the beta distribution has certain desirable properties, such as belonging to the exponential family of distributions, it has a cumulative distribution function which in general cannot be readily evaluated without the aid of extensive tables or electronic computers. The purpose of this paper is to develop a family of distributions, suitable for describing a random variable defined over the interval (0, 1), whose cumulative distribution function can be simply evaluated without the aid of tables or electronic computers.

Toward this end certain properties of the beta distribution will be viewed and then used to generate the family of modified beta distribution with the desired properties. The method used to generate the family of modified beta distributions from the family of beta distributions will then be used to generate the Weibull family of distributions from a member of the family of gamma distributions.

3.3.2 The family of beta probability distributions

The family of modified beta probability distributions to be developed depends strongly on the beta distribution. Therefore, it is convenient to first review certain properties of the beta distribution. Let X be a random variable defined on the unit interval (0, 1). If X belongs to the family of beta probability distributions with parameters a and b, then the cumulative distribution function (c. d. f.) of X is given by:

$$P(X \leq x) = I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1} (1-t)^{b-1} dt, \qquad (1)$$

$$0 < x < 1, \quad a > 0, \quad b > 0$$

where B(a, b), the complete function, is defined by:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt \qquad (2)$$

The complete beta function is related to the gamma function (defined by (14) below) by:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \qquad (3)$$

The density function of X is

$$f(x|a,b) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases} \qquad (4)$$

The moments of X are readily evaluated from (4) and (2). In particular the m^{th} moment of X is

$$E(x^m) = \frac{1}{B(a,b)} \int_0^1 x^{m+a-1}(1-x)^{b-1} dx = \frac{B(m+a,b)}{B(a,b)} \qquad (5)$$

In view of (3), (5) reduces to

$$E(x^m) = \frac{\Gamma(m+a)}{\Gamma(a)} \cdot \frac{\Gamma(a+b)}{\Gamma(m+a+b)} \qquad (6)$$

3.3.3 The genesis of the family of modified beta probability distributions

For most values of a and b the right-hand side of (1) cannot be integrated to yield a simple closed form expression. An exception to this statement is when a is unity and b is positive, for then it is easily shown that

$$I_x(1,\beta) = 1 - (1-x)^{\beta} \tag{7}$$

Although (7) is simple, it is not very flexible because only one parameter is available for curve fitting purposes.

More flexibility can be incorporated into (7) by using the following transformation

$$x = y^{\alpha} \qquad \begin{matrix} 0 < x < 1 \\ \alpha > 0 \end{matrix} \tag{8}$$

so that (7) becomes

$$P(X \leq x) = P(Y \leq y) = I_x(1,\beta) = 1 - (1-y^{\alpha})^{\beta} \equiv H_y(\alpha,\beta) \tag{9}$$

From (8) it is clear that the transformed random variable Y is defined over the interval (0, 1) and from (9) or (1) it is clear that $H_y(\alpha,\beta)$ is the c.d.f. for Y. The family of probability distribution $H_y(\alpha,\beta)$ will be called the modified beta family of probability distributions. The probability density function of Y for α, β positive is

$$f_Y(y|\alpha,\beta) = \begin{cases} \alpha\beta(1-y^{\alpha})^{\beta-1}y^{\alpha-1} & 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{10}$$

The moments of the transformed variable Y can be determined straightforwardly from (5). In particular, the n^{th} moment of Y is given by

$$E(Y^n) = E(X^{n/\alpha}) = \Gamma\left(\frac{n}{\alpha} + 1\right) \cdot \frac{\Gamma(1+\beta)}{\Gamma\left(\frac{n}{\alpha}+1+\beta\right)} \tag{11}$$

The right-hand side of (11) can be evaluated using tables for the gamma function.

When E(Y) is very small, which is often the case in practice, then an excllent approximation to $H_y(\alpha, \beta)$ can be obtained by a Taylor's expansion about the point y = 0. Since $H_o(\alpha, \beta) = 0$, and $H_y(\alpha, \beta) = f_r(y|\alpha, \beta)$ we have from (10)

$$H_y(\alpha, \beta) \cong \alpha\beta y^{\alpha}, \qquad 0 < y << 1 \tag{12}$$

Taking logarithms of (12) yields

$$\log H_y(\alpha, \beta) \cong \log \alpha\beta + \alpha \log y \tag{13}$$

From (13) it is clear that a plot of the logarithm of the c.d.f. versus the logarithms of y should be linear for small values of y with a slope equal to α and an intercept equal to $\log \alpha\beta$. This property can be used to estimate α and β graphically using log-log paper.

3.3.4 The genesis of the Weibull family of probability distributions

It is interesting to note that the family of Weibull distributions can be generated from the gamma distribution by the same method as was used above to generate the family of modified beta distributions from the beta distributions.

The c.d.f. of the gamma distribution is

$$F_x(a, b) = \frac{1}{\Gamma(a)} \int_0^x (bx)^{a-1} e^{-bx} d(bx), \qquad \begin{array}{l} 0 < x < \infty \\ a > 0 \\ b > 0 \end{array} \tag{14}$$

where

$$\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx \qquad a > 0 \tag{15}$$

The n^{th} moment of this distribution is

$$E(x^m) = \frac{\Gamma(m+a)}{b^m \Gamma(a)} \tag{16}$$

The integral on the right side of (14) assumes a particularly simple form

when a is unity for then it reduces to the exponential distribution

$$F_x(1,b) = 1 - e^{-bx} \qquad 0 < x < \infty \tag{17}$$

By making the transformation

$$= y^{\alpha} \qquad a > 0, \quad 0 < y < \infty \tag{18}$$

one obtains

$$G_y(\alpha,\beta) = F_x(1,\beta) = 1 - e^{-\beta y^{\alpha}} \tag{19}$$

The right-hand side of (19) is the c.d.f. of the Weibull distribution.

The moments of the Weibull distribution are readily obtained from (16) and (18). In particular, the n^{th} moment of Y is given by

$$E(Y^n) = E(x^{\frac{n}{\alpha}}) = \beta^{-\frac{n}{\alpha}}\Gamma(\frac{n}{\alpha} + 1) \tag{20}$$

Reference

[1] Wiederkehr, Robert R. V., "Search Characteristic Curves," Contract NSF-C 491, Westat Research, Inc., November 1967.

3.4 CONTINGENCY TABLES IN INFORMATION RETRIEVAL: AN INFORMATION THEORETIC ANALYSIS

by

R. H. Shumway

3.4.1 Introduction

Most often in the evaluation of an operating or proposed information retrieval system, the results of the data collection procedure are presented in categorical form. This means that the underlying observations are discrete random variables such as the number of relevant documents retrieved or the number of irrelevant documents not retrieved. In fact, a common practice is to present the results of an experiment in terms of a two-way contingency table whose entries have meaning in the information retrieval context. Table 3.4.1 shows the usual arrangement

Table 3.4.1. Contingency Table Characterizing Document Retrieval.

