73-23,919

PEDERSON, John Alvin, 1926-
PERFORMANCE - COST - VALUE DECISION PARAMETERS OF REFERENCE RETRIEVAL SYSTEMS.

The University of Oklahoma, Ph.D., 1973
Engineering, industrial

University Microfilms, A XEROX Company, Ann Arbor, Michigan

THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PERFORMANCE - COST - VALUE DECISION PARAMETERS OF REFERENCE RETRIEVAL SYSTEMS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

JOHN ALVIN PEDERSON

Norman, Oklahoma

1973

PERFORMANCE - COST - VALUE DECISION PARAMETERS

OF REFERENCE RETRIEVAL SYSTEMS

APPROVED BY

DISSERTATION COMMITTEE

ACKNOWLEDGEMENTS

I wish to acknowledge the comprehensive guidance and direction that I have received from Dr. B. L. Foote, which is evidenced by the existence of this dissertation.

Dr. Robert Shapiro's perspective for visualizing problems with the objective of providing rational solutions has been a stimulus on many occasions.

Professor Bruno's knowledge of libraries and their functional operations has provided several avenues for understanding reference retrieval systems.

Professor Astle presented many perspectives in the application of statistical techniques.

Mr. Baer's familiarity with the use of mathematical techniques and computer programming has been invaluable.

Professor Johnson's painstaking review of this manuscript has significantly reduced the number or the width of the gaps in the information flow of this document.

The additional work undertaken by my wife, Phyllis, for the seemingly endless revisions to the manuscript has made this dissertation possible. Also the loss of association with my daughters as "Dad" worked on this document is acknowledged.

TABLE OF CONTENTS

	Page
LIST OF TABLES	ix
LIST OF ILLUSTRATIONS	xi
Chapter	
I. INTRODUCTION	1
II. HISTORICAL BACKGROUND AND ANALYSIS OF INFORMATION RETRIEVAL SYSTEMS AND MODELS	8
Performance Factors	8
Index Terms	8
Frequency of Usage of Each Index Term	9
Index Term Vocabulary Size Versus Document-Collection Size	11
Search Terms	13
Query Intensity and Time-Rate Usage of Documents	13
Vocabulary Usage in User Query Formulation	15
Search Strategy and User Query Formulation	15
Evaluation of System's Performance	17
Recall-Pertinence, Relevance	17
Expansion of Recall-Relevance Ratios	20
Performance of Information Retrieval Systems	22
Aslib Cranfield	23
MEDLARS	25
Comparative Systems Laboratory	28

Page

Model to Calculate the Number of References to Documents to be Retrieved From a Retrieval System 30

Uses 33
Limitations 34

Total Cost-Value 35

Cost Models 35
Cost Data 41
Value 44
Optimization 48

Conclusions Based on Review 50

III. PROBLEM FORMULATION AND MODEL DEVELOPMENT 53

Analysis of Reference Retrieval Systems 53

Index Term Usage Errors 55
Index Vocabulary Usage Errors. . 66
Search Formulation 71

Intersection Search 71
Union Search 72

Level of Performance Model Output 73

Proposed Model 77

Reference Retrieval Model . . . 78

Error Determination Technique 78
Performance Model 78
Output Evaluation 83

Total Cost-Value 83

Total System Production Functions 84
Total Value of Output Benefits 85
Optimization 86

Page

IV. REFERENCE RETRIEVAL MODEL 87

Input Error Categorization Technique 88

Intellectual Phase 88

Discipline Designation . . . 88
Document Determination Concepts 90

Analytic and Development Phase 91

Functional Relationships . . 91
Testing Procedure and Experiment Conduct 93
Parameter Determination . . . 94

Performance Model 95

Input Description 97
Initial Operational Stages . . . 101

Error Determination 101
Designation of Search Term Rank 111
Category Designation of Document Index Terms . . . 113

Output Preparation 116

Interaction 116
Intermediate Output 121
Final Output Preparation. . . 122

V. REFERENCE RETRIEVAL SYSTEM APPLICATION 125

Objective 125

Model Stability 126
Input Errors 126
Output Dependence on Inputs. . . 127
Experimental Design for Performance Surface 130

Objective 130
Hypothesis 131

Page

Data Sources. 132

Error Determination. 132
Performance Model. 133

Results of Experiments. 136

Model Stability. 136

Number of Simulation Runs . . 136
Sample Size 138

Relationship of Variables. . . . 141

Input Errors. 141
Output Dependence 141

Performance Surface. 145

Conclusions 150

Performance Model Stability. . . 150
Error Sources. 151

VI. TOTAL COST OF FACILITY - TOTAL VALUE OF OUTPUT BENEFITS MODEL. . . 153

Total System Production Functions . 157

Total Facilities 157

Capital Equipment and Development 157
Operating Functions 160

Total Output Benefits Functions. 165

Total System Cost Functions 167

Total Cost of Facilities 168
Total Value Output Benefits. . . 175

VII. APPLICATION OF TOTAL COST-VALUE MODEL 177

Data Sources. 177

Profit Maximization 178

Page

Analysis of Results 182

Overall Review 183
Impact of Indexing and Search Terms 184
Profit Maximizing Values 188

VIII. CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER WORK. 192

Conclusions 193

Error Detection Technique Limitations. 194

Indexing and Searching. 196
Error Detection 197

Performance Model. 197

Optimization Procedure 198

Time Dependent Factors. 198
Value-Cost Considerations . . . 199

Specific Suggestions For Further Work. 200

APPENDIX 202

REFERENCES 203

LIST OF TABLES

		Page
2.1	Recall-Relevance	18
2.2	Recall-Relevance (pertinence) Weight Factors.	20
2.3	System Components - Constant and Varied. .	29
3.1	Parameters and Factors of the System . . .	79
4.1	Summary of Evaluated Experimental Data . .	96
4.2	Performance Model.	98
4.3	Search Term Input Values	103
4.4	Search Term Output	104
4.5	Index Term Input and Output Values	109
4.6	Index Term Categories and Cells.	110
4.7	Document Identification Number Vs. Search Terms.	114
4.8	Document Index Term Categories	115
4.9	Documents Indexed Under Search Terms . . .	117
4.10	Document Identification No's Recovered Vs. Input Categories.	119
4.11	Retrieved References to Number of Documents of Each Category.	120
5.1	Numerical Values for Parameters and Factors of the System	134
5.2	Experimental Data.	137
5.3	Test Data for Output Dependence.	143
6.1	Total System Production Functions	
	a. Total Facilities.	161
	b. Total Output Benefits	163
6.2	Total System Cost Function and Data Model Application	

Page

a. Total Cost of Facilities. 172
b. Total Value of Output Benefits. 174

7.1 Cost or Value of Various Items Needed for Application of TVOB-TCF System. . . 179

7.2 Solution Cost-Value-Profit Summary 189

LIST OF ILLUSTRATIONS

		Page
2.1	Pertinency - Relevance	19
2.2	Frequency Distribution of Term Usage in Vocabulary	32
3.1	Input Data Categories.	57
3.2	Ideal Relationship of Needed Index Terms .	59
3.3	Relationship of Used Terms	61
3.4	Combined Effect, All Categories of Terms .	63
3.5	Error Regions Vs. The Number of Terms Used in Indexing a Document	65
3.6	Linkages Between Documents, Index Terms and Their Frequency of Usage.	68
3.7	Vocabulary Index Term Frequency.	70
3.8	Performance Model.	82
4.1	Input/Output Classification Scheme	99
4.2	Search Term Category and Frequency Distribution.	107
5.1	Feasible Region.	128
5.2	Mean and Standard Deviation Vs. Variables.	139
5.3	Number of Terms Vs. Error.	142
5.4	Error in Output Related to Level of Input.	146
5.5	Feasible Region of Output.	148
6.1	Indexing Vs. Time.	158
6.2	Cost of Indexing	169
7.1	Plan View of Solution Profiles	185
7.2	Indexing Cost-Value-Profit	186
7.3	Searching Cost-Value-Profit.	187

PERFORMANCE - COST - VALUE DECISION PARAMETERS OF REFERENCE RETRIEVAL SYSTEMS

CHAPTER I

INTRODUCTION

The inauguration of the "information explosion" has expanded the data-handling industry and placed significant demands on the various related services. In the past these services such as acquisition, cataloging, classification, and retrieval of documents were the domain of the librarian. The ever-increasing needs and demands for information and data have brought this entire phase of specialized endeavor into direct contact with many other aspects of modern life on a much larger scope. In addition, more types of data are being handled systematically.

The differences in types of information indicate some distinguishing aspects of information levels in relation to form, content, and functions. Based on their degree of amenability to various types of mechanization, H. P. Luhn [1] has listed six levels of information in order of increasing complexity:

1. Ready reference look-up of facts; indexes, dictionaries, and catalogs,
2. Limited and narrowly defined categories of fact, especially where the categories are repetitive for each document (e.g., specification lists),

3. Inventories of uniquely definable structures and their interrelations and transformations (e.g., chemical structure),

4. 'Disciplined' concepts; mathematics, logic, and law,

5. Information about the exploitation of natural phenomena and applied services,

6. Unrestricted association of human notions (e.g., fiction).

The advent of computers has provided facilities for faster handling of larger volumes of data. Numerical data are usually much more amenable to manipulation than textual data, and their retrieval is referred to by Lancaster [2] as "data" retrieval." In contrast, textual data, which by nature is not as definitive, can be recovered by "information retrieval," and such a system could include the following:

1. document numbers,

2. citations,

3. full texts.

A "reference retrieval' system retrieves citations and document numbers whereas a "document retrieval" system, as indicated, retrieves the full text of selected documents, with the library evolving into an information system. The design and implementation of information retrieval systems for present needs, as well as consideration of the problems that will be created by impending growth demands of the future, are imperative. Channeling of interest patterns into defined areas has also drawn material from the library to

specialized information centers dealing with a specific subject area.

It is recognized that, in any given system such as that described by Sharp [3], the indexing terms for describing the document must be consistent with the recall terms if retrieval is to be obtained. In addition, the system must be capable of processing both of the above in a compatible manner. The size of present and future demands for resources designed to meet the user needs indicates that ever-increasing funds will be needed. Based on past experience, a planned system can be operated at a given level of efficiency. It would seem that a determination of the value of output in terms of quality and quantity would be a subject of interest to actual and to potential system users. Trade-offs of the various parameters of cost and value could be equated.

These factors of systems are described by Murdock and Liston [4] as cost, performance, benefits, and their interrelationships. Costs describe the expense of operating the information system in terms of dollars. Performance measures describe the attributes that are controllable by the system, such as accuracy, usage applicability, speed, quality, and extent of coverage. Benefits describe the consequences of the system in terms of (1) how human effort can be reduced, (2) how the system affects the behavior of persons in allowing new ideas to be formulated, and (3) how the system affects related systems, such as planning and decision

making. The authors, Murdock and Liston, also present several general approaches to optimizing these factors.

The necessity of having some measure of total system evaluation is shown by Johoda [5], who set up a series of n categories, A_1, A_2,---,n, each of which describe a reason why one or more organizations have changed from one system to another in a group of m systems, B_1, B_2, ---,m. Analysis of this data showed that in 80 percent of these n categories, instances were found where one organization had changed from system B_i to B_j and B_j to B_i. This change indicates inconsistency in defining the objectives of information retrieval systems by organizations.

In this study a model has been designed to relate two classes of activity of a reference retrieval system to the output. The objective is to maximize the efficiency of the system, given the relevant physical parameters and variables of the system and cost of the inputs and price of the outputs.

Data derived from literature and estimates based on real models are used to demonstrate the feasibility of application of the model.

The model consists of two operational segments which operate in conjunction with each other. One segment is the total cost-total value model for optimizing the levels of usage of the reference retrieval model. The other segment is the reference retrieval model, which consists of three

stages: the input error determination technique, the performance model, and the output evaluation procedure.

The first stage of the reference retrieval model is an error-determination technique. This procedure is designed to ascertain the amount and extent of errors incurred as documents are indexed and user searches are formulated by use of a fixed vocabulary with a prescribed level of indexing. Two types of errors encountered are those of commission and omission, which are the inclusion of unneeded terms to describe the contents of a document or to formulate a desired search and the lack of inclusion of terms needed. Errors are determined on review by an analyzer who ascertains the applicability or lack of applicability, assuming certainty on his part.

The performance model is the second stage of the series of phases. This technique simulates the operation of indexing documents for a reference retrieval system. This phase is accomplished by techniques that quantify the inputs and outputs.

The third stage consists of the application of the first two phases to determine the interrelationships of the inputs and their associated errors to the level of output of the performance model and its errors.

The exogeneous constants are specified, along with endogeneous constants and variables. The variables are the number of index and search terms used in indexing documents

and formulating searches. Searches can be formulated as can the indexing by drawing from an independent distribution of terms. The output consists of the number of references to documents which are classified into three categories that correspond to the inputs. These categories of output are the number of desired recalled references, number of desired unrecalled references, and the number of undesired recalled references. In addition this output is related in greater detail to the various combinations of inputs, correctly and/or incorrectly used. This technique allows the source and magnitude of output error to be related to the input levels of indexing and search term usage.