	Not Retrieved	Retrieved	
Not Relevant	X_{11}	X_{12}	$X_{1.}$
Relevant	X_{21}	X_{22}	$X_{2.}$
	$X_{.1}$	$X_{.2}$	N

representing the outcomes of an experiment where X_{11}, X_{12}, X_{21}, X_{22} are the numbers of observations in each of the four categories, with the dot notation symbolizing marginal totals. The two-way table yields measures of interest such as recall ($X_{22}/X_{2.}$) and precision ($X_{22}/X_{.2}$). One may collect a whole series of such two-way tables for competing methods or systems with the objective being a comparison between the different systems.

The purpose of this discussion is to propose the information statistic as a measure of association or effectiveness for a two-way contingency table. The notion of association is introduced as the lack of independence in the two-way table which associates the retrieved documents with the relevant documents. It is certainly true that almost any reasonable retrieval scheme will fail an independence test, but it is the extent to which a dependence (lack of independence) exists which makes one system better than another. The information statistic ranks the possible two-way tables in an order which is directly proportional to the likelihood of the table under the assumption of no association. The information statistic also provides the vehicle for partitioning this dependence between the various factors which influence the outcome of the search. In this way it performs for discrete data the same function as the analysis of variance does for continuous data. The partitioning of sums of squares by factors is simply replaced by a corresponding partition of the logarithmic information components. The approach may also be obtained by way of the likelihood ratio test and is approximately equivalent to the usual chi-square tests for independence in a contingency table. Several examples are presented to illustrate the versatility and applicability of the approach.

3 [illegible] Discussion

We regard the numbers generated in two-way tables as a multinomial distribution so that the probability of obtaining

$$P(X_{11}, X_{12}, X_{21}, X_{22}) = \frac{N!}{X_{11}!X_{12}!X_{21}!X_{22}!} P_{11}^{X_{11}} P_{12}^{X_{12}} P_{21}^{X_{21}} P_{22}^{X_{22}}$$

$$\sum_{ij} X_{ij} = N \qquad (1)$$

where P_{ij} is the probability that a document falls in the i^{th} row and j^{th} column of Table [illegible]. A measure of association must reflect clustering of numbers [illegible] of Table 3.4.1, since one is interested in associating those in the relevant with those in the [illegible]

category and the nonretrieved documents in the nonrelevant category. Clearly, such a number would be related to the departure from independence of the row and column classifications (i.e., the extent to which the relation $P_{ij} = P_{i.} P_{.j}$ is not true for all i and j). It should be noted in this context that non-independence components from nonrelevant documents not retrieved and relevant documents retrieved are not the only kinds of dependencies possible, but within most operating systems these are the likely ones. A measure can be defined which indicates the departure of the specific two-way table from one which would be generated if the row and column categories were independent. This argument ([3] p. 158) for an r x c table with observed cell entries X_{ij}, i = 1, ..., r, j = 1, ..., c leads to an information statistic

$$2\hat{I}(H_1 : H_2) = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} X_{ij} \log \frac{NX_{ij}}{X_{i.} X_{.j}} \qquad (2)$$

for testing the independence hypothesis $H_2 : P_{ij} = P_{i.} P_{.j}$ where $N = \sum_{ij} \cdot X_{ij}$, $X_{i.} = \sum_{j} X_{ij}$, $X_{.j} = \sum_{i} X_{ij}$. The dot notation applied to a probability indicates a marginal probability.

The preceding information number has asymptotically the chi-square distribution so that the relative significance of values for $2\hat{I}(H_1 : H_2)$ can be appraised from a probabilistic standpoint. For example, any information number computed from Table 3.4.1 has an approximate chi-square distribution with two degrees of freedom.

A statistic related directly to $2\hat{I}(H_1 : H_2)$ has been applied by R. Shirey, et. al. [5] to the following information retrieval situation. Suppose that we wish to evaluate the effectiveness of certain cues commonly associated with documents: in particular, their effectiveness as indicators of the relevance or nonrelevance of those same documents. The cues chosen are the conventional ones: citations, abstracts, first

paragraphs, last paragraphs and first and last paragraphs combined. In particular, in [5] a set of two-way tables relating the decisions made by the cues to decisions made by the judges is utilized as in Table 3.4.2

Table Relationship between cues and judges' decisions

		Cue Indication: Not Relevant	Cue Indication: Relevant	
Judge Decision	Not Relevant	X_{11}	X_{12}	$X_{1.}$
	Relevant	X_{21}	X_{22}	$X_{2.}$
		$X_{.1}$	$X_{.2}$	

Table 3.4.3 summarizes the results of the retrieval experiment given in reference [5]. The entries in the citations table show, for example, that 44 documents which the citations indicated as relevant were relevant, 112 documents which the citations indicated as not relevant were not relevant, 55 of the documents which the citations indicated as not relevant were relevant, and 16 documents designated as nonrelevant by the cues actually turned out nonrelevant.

Shirey, et. al. [5] present their equations in terms of the joint probabilities associated with Table 3.4.2 obtained by dividing each entry by N. It is convenient to retain their notation which substitutes the theoretical probability $P(D_{ij}C_j)$ for the sample frequency X_{ij}/N and identifies $P(D_{ij}C_j)$ as the joint probability that a document will be in the i^{th} relevance category while the cue is in the j^{th} relevance category. Marginal probabilities are defined as usual by summing the joint probabilities over the appropriate subscript. Then, as in conventional engineering communication theory, the original uncertainty about the document relevance is defined as

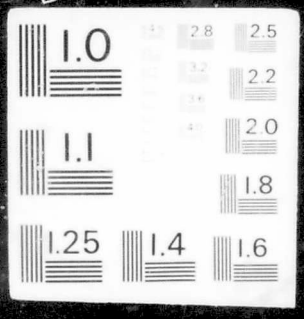


Table 3.4.3 Summary Data from a Retrieval Experiment Using Several Methods of Searching

Citations

Judge Decision \ Cue			
	44	55	99
	16	112	128
	60	167	227

First Paragraphs

Judge Decision \ Cue			
	55	43	98
	16	110	126
	71	153	224

Last Paragraphs

Judge Decision \ Cue			
	61	38	99
	20	106	126
	81	144	225

Abstracts

Judge Decision \ Cue			
	43	39	82
	18	113	131
	61	152	213

First and Last Paragraphs

Judge Decision \ Cue			
	63	31	94
	10	121	131
	73	152	225

$$H(D) = -\sum_{i} P(D_i) \log P(D_i) = -\sum_{ij} P(D_i, C_j) \log P(D_j) \tag{3}$$

The conditional uncertainty after the cues have been examined is

$$H(D|C) = -\sum_{ij} P(D_i, C_j) \log P(D_i|C_j)$$

$$= -\sum_{ij} P(D_i, C_j) \log \frac{P(D_i, C_j)}{P(C_j)} \tag{4}$$

where the vertical bar indicates a conditional probability.

The "uncertainty resolved" by knowing the cue is then given by

$$H(D) - H(D|C) = \sum P(D_i, C_j) \log \frac{P(D_i, C_j)}{P(D_i)\, P(C_j)} \tag{5}$$

A sample estimate for the information gain or loss in uncertainty is

$$\hat{H}(D) - \hat{H}(D|C) = \sum_{ij} \frac{X_{ij}}{N} \log \frac{NX_{ij}}{X_{i.}\, X_{.j}}$$

$$= \frac{1}{N} \hat{I}(H_1, H_2) \tag{6}$$

so that the two information approaches are consistent and the asymptotic distribution of the conditional information gain is also distributed as chi-square with one degree of freedom.