The segment of total cost-total value is predicated on a pure competition model of total value of output benefits versus total cost of facilities, where total value is similar to total revenue in normal economic considerations. Optimization is achieved by determining the maximum profit level of usage of index and search terms. Total costs include all money expenditures for initiating and operating a reference retrieval system. Total value is obtained by assigning a constant value per retrieved needed reference and a penalty per unretrieved reference at a fixed cost per unit for all user queries, along with the cost to the user of preparing a request and evaluating the output. These cost and value functions are formulated by use of production functions to relate the various inputs to the levels of indexing

and searching or to the outputs. Cost data are then used in conjunction with the production functions to determine the total costs and total value for each input prescribed. Summation of these costs for the pertinent factors of production will yield the final total cost of facilities and total value of output benefits for the system. The most profitable level of operation is then determined in relation to the number of index and search terms which are the decision variables. Various levels and configurations of systems in use can be simulated with this model.

An example of the application of the model assuming data based on realistic estimates and of data present in the literature in addition to judgment factors are presented. The feasibility of application of the model is thereby demonstrated.

CHAPTER II

HISTORICAL BACKGROUND AND ANALYSIS OF INFORMATION RETRIEVAL SYSTEMS AND MODELS

The various aspects of reference retrieval systems have their origins in the growth of libraries and the subsequent development of information systems. In this chapter several aspects of library growth and the development of new concepts of information handling will be considered. This background will include estimates of growth in resources, needs, costs, value, and other factors relative to providing information to those individuals that have need.

Considerable work has been done on evaluating the performance of real and synthetic information systems in terms of actual output (or the lack of) and causes of lack of output as related to the inputs to the system. Also, a review of a model, with algorithms to calculate the number of references obtained from a retrieval system, is presented, along with its uses and limitations.

Performance Factors

The performance of various factors of document acquisitions, usage of index and search terms, along with evaluation of system's performance is presented.

Index Terms

Indexing is mapping the document space into the index

space. This problem has been analyzed in depth by Landry and Rush [6], who have considered many theoretical aspects and prescribed several models to describe indexing and its various phases. One of their conclusions is expressed in the axiom "Accurate retrieval depends on the exactness of indexing."

The frequency of usage of each index term is used in subsequent model work to express the functional form of the distribution of index terms. It is later shown that the usage of terms in indexing a document and the usage of terms to formulate a search are independent but have similar conceptual considerations. The significance of several aspects of index terms has been investigated by various authors. The individual relationships will be discussed here.

Frequency of Usage of Each Index Term. The first work in this area was by Zipf [7], who plotted the frequency of term usage versus the rank relationships on log-log paper for all of the terms used by five English writers. The fact that he obtained essentially a straight line indicates that the product of frequency and rank is constant.

Houston and Wall [8] determined that Zipf's procedure was not applicable to a vocabulary with a limited number of terms such as that used for indexing. However, T. E. Boyle, of the Du pont Engineering Department, communicated in 1956 to Houston and Wall the suggestion that term usage might be a predictor in a retrieval system. They plotted frequency

of use on a log scale against the percentage of the cumulative number of terms on the normal probability scale. The results caused them to propose a log-normal distribution for the frequency of usage of terms in indexing. They suggest that the number of terms be unlimited prior to the indexing of documents and that, as the documents are indexed, new terms may be incorporated into the index term vocabulary. Significant deviation of actual results from the theoretical model occurs for the most frequently used terms. Based on the empirical data the authors would limit the application of the model to 95 percent of the less frequently used terms. Their work was based on document collections from 303 to 195,000 items, in which depth of indexing ranged from 5 to 32 terms per document. The number of terms in the vocabulary ranged from 1108 to 7730.

Arthur D. Little, Inc. [9], proposed a geometric distribution of the following form:

$$g_1(j) = \frac{(1-B)\,B^{j-1}}{(1-B^q)} \qquad (2.1)$$

$j = 1,2,---,q$

$g_1(j)$ = probability of using the jth term in indexing a document

q = total number of index terms in the collection

j = rank of the given term.

This equation provides for a distribution of a finite number of terms which can be simplified to

$$f(j) = (1-B)\, B^{j-1}, \text{ for large } q.$$

Morse [10] proposes the use of this latter geometric distribution based on an infinite number of terms. Raver [11] also uses a geometric distribution in his work.

Long, Barnhard and Levy [12], who analyzed the works used in radiological (x-ray) records, required a specialized vocabulary in an effort to determine key words for indexing these records. They treated all words as

1. key words, relevant used = information
2. discard words, nonrelevant used = persistent noise
3. unclassified words
    a. potential key words
    b. potential discard words
    c. noise, infrequently used nonrelevant words.

Their work showed that, after analyzing 40,000 words of text, 2,500 key words had been introduced and that the rate of introduction of new words was diminishing. However, after they had analyzed 100,000 words of text, new key words were still being introduced.

Index-Term Vocabulary Size Versus Document-Collection Size. A. D. Little, Inc. [13], who plotted the number of documents indexed as opposed to the number of index terms for several indexing systems, concluded that the minimum vocabulary size for a large document collection could be expressed as shown in the following equation:

$$I = 18\sqrt{D} \qquad (2.2)$$

I = vocabulary size = number of index terms

D = number of documents in the collection.

As a lower numerical limit, 10,000 terms was proposed, an implication that large systems will require a large number of terms. Houston and Wall [14], in connection with their previously cited work, have concluded that

$$I = 3{,}330 \log (\pi_I + 10{,}000) - 12{,}600$$

for $10{,}000 < \pi_I < 1{,}000{,}000$

I = number of index terms

π_I = total number of term uses in indexing

$\bar{X}$ = average number of terms used to index a document

D = number of documents in the collection

$$\pi_I = \bar{X} D.$$

Therefore, $I \approx \sqrt{\bar{X} D}.$ (2.3)

This relationship suggests that a large vocabulary will result from the growth of two factors: (1) increase in number of documents indexed, and (2) increase in the number of terms per document. These results were obtained by an empirical study of the available data, and their mathematical implications have not been fully explored. The limits they obtained are based on 8 index terms per document items for a collection size of 10,000 documents to 70 index terms per document for a collection size of 1,000,000 documents. This technique also allows investigation of the rate of growth of

vocabularies. The rate of increase in the number of index terms per document is

$$\frac{dI}{dD} = \frac{d}{dD}[3{,}330 \log (\pi_I + 10{,}000) - 12{,}600]$$

$$= \frac{1440}{D+\frac{10^4}{\bar{X}}} .$$

Search Terms

The interrelationships of search formulation terms, document collection size, and vocabulary are needed to evolve quantitative measures for implementing the searching aspect of the subsequently developed reference retrieval model. It is particularly necessary to have an analysis of query intensity and vocabulary usage so that these factors can be interrelated and related to their counterparts in indexing.

Query Intensity and Time-Rate Usage of Documents. Query intensity measures the frequency of usage of a given area of knowledge. Therefore, analysis of the frequency of usage of documents provides some measure of the need for dissemination of information in a prescribed subject area.

Arthur D. Little, Inc. [15] plotted the library statistics of large colleges and universities for 1956-57 and concluded that a relationship of requests to the size of the collection was similar to the relationship of additions to collection size as shown.

$$R = dU^{e}$$

R = intensity of requests

d = constant

U = size of collection

e = exponent, less than 1 = slope.

Thus, the rate of queries would decrease with the increasing size of the library collection.

Leimkuhler's [16] work suggests that the circulation rate for materials at Purdue, recorded for 40 years, is parallel to acquisition and holdings (collection size). There are, however, considerable short-term fluctuations in all three factors.

Studies by Fussler and Simon [17] and Jain [18] suggest a decrease in the average circulation rates of items as their age increases. Jain's work, based on limited sampling, suggests an annual decline of 4.5 percent since publication and 6 percent since acquisition.

The work of Morse [19], which proposes a model using the Markov Process, indicates that the expected circulation rate has the following relationship:

$$R(m) = \alpha + J(m)$$

R(m) = projected circulation rate for the current year

m = circulation rate of the previous year.

α and J are parameters based on the various document classes or discipline.

The parameter α is considered to be essentially constant for approximately the first five years, then declines to one-half of its original value by the tenth year. Apparently the average usage of a given book decreases with age.

Vocabulary Usage in User Query Formulation. Inasmuch as user queries must be formulated to be retrieved, they require formulations that use terms in order to identify desired items and to retrieve them from the system. Arthur D. Little, Inc.[20] provides two alternative hypothesis on the distribution of index terms for retrieval:

1. the assumption that each term in the vocabulary has an equal probability of being chosen,

$$g_2(j) = \frac{1}{q}$$

where q = total number of index terms in the collection.

2. the assumption that the likelihood of an item to be chosen is a function of its probability of usage in indexing documents, from equation (2.1).

$$g_1(j) = \frac{(1-B)\,B^{j-1}}{(1-B^q)}.$$

This latter approach suggests that the document file is formed according to the needs of the users and that it provides material proportional to the intensity of interest in the various areas. It also infers that these documents are available, which is questionable.

Search Strategy and User Query Formulation. The type of search strategy possible and its effect on output of a

reference retrieval system are a consequence of the searching vocabulary and the method of formulation of searches. Therefore, the formulation of searches is investigated. As indicated by Arthur D. Little, Inc. [21], intersections of one to four terms were used to investigate the various search strategies.

By use of the equations previously proposed, the search strategy is used to formulate the expression for the expected number of citations of documents to be retrieved, which is

$$\bar{Z} = \sum_{j=1}^{q} g_1(j)\ \bar{X}\ D$$

which reduces to

$$\bar{Z} = (1-B)^2\ \bar{X}\ D$$

where $\bar{Z}$ = expected number of documents to be retrieved per completed user query.

However, applying this formula in the system being evaluated gave the number of documents (citations) far in excess of that noted in actual situations. Therefore, approaches for limiting the number of citations were then considered. The basic equation was also expanded to include varying numbers of terms in the search formulation and is discussed in greater detail later in this chapter.

Uhlmann [22] discusses some aspects of intersection of user request and file items, using Boolean logic and probability theory to devise a probabilistic search-strategy in a coordinate indexing system. This procedure investigates

and evaluates relations between secondary sets, produced by operations between them during the formulation of specifications, either for documents or for requests. The search operation seeks to ascertain which and how many members or subsets a request and a document specification have in common.

Raver's work [23], which includes retrospective searches, developed procedures for reducing search time based on the method of formulation of search strategy.

Evaluation of System's Performance

The evaluation of the system's performance includes factors that are endogeneous to the system. These factors are reflected in retrieval efficiency and a desired output in usable form.

Recall-Pertinence, Relevance. Perry, Kent, and Berry [24] devised a series of factors, including the recall factor and pertinency factor, which can be defined with the aid of Table 2.1. These factors were also basic to Cleverdon [25, 26,27] in his comprehensive analysis of four indexing systems in the Cranfield Project.

Lancaster [28] used the same procedure in his evaluation of MEDLARS (MEDical Literature Analysis and Retrieval System) but applied the term precision instead of relevance inasmuch as he was evaluating user need.

Montague [29] used recall and relevance while relating

and evaluates relations between secondary sets, produced by operations between them during the formulation of specifications, either for documents or for requests. The search operation seeks to ascertain which and how many members or subsets a request and a document specification have in common.

Raver's work [23], which includes retrospective searches, developed procedures for reducing search time based on the method of formulation of search strategy.

Evaluation of System's Performance

The evaluation of the system's performance includes factors that are endogeneous to the system. These factors are reflected in retrieval efficiency and a desired output in usable form.

Recall-Pertinence, Relevance. Perry, Kent, and Berry [24] devised a series of factors, including the recall factor and pertinency factor, which can be defined with the aid of Table 2.1. These factors were also basic to Cleverdon [25, 26,27] in his comprehensive analysis of four indexing systems in the Cranfield Project.

Lancaster [28] used the same procedure in his evaluation of MEDLARS (MEDical Literature Analysis and Retrieval System) but applied the term precision instead of relevance inasmuch as he was evaluating user need.

Montague [29] used recall and relevance while relating

Table 2.1. Recall - Relevance

	P (Relevant) (Pertinent)	$\overline{P}$ (Not Relevant) (Not Pertinent)	Total
R (Retrieved)	a	b	a+b
$\overline{R}$ (Not Retrieved)	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$\text{Recall} = \frac{a}{a+c}$$

$$\text{Relevance (Pertinence)} = \frac{a}{a+b}$$

expense versus depth of indexing.

At this point it seems appropriate to distinquish between system effectiveness and user effectiveness, which is well defined by Rees[1], who says

> The difference between relevancy and pertinency is that relevancy is a property which corresponds to a question, while pertinency is a property which corresponds to a need. Relevancy is associated with the relationship between a document and a question,

[1]Alan M. Rees, "Semantic Factors Role Indicators et Alia: Eight Years of Information Retrieval at Western Reserve University," Aslib Proceedings, Vol. 15, No. 12 (December, 1963), p. 358.