For the simple two-way table, then, the usual statistic for testing independence of row and column classifications is identical with the measure developed from the conditional entropy communication channel approach. In the following section we shall show how the two-way table information measure can be extended to a more general situation such as the data in Table 3.4.3. General material on the use of the information theoretic measure for testing several hypotheses in multidimensional contingency tables can be found in references [3] and [4].

3.4.3 Examples

The data in Table 3.4.3 will be represented in the form X_{ijk} where i and j refer to the two-by-two table indexes and k identifies the type of cue (i, j = 1, 2; k = 1, ..., 5). The first question that will be considered is the same as considered by Shirey, et. al. [5], and relates to measuring the association between cue indications and relevance for each particular cue (i.e., citations, abstracts, etc.). Now, if P_{ijk} is the probability of being in the i^{th} row and j^{th} column of the k^{th} two-way table, with $P_{i.k}$, $P_{.jk}$, $P_{..k}$, the corresponding marginal probabilities, we may envision the hypothesis for testing independence in the form

$$H_2: \qquad P_{ijk} = \frac{P_{i.k} P_{.jk}}{P_{..k}} \tag{7}$$

This hypothesis tests whether the judge's decision is independent of the cue for a given type of cue. The information measure of association is the analog of the partial correlation coefficient between judges' decisions and cues, with the effect of method eliminated. Symbolically, we may write the conditional information component as

$$2\hat{I}\left((D|K) \times (C|k)\right) = 2 \sum_{ij} X_{ijk} \log \frac{X_{..k} X_{ijk}}{X_{i.k} X_{.jk}} \tag{8}$$

which has the chi-square distribution with one degree of freedom. Computations are facilitated with the table of 2n log n in [4]. The results shown in Table 3.4.4 are all highly significant, since 12.16 is the chi-square critical point at the 99.95% level of significance. This is to be expected, but the ordering of the conditional information components is interesting, with two paragraphs yielding approximately twice the information in either the first or last paragraph. These components are all highly significant so that, given that the k^{th} method is assigned,

there is a very high degree of association between the cue indicator and the judge's decision as to the relevance of the document.

The above comparisons are not particularly illuminating, since they yield only cursory information about the significance of the differences between methods. In order to test for these effects and others we regard the data in Table 3.4.3 as a three-dimensional contingency table, with the observation X_{ijk} denoting the number of documents in the i^{th} decision and j^{th} cue category for the k^{th} method (i, j = 1, 2, k = 1, ..., 5). In this case we may regard the total association information for the three-way table as being specified by the hypothesis

$$H_3: P_{ijk} = P_{i..} P_{.j.} P_{..k} \tag{9}$$

which tests the independences of the three categories, decision, cues, and methods. By [3], p. 162, this yields the information component

$$\hat{I}(DxCxM) = \sum_{i,jk} X_{ijk} \log \frac{N^2 X_{ijk}}{X_{i..} X_{.j.} X_{..k}} \tag{10}$$

and the resulting number is shown in Part II of Table 3.4.4 as significant at the 0.001 level. Now, given that there is a high degree of association between the three categories, the natural question is the extent to which the association between i, j, and k is influenced by the association between i and j. If this component is the only significant contributor to the decisions-by-cues-by-methods (DXCXM) component the effect of method on the retrieval performance can be discounted. Hence, we partition the independence component $\hat{I}$ (DxCxM) into an information component depending strictly on the association between the judge decision and cue, say, $\hat{I}$(DXC), and an information component reflecting the association between methods and the decision cue two-way table, say, $\hat{I}$(DCxM). This effectively tests the homogeneity of the two-way tables for different methods. The hypotheses of interest are

$$H_4: P_{ij.} = P_{i..} P_{.j.} \tag{11}$$

Table 3.4.4. Analysis of Information

I. Analysis of Information (Conditional Components)

	k	$\hat{I}(D\|k)x(C\|k)$
Citations	1	29.725***
Abstracts	2	36.763***
First Paragraph	3	49.504***
Last Paragraph	4	51.923***
First and Last Paragraphs	5	93.712***

II. Analysis of Information (Independence Component)

Source	Information	d.f.
$\hat{I}(DxC)$	250.928***	1
$\hat{I}(DCxM)$	18.309	12
$\hat{I}(DxCxM)$	269.237***	13

III. Analysis of Information (I(DCxM) Method Component)

	$\hat{I}(DCxM)$	d.f.
Within Abstracts and Citations	2.427	3
Within Paragraph Methods	7.064	6
Between Abstracts-Citations, and Paragraphs	8.818*	3
	18.309	12

* Significant at 0.05 level
** Significant at 0.01 level
*** Significant at 0.001 level

$$P_5 : P_{ijk} = P_{ij.} P_{..k} \qquad (12)$$

In the above, H_4 corresponds to the hypothesis of no correlation between decisions and cues, with H_5 indicating a hypothesis of no multiple correlation between methods and the decision cue combination.

The information measures ([3], p. 167) are

$$\hat{I}(DxC) = 2 \sum_{ij} X_{ij.} \log \frac{NX_{ij.}}{X_{i..} X_{.j.}} \qquad (13)$$

$$\hat{I}(DCxM) = 2 \sum_{ijk} X_{ijk} \log \frac{NX_{ijk}}{X_{ij.} X_{..k}} \qquad (14)$$

The numerical results for (13) and (14) applied to the Table 3.4.3 data are shown in part II of Table 3.4.4. The component testing the independence of cues and judges' decisions is highly significant whereas the component yielding information on methods is not significant. Thus one would tend to conclude on the basis of the above that methods are not inherently different. The conditional information components in Part I of Table 3.4.4 indicate that if one pools methods into a group containing paragraph methods and a group containing the abstract and citation method results a significant effect might result. Therefore, in part III of Table 3.4.4 we have partitioned the DCXM component into two within group components and a between group component which is significant at the 0.05 level.

To summarize, all methods are decidedly better than a random method would be, as evidenced by the uniformly high conditional information components. The association, as would be expected, is mainly between the particular cue and whether or not the document is judged relevant. However, a further examination of the independence component implies that there is a significant difference between the methods if they are grouped into a method involving only abstracts and citations and a method

involving only paragraphs. Hence, one tends to conclude that methods involving paragraphs are significantly better than methods involving citations and abstracts.

3.4.4 Conclusions

The purpose of this discussion has been to examine the performance of contingency table techniques as they might be applied in the field of information retrieval. The methods rest, in theory, upon the assumption that the data is categorical, an assumption which is certainly well satisfied in most evaluation study contexts. Then, based on multinomial theory, an approach developed either from the likelihood ratio or the information statistic is presented which yields a sequence of tests for association between retrieved and relevant data in a typical information retrieval experiment. It is shown that the tests can give answers to many of the questions which arise about the effectiveness of certain information retrieval systems. In addition, the information measure presented appears to have an advantage over recall and precision in that all of the data in the two-way table is used in a single measure of association. Relations between the information measures and the usual simple, partial and multiple correlation functions are given reinforcing the intuitive appeal of the information approach.