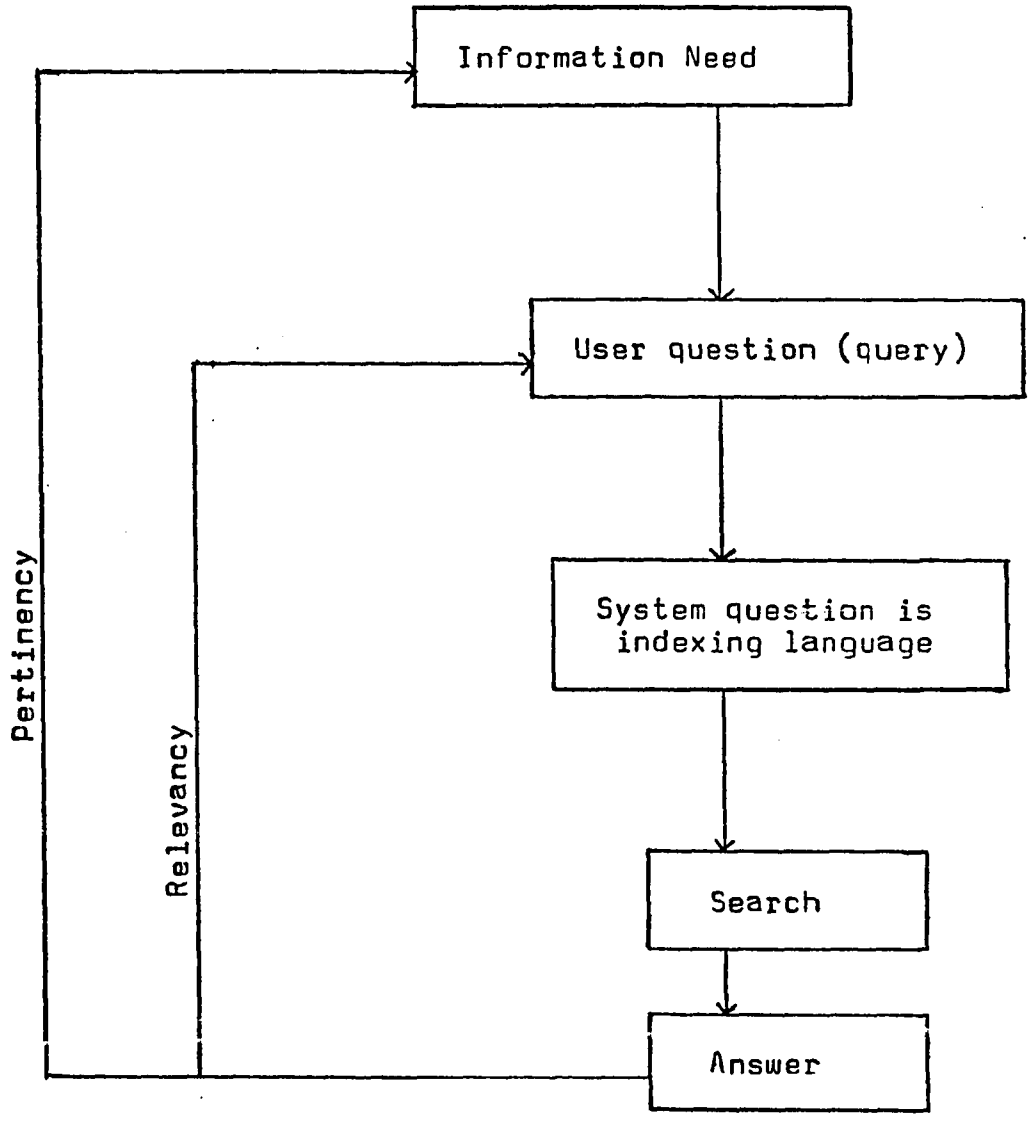


Fig. 2.1.--Pertinency-Relevance
(from Rees)

whereas pertinency is associated with the information need of which the question is a formal representation. The degree to which relevancy and pertinency coincide can be considered as a measure of the questioner's ability to represent his need in terms of a formal statement, [See figure (2.1)] the skill of the question analyst and the effectiveness of the indexing language.

On this basis it can be assumed that it is the responsibility of the system to retrieve the documents or the references to the documents or items responding to the user's articulated query or those which are the subject of his query.

Expansion of Recall-Relevance Ratios. Recall-relevance ratios can also be expressed in a slightly different manner. Therefore, a corollary to Table 2.1 is presented in Table 2.2.

Table 2.2 Recall-Relevance (Pertinence) Weight Factors

	Relevant (Pertinent)	Not Relevant (Not Pertinent)
Retrieved	K	M
Not Retrieved	L	J

K = value of retrieving a relevant document

J = value of not retrieving a non-relevant one

M = cost of retrieving a nonrelevant document

L = cost of failing to retrieve a relevant one.

Combining the two tables in the manner proposed by Verhoeff, Goffman, and Belzer [30] produces the measure of efficiency E.

$$E = aK - bM - cL + dJ.$$

However, at the time this measure was proposed, there did not appear to be any quantitative information to test the equation.

Swets [31] proposed the measure obtained by plotting the values obtained from Table 2.1.

$$\frac{a}{a+c} \text{ vs. } \frac{b}{b+d}$$

where

$$\frac{a}{a+c} = \text{recall}$$

and

$$\frac{b}{b+d} = \frac{\text{retrieved non-relevant}}{\text{(retrieved and not retrieved non-relevant)}} .$$

One advantage of this measure is that it incorporates all four fields of relevance and recall. In subsequent work Swets [32] did present examples using numerical data. However, the number of sample values needed for each of several data points has precluded widespread adoption.

This concept is expanded by Salton [33], who in addition uses the terms "fallout" and "generality." The expression used above is designated as follows:

$$\frac{b}{b+d} = \text{fallout},$$

the proportion of nonrelevant documents actually retrieved while searching the document collection. The proportion of relevant documents available for retrieval as a function of the total number of documents is

$$\frac{a+c}{a+b+c+d} = \text{generality}.$$

Bourne, and others [34], propose a system that includes the standard recall and relevance ratios. Also included are a series of user parameters. Their data show that various users will tend to disagree as to the relative importance of the various parameters.

Limitations on the use of recall versus relevance ratios as related to users is demonstrated by Pollock [35], who argues that sometimes one document is adequate whereas, at other times, many documents are required. Therefore, Pollock has developed a model to determine the expected number of documents needed to satisfy the user's query.

It is reasonable to assume that the number of relevant documents needed is proportional to the levels identified by Luhn [36]. This aspect is discussed by Wall [37], who also suggests that there is merit in ranking the various documents as to their usefulness or relevance.

Performance of Information Retrieval Systems

The primary tests for performance of indexing systems

are based on the results obtained by Cleverdon, [38,39,40]. The most comprehensive test of an operating reference retrieval system is that of Lancaster's [41] evaluation of MEDLARS. Saracevic, and others [42], evaluated a series of factors in their work at Comparative Systems Laboratory.

Aslib Cranfield

The work of Cleverdon and his associates in the Aslib Cranfield Research Project, involved determining the efficiency of four descriptor languages in a test situation based on a real industry. All of the indexing and the questions to interrogate the system were formulated for the purposes of the experiment. The descriptor languages were real, and the conditions of the experiment were well controlled.

The general subject was aeronautics, with half of the documents being articles in scientific and technical journals and the other half being research reports. The basic measure of efficiency was recall-relevance ratios. Tests were also conducted to measure efficiency based on the time allowed for indexing. By use of a fixed number of documents in the sample for each descriptor language, these documents were then indexed by individuals who were allotted a fixed time interval to accomplish their work. The experiments were repeated at time intervals of 2,4,8,12, and 16 minutes per document. The results of this experiment were not evaluated as to the number nor the applicability of each term

applied by the indexers. However, the effect of the time allowed for indexing was evaluated based on retrieval. The retrieval efficiency was based on the number of documents retrieved in response to a fixed number of questions for each test group (consisting of all the descriptor languages at each level of time allocated for indexing).