References

[1] Cochran, W. G., (1952), "The X^2 Test of Goodness of Fit, Annals of Mathematical Statistics, Vol. 32, pp. 12-40.

[2] Good, I. J., (1963), "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," Annals of Mathematical Statistics, Vol. 34, pp. 911-934.

[3] Kullback, S., (1959), Information Theory and Statistics, Wiley, New York.

[4] Kullback, S. M., M. Kupperman and H. H. Ku, (1962), "An Application of Information Theory to the Analysis of Contingency Tables with a Table of 2N log N," J. Research, National Bureau of Standards, 66B, pp. 217-243.

[5] Shirey, Donald L. and M. Kurfeerst, (1967), "Relevance Predictability II," Chapter 14 in Electronic Handling of Information: Testing and Evaluation, Academic Press, London.

3.5 A NET BENEFIT MODEL FOR EVALUATING ELEMENTARY DOCUMENT RETRIEVAL SYSTEMS

by

Robert R. V. Wiederkehr

3.5.1 Introduction

Most efforts to evaluate information retrieval systems have dealt with systems in the research and development stage, and have been concerned primarily with demonstrating the technical feasibility of certain subsystems, such as indexing. The performance measures used in these evaluations have emphasized the recall (fraction of relevant documents that are retrieved) and the precision (fraction of retrieved documents that are relevant) or closely related measures, and have deemphasized or ignored factors of cost and benefits. In the research and development stage of a system, where the primary objective is to demonstrate technical feasibility, this approach appears appropriate. However, when a large-scale system advances to the operational stage, the primary objective should shift toward demonstrating economic feasibility, and performance measures should reflect both cost and benefits. The purpose of this paper is to develop a framework for quantitative economic evaluation of information retrieval systems.

3.5.2 A rationale for the economic evaluation of document retrieval systems

The evaluation of a system in general and an information retrieval system (or subsystem) in particular, may be considered to be a problem of economic choice which, according to Hitch and McKean [1]* has the following five elements:

---

*Hitch and McKean's book has been widely accepted as a standard reference by evaluators and systems analysts in the Department of Defense and more recently in other federal agencies.

1. An objective or objectives. What aims are we trying to accomplish with the equipment and personnel that the analysis is designed to compare?

2. Alternatives. By what alternative equipments, procedures, and/or personnel may the general objectives be accomplished? The alternatives are frequently referred to as systems.

3. Costs or resources used. Each alternate system requires certain costs or the using up of certain resources.

4. A model or models. Models are abstract representations of reality. They may assume the form of a complex set of mathematical relationships, or they may assume the form of small-scale representations of reality. For evaluation purposes the essential relationships are those between system inputs and system outputs. These are the relationships that should be represented by models.

5. A criterion. A criterion is a decision rule or test by which one chooses one alternative system rather than another.

Therefore one can conduct a comparative evaluation by applying the criterion to each of the various alternative systems. Furthermore, since an alternative that always exists is the "system" that would exist in the absence of any new or proposed systems, one can use this "system" as a frame of reference and thereby conduct an evaluation of a single new or proposed system.

Hitch and McKean have considered the above five elements in connection with evaluating military systems. We shall consider the five elements in connection with evaluating information retrieval systems.

3.5.2.1 Objectives

The objectives of an information retrieval system may be viewed from a number of levels. A higher level objective is to provide informational support in meeting organizational goals. Lower level objectives are to satisfy users' informational needs, requests, or wants. Regardless of the level of viewpoint, information retrieval systems play

a support role and are of value only insofar as they assist organizations or individuals in meeting their goals.

3.5.2.2 Alternatives or systems

A document retrieval system may be defined to be a combination of equipment, people and procedures for performing the following functions:

a. Document composition
b. Document production
c. Document storage
d. Document identification and location
e. Document presentation
f. Document assimilation

[See Part II, Measurement]

3.5.2.3 Costs or resources used

In arriving at estimates for the costs of a document retrieval system, it is convenient to consider two categories of costs: fixed or investment costs, and operating costs. The fixed costs are the one-time costs of developing, constructing and setting up the system while the operating costs are the recurring costs of running and maintaining the system year after year.

A convenient breakdown of operating costs for an information retrieval system was developed by Hertz [2] who considered the cost components for each of the following functions:

a. Encoding objects (or items)
b. Inserting them into storage
c. Encoding the search request
d. Preparing it for search
e. Searching the store

f. Identifying the retrieved objects

g. Appraisal of search results

h. Obtaining source objects

i. Reformulation or withdrawal of the search request

The operating costs, as well as operating time, depend strongly on the level of activity of the information retrieval system.

An acceptable cost analysis should involve: (1) selecting a planning period, and an interest rate, (2) estimating the fixed costs, the operating costs for each year of the planning period, and the residual value of the system at the end of the planning period, (3) on the basis of (1) and (2) computing the overall cost - either by the present value method or average annual cost method. A complete description of a cost analysis is beyond the scope of this paper. For more details of a cost analysis see Grant and Ireson [3].

In addition to cost, an important resource consumed in operating an information system is the user's time. This time is either time consumed in interacting with the system or waiting time for the system to respond to the search request.

3.5.2.4 Models

For an information retrieval system, the main outputs are the wanted information retrieved and the value derived therefrom. The main inputs are the system cost and the user's search effort required to retrieve the wanted documents.

3.5.2.5 Criterion

When both benefits and costs can be expressed in the same units, a generally acceptable criterion is to select the alternative which maximizes the net benefit, i.e., the benefit minus the cost. This choice is equivalent to maximizing the profit.

Although it is not generally possible to express the benefits objectively in the same units as the cost, it is instructive to assume that appropriate factors can be selected which convert the benefit expressed in non-monetary units to benefits expressed in monetary units. In this paper, therefore, it will be assumed that benefits and costs can be expressed in the same units, and that the net benefit is a suitable measure for evaluating alternatives.

A second measure which is often employed is the benefit-to-cost ratio. For example, Murdock and Liston [4] have proposed using the benefit-to-cost ratio as a suitable measure for evaluating information retrieval systems. It is therefore of considerable interest to determine the relationships between the net benefit and the benefit-to-cost ratio. Such relationships are developed below.

3.5.2.5.1 Relationships between the common cost-benefit measures

For an information system operating at a given performance level, let C and B denote the costs and the benefits, respectively. As the performance level varies, C and B also vary and can be conveniently described by a curve of B versus C. The bold lines on Figure 3.5.1 illustrate how B and C might vary for two alternate information systems, I and II.

Two measures which are readily derived from B and C are the net benefit or profit, P, and the benefit to cost ratio, R. These measures may be computed as follows:

$$P = B - C \qquad (3.5.1)$$

$$R = B/C \qquad (3.5.2)$$

From (3.5.1) it follows that lines of constant P on Figure 3.5.1 are lines with a slope of unity and intercept of P. From (3.5.2) it follows that lines of constant R on Figure 3.5.1 pass through the origin and have a

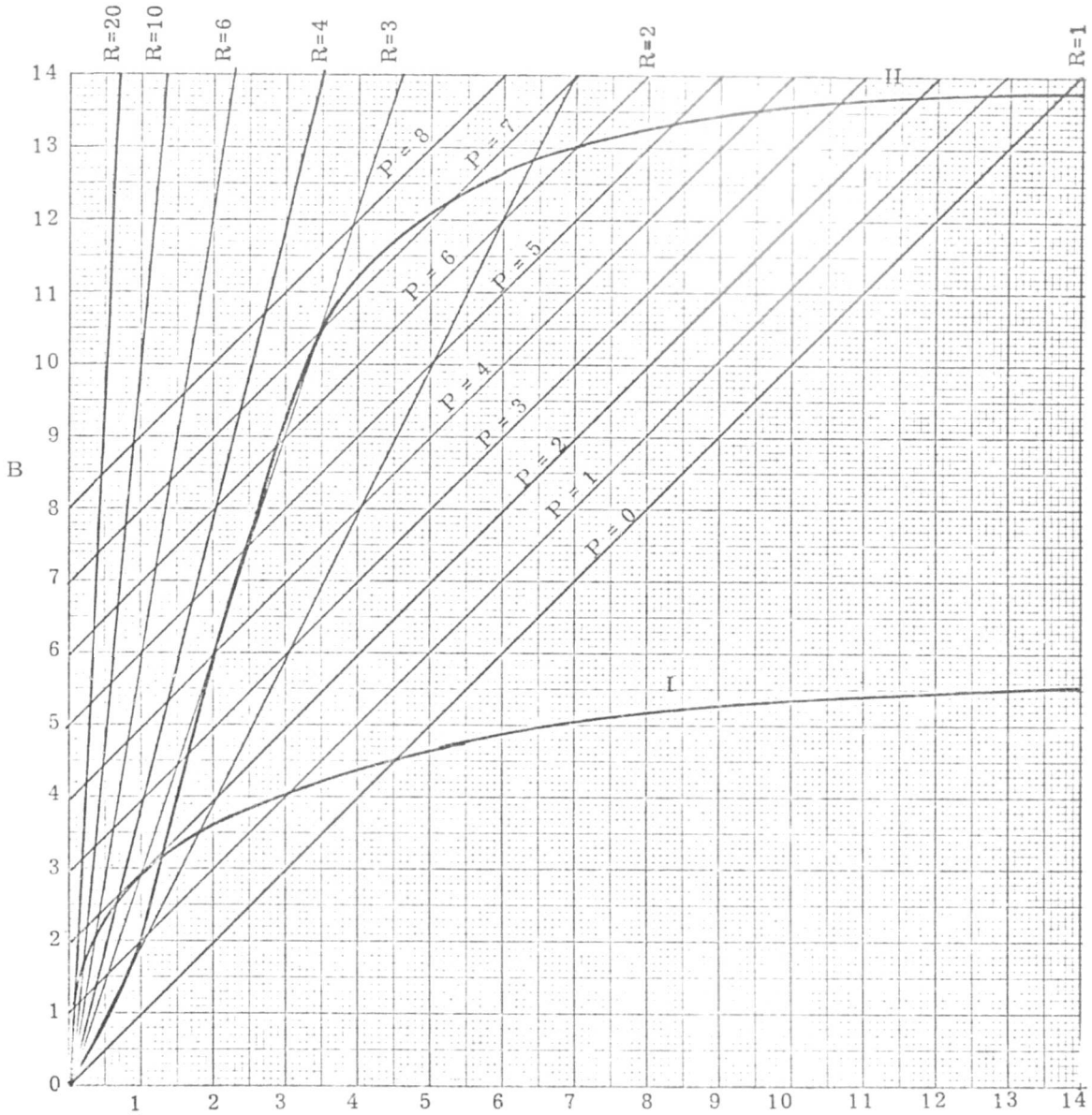


Figure 3.5.1. Benefit-Cost Relationships for Two Alternate Information Systems.

slope of R. Consequently Figure 3.5.1 can be used to study how the benefit, the net-benefit, and the benefit-to-cost ratio simultaneously vary with cost (and therefore with performance level).

For a situation where adequate resources are available for the selection of one of several alternate systems, the generally accepted measure of relative economic worth is the net benefit, i.e., the benefit minus the cost. The optimum choice in this case consists of selecting the alternative system which maximizes the net benefit, P.

For the two information systems represented in Figure 3.5.1 one can see that for System I the maximum net benefit 2 and occurs at a cost of 0.8 while for System II the maximum net benefit is over 7 and occurs at a cost of 5. Hence, System II is preferable to System I if only one of these systems is to be selected.

In general maximizing the benefit to cost ratio does not yield the optimum choice. However, one instance where this choice would be optimum is the following: A number of Type I or Type II systems are to be purchased for a fixed budget and the total benefit equals the benefit per system times the number of systems. Here the total net benefit is maximized by selecting the number of Type I systems which just consume the fixed budget.

3.5.3 Models for evaluating the performance of document retrieval systems in conducting elementary searches

An information retrieval system may be used for a number of different purposes, in a number of different subject areas, and by a number of different types of users. For example, an information retrieval system may be used for any of the following purposes:

a. To retrieve a specific document whose author and title are known.

b. To retrieve a specific document known to exist but where author and/or title are unknown.

c. To retrieve some documents containing information on a given subject.

d. To retrieve all documents within the physical file containing information on a given subject (internal bibliographic search).

e. To retrieve all documents both within and without a physical file containing information on a given subject (internal and external bibliographic searches).

f. To receive a sampling of recent documents containing information on one or more subjects to become aware of current developments.

g. To browse.

Use may also be classified according to subject matter, user characteristics, and so on.

An information retrieval system may perform well for some types of use and poorly with respect to others. Also, the value of the output may be great for some types of searches and low for other types of searches. A well designed information retrieval system should have the property that it performs well for those types of searches which have a high value. Therefore, in evaluating an information retrieval system it is important to distinguish between the various types of use.

Precisely how these distinctions should be made, i.e., what the ultimate breakdown or categorization of use should be, is a matter which should be resolved by further research. However, whatever the ultimate breakdown of use may be, it will consist of a number of elementary types of use. For example, if use should be categorized only according to type of purpose, and if the suitable categories turned out to be a. to g. above, then there would be seven elementary types of use.

In this section attention is focused on a number of searches, all of which are restricted to a particular elementary type of use. These

searches are called elementary searches. For example, if the above categorization of uses were considered appropriate, then the collection of all internal bibliographic searches, i.e., d above, would constitute an elementary class of searches. Corresponding to the other six categories would be six more elementary classes of searches.

The simplest type of document retrieval system is one which performs searches belonging to only one elementary class of searches. This type of system may be called an elementary document retrieval system. Investigators who make no distinctions between various types of uses, in effect, are assuming that they have an elementary document retrieval system.

The remainder of this section considers four types of models for elementary document retrieval systems:

a. The Search Characteristic Curve, a model that relates the search return to search effort for various levels of system performance

b. The Acquisition Characteristic Curve, a model that relates the return to the file size

c. The User Benefit Curve, a model that relates user benefit to search return, and

d. The Net Benefit Model, a model that relates net benefit to the level of system performance.