Some descriptor languages show a decrease in retrieval efficiency as a function of time available for indexing; others show an increase in efficiency. In addition, the general level of retrieval efficiency varies between the various descriptor languages. Analysis of the successful and unsuccessful searches shows a general increase in the number of terms applied to a document as the amount of time allocated for indexing increases. The variation in the number of terms per document is greater between the various descriptor languages than it is between the number of terms per document for the successful and the unsuccessful searches.

In addition, the number of not needed documents that would be recovered as the number of terms applied in indexing increases is not known. Therefore, it can be concluded that, as time allowed for indexing increases (within some limit), the number of terms applied to a document will increase. The applicability of these terms to describe the contents of the document and their effect on retrieval cannot be ascertained by this experiment. Furthermore this evaluation used synthetic questions to evaluate the system. Overall data

showed that all four systems operated at a recall of between 60 and 90 percent, with an average of 80 percent. All groups ranged from 74 to 82 percent for the individual recall ratios. Of the failures to retrieve, 60 percent were caused by indexing failures, 34 percent by question-and-search failure, and 6 percent by system failure (indexing system).

Aitchison and Cleverdon's [43] evaluation of Western Reserve indexing showed 30 percent of the failures to be in the indexing language, an indication that, in both the Cranfield and Western Reserve University evaluation, the major source of errors was either the indexer or the searcher, that is to say, a human error.

MEDLARS

Lancaster [44], in his work with the National Library of Medicine, evaluated the performance of MEDLARS, which had been in operation four years. MEDLARS is a multipurpose output system of the National Library of Medicine. In this paper there is concern only with the "demand search" aspect (i.e. requests formulated in response to a qualified user's request [demand for information]). The indexing vocabulary MeSH (Medical Subject Heading), consisted of approximately 7000 pre-coordinate subject headings in thirteen subject areas. A hierarchical classification was available, and subheadings were introduced during the operation of the system. Approximately 200,000 documents were indexed annually, at an

average of 6.7 terms per document. Tapes of the input are available monthly to a group of cooperating medical centers located over the United States and some foreign countries. The searches are formulated by use of an intersection of terms (A and B), or a union of terms (A or C), or a combination (A and B or A and C). Output is in the form of a computer printed bibliography. Therefore, at the time of Lancaster's evaluation there was (1) an existing reference retrieval system with a specified vocabulary, (2) an inventory of references to documents which were recallable, and (3) a working group of users who were applying the output of the system to their particular needs. These users could be considered as sources of measurement of the desired characteristics of the system. While the total objectives of the evaluation were broader than those given here, two of the test requirements were to measure (1) its recall power (i.e. its ability to retrieve "relevant" documents, --- documents of value in relation to an information need that prompted a request to MEDLARS) and (2) its precision power (i.e. its ability to hold back 'non-relevant' documents).

The efficiency was measured over a twelve-month interval of operation of the system in conjunction with a select group of users who were unaware of the existence of an evaluation program until they submitted their requests; in this manner "real" requests were assured. The cooperating users were presented their normal output and an auxiliary output with a

limited number of references to documents. This auxiliary output consisted of the "precision set" and the "recall set," with no restrictions as to being mutually exclusive, or partially or wholly contained within another. Practically, there was usually some degree of intersection of both sets.

The "precision set" was a subset of the output listing, chosen by random number limited to 25-30 references, normally presented to the user. The user was presented with the documents corresponding to the references in the sample. He was asked to evaluate these documents and state whether they were of value to him in his specific request and thereby relevant. The ratio of the number of documents, judged by the user to be of value to him, divided by the number of documents in his precision set produced the precision ratio and was assumed applicable to the entire output.

The "recall set" consisted of a listing of references to documents obtained from other sources such as (1) those known by the user at the time of submitting the request, (2) local librarian, and (3) other sources. The references in this set, which were judged relevant by the user, formed the denominator of the recall ratio, and the number of these documents listed in MEDLARS formed the numerator. The recall ratio was assumed to be applicable to the entire output of that user. This procedure was adequate because, theoretically, it would have been necessary to review the entire listing of documents to ascertain whether they should have

been indexed in response to a given request.

Much effort was used to ascertain the causes of failures of formulated searches, both the omission of references to desired documents and the inclusion of references to undesired documents. Effort was made to distinguish between the types of error; those of indexing versus those of searching, and they were treated independently. However, in each of the types of error, one specific reason for failure was generally assigned. The various sources of error and their frequency and type of error, as related to the number of formulated searches, were presented. However, since the compiled output data did not distinguish between the various input formats of intersection, union or combination of both, it is not possible to relate the output of the number of references to documents to the number of search terms in a quantified manner. Similarly, it is not possible to relate output errors to the number of terms used in indexing except in a qualitative manner. The procedure used in evaluating the output involved, having two subsets of the output, the "recall set" and the "precision set," had unequal sized samples. Therefore, any statistical calculations about the output will, of necessity, have different confidence limits for the same population.

Comparative Systems Laboratory

Saracevic, et al [45], have done considerable work on determination of the source of error in documentation systems

at the Comparative Systems Laboratory. Their testing procedure was based on the data presented in Table 2.3.

Table 2.3. System Components - Constant and Varied

Function Components	How Treated
1. Acquisition - policy	Constant
2. Source of input, (i.e. degree of completeness)	Varied
3. Indexing language - vocabulary set of index terms with a set of rules	Varied
4. Coding - symbolic representation of index terms	Varied
5. File organization - order of file contents	Constant
6. Question analysis - formulating query concepts into indexing language	Varied
7. Search strategy - procedure to search the file	Varied
8. Format of output - physical form and degree of representation of document presented to user	Varied
Purpose Components	**How Treated**
1. Class of user	Constant
2. Discipline	Constant
3. Size of file	Constant

As can be seen from this table, this is a significant number of components to evaluate. The function components

were varied to determine the effect on the purpose components which were held constant. Their work does provide a good basis for ascertaining the effect on other aspects of the system of varying parameters, and their conclusion is that the human factor has the highest variability in most components of a retrieval system.

Model to Calculate the Number of References to Documents to be Retrieved From a Retrieval System

The model for this calculation is described in A. D. Little's [46] work. In such a system a set of terms is applied to each document. A search is made by specifying a set of terms, and only the documents listed under all of the terms in the set are obtained as formulated search output. This is a Boolean algebra approach, in that, the search is formulated by using the intersection of terms. This model was conceived as appropriate for determining the expected number of items to be retrieved, given the following:

1. the number of documents in the collection,
2. the number of terms in the indexing vocabulary,
3. number of terms used in indexing a document,
4. number of terms per formulated search.