3.5.3.1 The Search Characteristic Curve

Each use of a document retrieval system by a user may be considered to be a search. Browsing, for example, is a haphazard kind of passive search while dissemination of information for current awareness is an active type of search. In any event, a search supplies the user with a number of documents or document representations, such as abstracts or titles. For a given class of elementary searches it will be assumed that the output of the system consists of a number of items that

are examined at least in part by the user so that he judges each item with respect to relevance of the item to the user's search wants. In general, there will be k different levels of relevance ranging from not relevant to extremely relevant. A very common approach is to assume that k is equal to 2, and that the two levels of relevancy of items are "relevant" and "not relevant". For the moment it will be assumed that k = 2.

The total number of items examined by the user is a measure of the user's search effort. As a result of this search effort the user will reap certain returns. A measure of these returns is the number of relevant documents found among those examined. The relationship between the search return, expressed by the number of relevant documents examined, and the search effort, expressed by the number of documents examined, is called the "search characteristic curve."

Properties of the search characteristic curve have been studied and presented previously [Section 3.2]. As a result of this study, the following useful normalized form for a search characteristic curve was derived:

$$r_F = 1 - (1 - f^{1/k})^b \qquad (3.5.3)$$

where

$r_F = \frac{m}{M_F}$, the fraction of relevant documents in the file examined (or recall)

$f = \frac{n}{N_F}$, the fraction of the file examined

N_F = the total number of documents in the file

M_F = the number of relevant documents in the file

n = the total number of documents examined

m = the number of relevant documents examined when the number of documents examined is n

k, b = the search performance parameters, i.e., parameters which describe the performance of an information retrieval system. The larger the value of these parameters the better the performance of the information retrieval system.

The equivalent unnormalized form of the search characteristic curve is:

$$m = Mr_F = M_F \left\{ 1 - \left[1 - \left(\frac{n}{N_F} \right)^{1/k} \right]^b \right\} \qquad (3.5.4)$$

Equation (3.5.4) permits one to compute the expected number of relevant documents that will be retrieved for a given elementary class of searches as a function of the number of documents examined, n, the number of relevant documents in the file, M_F, the size of the file, N_F, and the search performance parameters. Thus, measures described in Part II, Measurement, yield inputs to the models described in this section.

3.5.3.2 Acquisition characteristic curves

Search characteristic curves relate the returns resulting from a given search policy to the number of documents examined. A similar relationship exists between the returns resulting from a given acquisition policy and the number of documents acquired, indexed, and filed. This latter relationship will be called the acquisition characteristic curve. Just as the search characteristic curve describes the effectiveness of a search policy, the acquisition characteristic curve describes the effectiveness of the acquisition policy. The purpose of this subsection is to develop the notion of the acquisition characteristic curve.

For a given body of knowledge which is to be covered by a given information retrieval system for a given body of users, let N be the total number of documents, and let M be the number of documents wanted by the body of users. The total number of documents N can be

subdivided into the M wanted documents and the remaining N-M unwanted documents. For example, documents written in a foreign language which cannot be translated feasibly would constitute a part of the unwanted documents.

If the acquisition policy is such that every single one of the N documents is acquired, then all of the M wanted documents will be a part of the file of the information retrieval system. Although such a policy provides perfect coverage of wanted documents, it can be exceedingly costly, especially when N is very large, and therefore that policy is rarely adopted.

A more feasible acquisition policy involves acquiring only a subset of the N documents. Let N_F be the number of documents in such a subset, i.e., the number of documents that are acquired, indexed and placed in the file. If N_F is less than N, there is, of course, a chance that not all of the M wanted documents will be a part of the file, i.e., the number of wanted documents in the file is less than M. Let M_F be the number of wanted documents which are part of the file. Note that M_F is a measure of the return resulting from certain acquisition.

The relationship between M_F and N_F for a family of acquisition policies will be defined to be the acquisition characteristic curve, and is represented graphically in Figure 3.5.2.

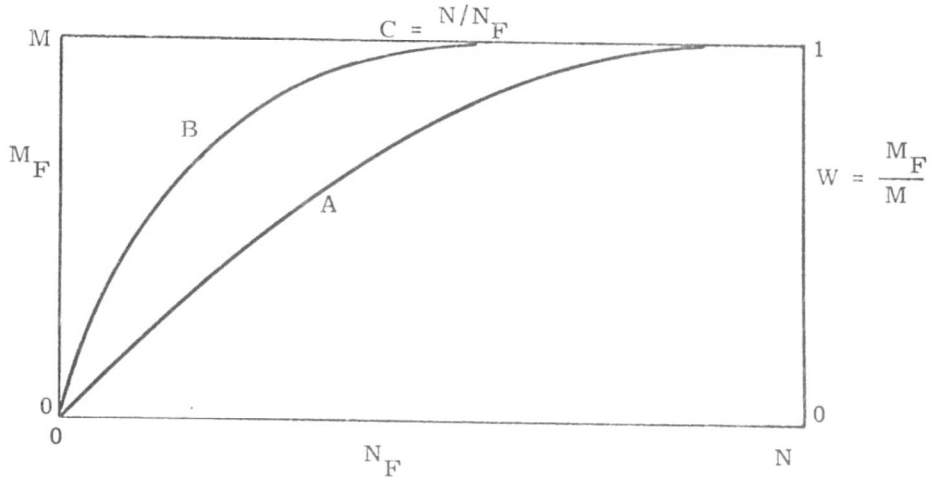


Figure 3.5.2. Acquisition characteristic curve.

The normalized version of the acquisition characteristic curve is also shown in Figure 3.5.2 and is obtained by dividing the abscissa by N and the ordinate by M. The resulting ratios may be defined to be the coverage, C, and the fraction of wanted documents contained in the file, W, and are given by the following equations:

$$C = \frac{N_F}{N} \tag{3.5.5}$$

$$W = \frac{M_F}{M} \tag{3.5.6}$$

Applying the same rationale used to obtain analytical expressions for the search characteristic curves, it can be shown that the acquisition curves of Figure 3.5.2 can be represented by the following equation:

$$W = 1 - (1 - C^{1/h})^{a} \tag{3.5.7}$$

3.5.3.3 User benefit curves

Each search generates a number of relevant documents from which the user derives a certain amount of benefit. If the user obtains all the relevant documents as a result of a search, the user benefit derived from the search is a maximum. On the other hand, if the user obtains only a fraction of the relevant documents as a result of a search, the user benefit derived from the search is a fraction of the maximum benefit. Typical reasons for a search producing only a fraction of the maximum number of relevant documents are the following:

a. The file against which the search is processed contains only a fraction of the relevant documents, i.e., the file coverage is not 100 percent.

b. Errors were made either in indexing the document or in preparing the query.

c. The depth of search was not sufficiently great to retrieve all the relevant documents in the file.

Consider how the user benefit derived from a search varies as the fraction of relevant documents examined varies from zero to unity. If none of the relevant documents is retrieved, we may assume that the user benefit is zero. As more and more of the relevant documents are retrieved, the user benefits rise continuously. When all of the relevant documents are retrieved, the user benefits will certainly achieve their maximum value.

Because it is difficult and perhaps unreasonable for the user to determine the maximum value of the benefit derived from a search that gives him all the relevant documents, it is not convenient for the user to think in terms of the fraction of maximum user benefit. The fraction of maximum user benefit, v, will increase steadily from zero to unity as the fraction of relevant documents examined, r, varies from zero to unity. A method for arriving at a relationship between v and r is by asking the user to estimate subjectively values of v for each of a series of values for r. But first it is helpful to relate r, the fraction of the universe of relevant documents, to r_F, the fraction of relevant documents in the file, and W, the file coverage. By definition, it follows that

$$r = \frac{m}{M} = \frac{m}{M_F} \cdot \frac{M_F}{M} = r_F \cdot W \qquad (3.5.8)$$

Since subjective estimates may be expected to lack precision, it appears appropriate to express the relationship between v and r by the following equation:

$$v = 1 - (1 - r)^u \qquad (3.5.9)$$

where u is a parameter that describes the rate of increase of v with respect to r. For small values of r, u is approximately equal to the rate of increase of v with respect to r. The relationship between v and r is called the fractional user benefit curve.

A few examples will clarify the meaning of fractional benefit

curves and the parameter u. As a first example, consider a user interested in finding one or at most two relevant documents on a given subject. Additional relevant documents will be of very little additional benefit for his search needs. Furthermore, suppose that the total number of documents relevant to his search request is 50. Since the user is interested in obtaining at most 2 of 50 relevant documents, he is interested in a recall, r, of only 2/50 or 0.04, a sufficiently small value of r for Equation (3.5.9) to be applicable, i.e.,

$$u \cong \frac{v}{r} = \frac{1}{0.04} = 25 \qquad (3.5.10)$$

As a second example, consider a user who is interested in conducting an exhaustive bibliographic search on a given subject. Each additional relevant document, on the average, will add equally to his benefit, and only when he obtains all of the relevant documents will his benefit reach its maximum. In this case the fractional user benefit is equal to the fraction of relevant documents retrieved, i.e., v = r. This is equivalent to using a value of u equal to unity in (3.5.10).

3.5.3.4 A net benefit model

The net benefit is defined as the benefit minus the costs. For this subtraction to be meaningful, the benefit and the cost should be commensurable, i.e., they should have the same units. To convert the fractional user benefit to the user benefit expressed in dollars, a dollar value should be assigned to the maximum user benefit, and the fractional user benefit should be multiplied by this dollar value.

And what should this dollar value be? Assuming that the objective of the information retrieval system is to provide informational support in meeting organizational goals, this dollar value should be the price that the managing body of the organization is willing to pay for the service provided by the particular type of search for the particular type of user.

Usually there are other means of obtaining similar services for which the price can be readily established, and this can be used by the managing body as a frame of reference for arriving at an appropriate price. For example, additional research assistants can be hired to conduct searches using services that would be available in the absence of the information retrieval system. High caliber research assistants who provide a high quality probably could be attracted by high salaries, and it should be possible to match approximately the quality of search (except for response time) of the system with an appropriate caliber research assistant. From the salary and time requirements of such a research assistant and any additional expenses to be incurred in executing a search, the cost of a search can be readily computed.

If the response time of the search conducted by the research assistant and the search conducted by the system were the same, the cost of the search conducted by the research assistant would be a reasonable price to use. However, if the system's response time for a search is substantially less than the research assistant response time for a search, it may be desirable to pay a premium for more rapid service, the amount of the premium depending on the urgency of the search results. In any event, the managing body of an organization could reasonably establish an appropriate price to pay for the maximum user benefit derived from a particular type of search by a particular type of user. Let this price be denoted by B. Then for a search that yields a fractional user benefit of v, the user benefit per search expressed in dollars is $B \cdot v$. For s searches, therefore, the total benefits, B_T, is simply:

$$B_T = B \cdot v \cdot s \qquad (3.5.11)$$

Although (3.5.11) appears to be a simple expression, it should be remembered that v requires a string of relationships, viz (3.5.3), (3.5.7), (3.5.8), and (3.5.9) to be evaluated. This point will be brought out subsequently in an example.

3.5.4 A comprehensive cost model

The cost for a simple document retrieval system is now considered. The average annual cost of a simple document retrieval system consists of two components: a fixed cost and a variable cost. How the cost is resolved into these components depends on what management policies are adopted.

In this paper, it is assumed that the costs vary with respect to three factors: (1) the number of documents in the file, N_F, (2) the number of searches per year, s, and (3) the number of documents examined per search, n. All other costs are assumed to be fixed. More specifically, it is assumed that the total annual cost, C_T, may be expressed as follows:

$$C_T = C_0 + C_1 \cdot N_F + (C_2 + C_3 \cdot n) \cdot s \qquad (3.5.12)$$

where C_0 = fixed cost

C_1 = the marginal cost of increasing the size of the file by one document

C_2 = the contribution to the cost per search which does not vary with the number of documents examined

C_3 = the marginal cost per search of increasing by one the number of documents examined in a search.

The annual net benefit, P_T, is defined to be the difference between B_T and C_T, i.e.,

$$P_T = B_T - C_T \qquad (3.5.13)$$

Substituting (3.5.11) and (3.5.12) into (3.5.13) yields the following expression for the net benefit:

$$P_T = (B \cdot v - C_2 - C_3 \cdot n) \cdot s - C_1 \cdot N_F - C_0 \qquad (3.5.14)$$

One application of equation (3.5.14) is to compare the relative merits of alternative information retrieval systems. If all of the cost and benefit parameters required to compute P_T are available for one or more simple information retrieval systems, the value of P_T can be estimated for each system using (3.5.14).

For several simple information retrieval systems designed to fulfill the same objectives, the system with the largest value of P_T is preferred. However, since P_T may be viewed as a random variable, the uncertainty in the P_T values of the various alternative systems should also be considered before one system is claimed to be better than all the others -- for there may be no statistically significant difference between the values.

Another use of (3.5.14) is to predict improvements in the operation of a given information retrieval system by estimating the increase in net benefit associated with changes in various system parameters and decision variables. For small changes in these quantities the rate of increase in net benefit with respect to the quantity can be obtained by taking derivatives of P_T given by (3.5.14).

These two applications of Equation (3.5.14) are illustrated by a hypothetical example below.

3.5.5 Hypothetical example

Assume that the performance of a simple information retrieval system is characterized by the following relationship and data:

Search Characteristic Curve

$$r_F = 1 - (1 - f^{1/2})^{10} \qquad (3.5.15)$$

Acquisition Characteristic Curve

$$W = 1 - (1 - C)^5 \qquad (3.5.16)$$

File Size

$$N_F = 100,000 \text{ documents}$$

Total Number of Documents

$$N = 200,000 \text{ documents}$$

User Benefit Curve

$$r = W \cdot r_F \tag{3.5.17}$$

$$v = 1 - (1 - r)^2 \tag{3.5.18}$$

Estimated Benefit Per Search, B

lower estimate -- \$10

upper estimate -- \$20

Costs

overall fixed costs, C_0 = \$25,000

cost per document, C_1 = \$0.25

fixed cost per search, C_2 = \$2

cost per search per document, C_3 = \$0.03

Number of Searches Per Year

$s = 10,000$

Typically, the number of documents examined, n, is 500. The above parameters and data are summarized in Table 3.5.1 below.