Based on analysis, plus work on previous systems, it was demonstrated that the geometric distribution of the frequency of the use of each term in the vocabulary in indexing was applicable. This approach ranks terms in decreasing

probability of their application of use in indexing documents, which is shown graphically in Figure 2.2, and is expressed in equation (2.1)

$g(j)$ = probability of usage of the jth term in indexing

$$g(j) = \frac{(1-B)\,B^{j-1}}{(1-B^{q})}.$$

The actual number of documents to be indexed under the jth term having a rank of j can be expressed as follows:

$$V_j = \frac{(1-B)\,B^{j-1}\,\bar{X}\,D}{(1-B^{q})}. \qquad (2.4)$$

It is assumed that the probability of usage of any terms in searching is directly proportional to its probability of usage in indexing. Therefore, the expected number of references to documents in a search formulated by using one term is

$$\bar{Z}_1 = \frac{(1-B)^2(1-B^{2q})}{(1-B^2)(1-B^{q})^2}\,\bar{X}\,D$$

assuming that q is large

$q \longrightarrow \infty$, $B^{q} \longrightarrow 0$.

Therefore, this equation can be simplified as follows:

$$\bar{Z} = \frac{(1-B)^2\,\bar{X}\,D}{(1-B^2)}.$$

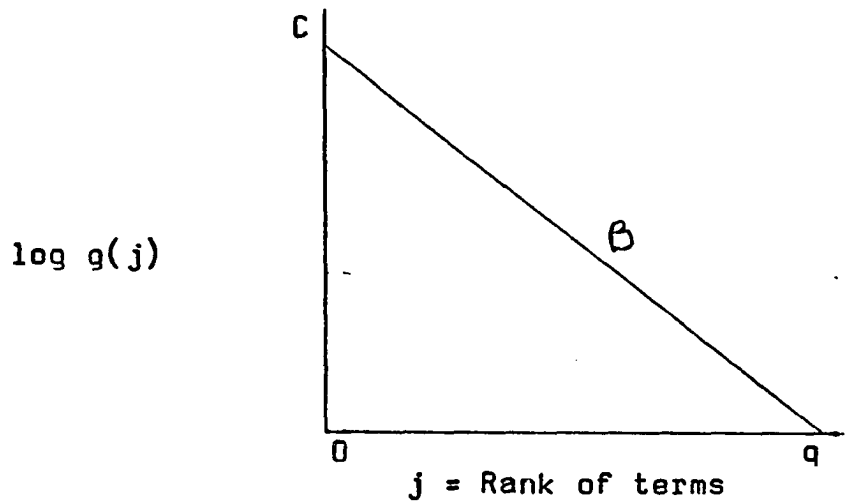


Fig. 2.2--Frequency Distribution of Term Usage in Vocabulary

For a search formulated by using two independent terms, the probability distribution equation (22) from page IIA-18 is

$$h(j_1, j_2) = \frac{(1-B)(1-B^2)\ B^{j_1+j_2}}{B^3},$$

$$j_1 \neq j_2 = 1,2,---,q.$$

This formula is correct if the expression for the number of terms for use in indexing is defined as

$$j_1 = 1,2,---,q$$

$$j_2 = j_1+1,\ j_1+2,---,q.$$

Therefore, the first form of equation (27), page IIA-21 is correct,

$$\bar{Z}_2 = \sum_{j_1=1}^{\infty} \sum_{j_2=j_1+1}^{\infty} (1-B)(1-B^2)\ B^{j_1+j_2}\ F_2 \frac{(1-B)^2}{B^2} \cdot B^{j_1+j_2}\ \bar{X}^2 D$$

in addition a correlation factor, F_2, has been added.

The final form of equation (27), given as

$$\bar{Z}_2 = \frac{(1-B)^2\,\bar{X}^2 D\,F_2}{(1+B)(1+B^2)}$$ omits a factor of B in the numerator and should read:

$$\bar{Z}_2 = \frac{B(1-B)^2\bar{X}^2 D\,F_2}{(1+B)(1+B^2)}\ .$$

This develops into the general form of the equation (32), page IIA-23,

$$\bar{Z}_q = \frac{(1-B)^q\bar{X}^q D\,F_q}{B^q} \prod_{j=1}^{q} \frac{B^j}{1+B^j}\ .$$

The correlation factor, F_q, can be expressed as θ^{n-1}, where θ is approximately 3.

So the final equation as expressed in No. 33, page IIA-26, is

$$\bar{Z}_q = \left[\frac{(1-B)\,\bar{X}\theta}{B}\right]^q \frac{D}{\theta} \prod_{j=1}^{q} \frac{B^j}{1+B^j}\ .$$

This is correct if $j_1 = 1,2,---,q_1$

$j_2 = j_1+1,---,q_1$

$\vdots$

$j_q = j_q-1+1 = j_q$.

Uses

The model, as outlined, can evaluate and determine the number of documents to be retrieved at various levels of usage of terms in both indexing and searching. As the number of terms of either indexing or searching separately is

varied, the output will express this change. The model is based on the assumption that the ranks of the index terms are equal to the ranks of the search attempts.

Limitations

The assumption ofequality of ranks is a limitation and other limitations are discussed below.

1. The system, as constructed, assumes that the rank of the search terms is identical to that of the index terms. This assumption implies that the available data correspond to need. Carrying this concept one step further implies that present needs of users were anticipated by previous document contributors, and these documents were processed in accordance with future need. Therefore, any new area of knowledge development is precluded, because all future needs are foreseen, and there will not be any change in demand.

2. Intersection of search sets has been formulated, but the model itself does not provide for a union type of search.

3. The value of the equation is in terms of the expected value. That is to say, it is a single number without any variance or range.

4. If the ranks of the search terms are unequal to that of the index terms, the application of the equation tends to become rather cumbersome. If the rank of the index terms and the rank of the search terms are dissimilar, these differences must be related.

5. The procedure prescribed by the model assumes that all the terms used and only those terms are applied in both indexing and searching (independently). Therefore, there is no recognition of the existence of errors in either indexing or searching or both, and their consequences in the output of retrieved citations can not be quantitatively evaluated. A qualitative approach, however, is described.

Total Cost-Value

The concepts of cost and value of a system or product must be considered if optimum allocation of resources to benefits are to be obtained.

Cost Models

The significance of costs in any real system must be considered. Therefore, they are investigated from several concepts. For the purpose of this report, they include (1) monetary expenditures necessary to operate a system including the installation expense, and (2) the cost to the user of preparing his request and analyzing the output of the list of references obtained in response to his query.

Considerable analysis of cost data for information and documentation retrieval systems has been expressed by Landau [47] in his article "The Cost Analysis of Document Surrogations of Literature Review." The essence of this report is that very little cost information is available in the forms of (1) structures for using the cost information, (2) procedures for recording and obtaining cost information, or (3) numerical values to express the costs.