Table 3.5.1 Summary of hypothetical parameters and data

Search	$k = 1/2$ $b = 60$ $n = 500$
Acquisition	$h = 1$ $a = 5$ $N_F = 100,000$ $N = 200,000$
Benefits	B = \$20 to \$40 $u = 5$
Costs	C_0 = \$25,000 C_1 = \$0.25 C_2 = \$2 C_3 = \$0.03
Demand	$s = 10,000$

Since f is 500/100,000, r_F is 0.52 from (3.5.15). Since C is 0.5, W is 0.969 from (3.5.16), and from (3.5.17) r is 0.50. For r equal to 0.50, v equals 0.753 by (3.5.18). According to (3.5.11), therefore, B_T should be between

$$B_T^l = \$20 \cdot (0.753)(10,000) = 150.6K \qquad \text{low estimate}$$

and

$$B_T^n = \$40 \cdot (0.753)(10,000) = 301.4K \qquad \text{high estimate}$$

The costs computed according to (3.5.12) are:

$$C_T = \$25K + \$0.25 \times 100K + \$(2 + .03 \times 500)10K$$
$$= \$220K$$

Hence the net benefit lies between

$$P_R^l = \$199.6K - \$220K = -\$69.3K \qquad \text{low benefit}$$

and

$$P_T^h = \$399.3K - \$220K = \$81.4K \qquad \text{high benefit}$$

Thus, one can see that the system under the conditions specified above is questionable and not clearly justified from a net benefit analysis standpoint. Suppose, however, that the system is modified to yield the same recall (0.52) with total retrieval of 500 documents reduced to 100 documents. Also, assume that this improvement increases s to 20,000 searches, increases fixed costs to \$50,000, and increases fixed cost per search to \$4.

$$C_T = \$50K + 0.25 \times 100K + \$(4 + 0.03 \times 100)10K = \$215K$$

$$B_T^l = \$20(0.753)(20,000) = \$301.4K \qquad \text{low estimate}$$

$$B_T^n = \$40(0.753)(20,000) = \$602.7K \qquad \text{high estimate}$$

$$P_T^l = \$301.4K - \$215K = \$86.4K \qquad \text{low benefit}$$

$$P_T^h = \$602.7K - \$215K = \$387.8K \qquad \text{high benefit}$$

Thus, the system modification would appear to be entirely justified from the standpoint of net benefit.

As indicated in the first example, the uncertainty in benefits per search may be crucial in estimating the net benefit. However, when comparing one system with another, this uncertainty is not nearly so important, for one would expect that the same relative ranking of systems according to net benefit would be preserved whether or not the high or low estimate is employed.

To continue with the first example, assume that the benefit per search is equal to the break-even value ($220K), i.e., the value for which P_T is zero. This value is readily found to be approximately B = $29. For this value of B, questions such as the following may be asked:

a. What benefit could be gained by enlarging the size of the file, and maintaining the same policy?

b. Should users, in general, examine more than 500 documents, the typical number or less?

c. If the search policy can be improved to the extent that b is increased by one unit, how much would the net benefit increase?

d. If the acquisition policy can be improved to the extent that a is increased by one unit, how much would the net benefit increase?

Answers to questions such as these should be particularly useful in deciding what direction of development appears to be most fruitful.

Questions a and b are now considered in terms of the above hypothetical information retrieval system.

a. The effect of file size

The effect of file size on the net benefit can be obtained by differentiating P_T with respect to N_F while holding the other parameters

in Table 3.5.1 fixed. From (3.5 14) and (3.5.12) it follows that:

$$\frac{\partial P_T}{\partial N_F} = P \cdot S \cdot \frac{\partial v}{\partial N_F} - c \qquad (3.5.19)$$

From (3.5.3) through (3.5.9) and for k = 1/2, h = 1 it can be shown that

$$\frac{\partial v}{\partial N_F} = -u(1-r)^{u-1}\left[w \cdot b \cdot (1-f^k)^{b-1} \cdot \frac{n}{N_F^2} \cdot kf^{k-1} r_F\, a(1-c)^{a-1} \frac{1}{N}\right]$$

Using values from Table 3.5.1 for these parameters yields:

$$\frac{\partial v}{\partial N_F} = -2(1-r)\left[w \cdot 10(1-f^{1/2})^9 \cdot 1/2 f^{-1/2} \cdot \frac{n}{N_f^2} - r_f \cdot \frac{5 \cdot (1-c)^4}{N}\right]$$

and in view of (3.5.19)

$$\frac{\partial P_T}{\partial N_F} = (29)(10,000)(.00032) - (0.25)$$
$$= \$93.40$$

Therefore, at the operational conditions shown in Table 3.5.1, increasing the file size by one document will increase the net benefit by $93.40.

b. Effect of examining more documents

From (3.5.14) it follows that

$$\frac{\partial P_T}{\partial n} = B \cdot s \cdot \frac{\partial v}{\partial N} - c_3 \cdot s \qquad (3.5.20)$$

From (3.5.9), (3.5.8) and (3.5.3) for R = h = 1 it can be shown that

$$\frac{\partial v}{\partial n} = u \frac{(1-r)^{u-1}(1-f)^{b-1}}{N_f} b \cdot w$$

and in view of (3.5.20)

$$\frac{\partial P_T}{\partial n} = (29)(10,000)(-.000092) - (.03) \cdot (10,000)$$
$$= \$-326.80$$

Therefore, at the state of operation shown in Table 3.5.1, if each user examines one more document the net benefit should increase by $-326.80.

Similarly, by differentiating P_T with respect to other parameters one can predict how much the net benefit should increase per unit increase in these parameters. For example, $\frac{\partial P_T}{\partial b}$ is an estimate of the increase in net benefit per unit increase in b, a parameter that characterizes the effectiveness of the search policy; $\frac{\partial P_T}{\partial a}$ is an estimate of the increase in net benefit per unit increase in a, a parameter that characterizes the effectiveness of the acquisition policy; $-\frac{\partial P_T}{\partial C_i}$ for i = 0, 1, 2, 3 are estimates of how much the net benefit will decrease per unit increase in the cost C_i; and $\frac{\partial P_T}{\partial s}$ is an estimate of how much the net benefit will increase if the number of searches per year can be increased by one. From a knowledge of these estimates, one obtains an idea concerning what areas of research and development should yield the greatest payoff measured in terms of net benefit. For example, improving the acquisition policy to the extent that the parameter a increases by one unit may yield a much greater increase in net benefit than the same increase in b. This would suggest that doing more research and development in improving the acquisition policy should be more fruitful than doing research and development in improving the search policy.

References

[1] Hitch, Charles J., and Roland McKean, (1960), The Economics of Defense in the Nuclear Age, Cambridge, Mass.: Harvard University Press.

[2] Arthur Andersen & Co., "Research Study of Criteria and Procedures for Evaluating Scientific Information Retrieval Systems". Prepared for the Office of Science Information Service, National Science Foundation, Final Report March 1962.

[3] Grant, Eugene L. and W. G. Ireson, (1964), Principles of Engineering Economy, 4th edition, New York, Ronald Press.

[4] Murdock, J. W. and David M. Liston, Jr. (October, 1967), "A General Model of Information Transfer", American Documentation, 18, 197-208.

END

DATE FILMED

4-12-69