Lancaster [48] develops a conceptual procedure for trade-offs between input and output costs, and other aspects of surrogation. He discusses (1) cost effectiveness in terms of how effective a system is in satisfying its objective, and (2) cost benefits, which relate to the justifi-

cation of the existence of a system. He also discusses two kinds of variables costs: (1) those that are a function of the number of transactions, and (2) those that are a function of the manner of conducting operations, both of which are subsequently incorporated in the model developed later in this report. Lancaster's report is quite comprehensive in covering the various aspects of overviewing an existing or contemplated retrieval system; but it does not present any specific functional form of relating the various phases of a surrogation system so that a given situation can be quantified.

Keith [49] presents a general model for evaluating information storage and retrieval systems. His model is expressed in functional form as follows:

$$E(C_t) = E(C_i) + E(C_m) + E(_{in}) + E(C_{op}) + E(C_{out}).$$

Where, C_t = Total costs

C_i = System initialization cost

C_m = Maintenance cost

C_{in} = Input cost

C_{op} = Operation cost

C_{out} = Output cost.

This primary group is divided into phases as follows:

$$E(C_i) = E(C_{imp}) + E(C_{st}) + E(C_{eq}) + E(C_{con}).$$

Where, C_{imp} = Software acquisition cost

C_{st} = Staff training cost

C_{eq} = Equipment acquisition cost

C_{con} = File record processing cost.

$$E(C_m) = E(C_{upd}) + E(C_{edt}).$$

Where, C_{upd} = File updating cost

C_{edt} = File editing cost.

$$E(C_{in}) = E(C_{qpr}) + E(C_{pro}).$$

Where, C_{qpr} = Query preparation cost

C_{pro} = Query processing cost.

$$E(C_{op}) = E(C_{opr}) + E(C_{dly}).$$

Where, C_{opr} = Operating time cost

C_{dly} = Delay cost.

$$E(C_{out}) = E(C_{for}) + E(C_{list}).$$

Where, C_{for} = Format cost

C_{list} = Listing cost.

Bloch and Ofer [50], working on a selective dissemination information system, have presented a procedure for relating data through stages of preparation in a functional form as a step toward obtaining a total cost relationship. One of their applications is ascertaining the value of computer time and allocating it to various functions as follows:

$$T = (F_1 \cdot a + F_2 \cdot b + F_3 \cdot c + F_4)\ N.$$

Where, T = Time for processing N cards

N = Number of cards processed

F_1 = Time parameters

a = Number of n-tuples

b = Number of sentences

c = Number of microprofiles.

Therefore, an application of production functions is presented, but the concept is not expressed.

Bourne, et al [51] did considerable work in evaluating real and potential users of information retrieval systems. Much effort was expanded in categorizing the users and obtaining their opinions on a range of questions concerning efficiency, timing, effort, value, and other factors. The factors held constant by Bourne and Ford are

1. size of the file items (number of pages or characters used),
2. initial file size,
3. amortization period for equipment purchased,
4. rate of return for amortization calculations,
5. burden and overhead percentage.

While the list is adequate for the purpose used by the authors, the data were organized on a unit of output basis versus cost. The units of input and cost per units of input are included simultaneously. Therefore, determination of variation of use of input or changes in cost per unit of input cannot be ascertained separately. Application of three of the factors, initial file size, amortization period for equipment purchased, and rate of return are included in the model, which is subsequently developed. Also, it is assumed

that the size of the file, item 1, is held constant.

Bourne and Ford [52] have presented an annual cash flow and an equivalent annual cost procedure for structuring the accumulated cost data as it relates to the inputs versus the outputs and is designed to handle a series of types of monetary expenditures. Considerable thought and effort was devoted to evolving a means of relating time versus other constraints of an information system in terms of cost to the user. These are as follows:

1. The necessary time to prepare the input requests,
2. The time delay necessary to prepare and provide the output,
3. The time to analyze the output,
4. The time to reformulate the search and go through steps (1), (2), and (3), if the first search is not successful or, if necessary, to obtain the required information from other sources.

The authors did investigate user needs, costs, and a procedure for expressing the monetary expenditures based on time. However, there is no procedure for determining the monetary expenditures needed to generate a system and when they will be incurred. Neither are these expenditures and the assso-ciated user costs related, nor is there a quantified procedure for relating the values of the system directly to costs. Also, there is no procedure for relating indexing and searching term characteristics as they affect output performance and cost.

Marron and Snyderman [53,54], working on a reference retrieval system using computer storage and retrieval, have evolved a time-effort distribution analysis of the form

$$T = M + B + S,$$

Where T = Computer usage, hours

M = Maintenance time, hours

B = Batched usage jobs, hours

S = Simgly run jobs, hours.

They determined computer usage cost by making a linear distribution of costs with computer time. Maintenance time was distributed linearly between batched and singly run jobs based on the time involved for each phase which implies that cost is a function of computer time. In addition, allocated costs are also linearly related to the basic function.

Kuney [55], evaluating the feasibility and economics of computer typesetting for scientific publications, presents data relating rate of keystrokes and computer processing speed in cost per unit of each. These functions have the appearance of a rectangular hyperbola. Their general form is applicable to reference retrieval file data processing, but the specific values have limited application because they are based on a continuous high volume-input, not intermittant input as in the proposed model.

Stanwood [56] discusses the cost and time involved in operating a computerized information retrieval system that provides selective dissemination of information to a series

of users. Statistical data pertaining to a number of operations on a percent of time basis is presented. While Stanwood's analysis ascertains the time necessary for the number of procedures involved, it is not presented in a manner that is readily related to the volume throughput of the entire system or any of its phases. In addition, it is related to selective dissemination of information not retrospective or demand searches.

The work of Aslib [57,58,59] indicates that the incremental number of terms used to index a document decreases as a function of time.

Barish [60] presents a technique showing the development and application of equivalent annual cost, which is used to relate initial installation and equipment expenditures to the annual costs and benefits of operating a reference retrieval system.

Cost Data

Montague [61], having done work relating costs, relevance, and recall for a patent reference system, also presents some numerical values for the three systems under study. She also goes into the cost versus depth. This implies that after some particular level is reached the cost of indexing will increase without limit and without any appreciable increase in depth of indexing. Input costs per document varied from $4.00 to $15.00, while the search costs

varied from $32.00 to $102.00 per user question.

Overmyer [62], at the Center for Documentation and Communications Research, in connection with the American Society for Metals Documentation Services, has done considerable work in costing various types of outputs of their system in addition to covering installation costs. A constant cost per unit of input is presented. Extensive cost of various ranges of output and input factors is presented. For the three systems presented, the total search costs were $105.48 to $150.48 per response to a particular user query.

Costello [63] shows a summary from five data retrieval installations. A description of attributes of each installation, based on a large number of variables, is presented. A review of the presented costs of actually conducting a search shows a variation from $5, $19, $45, to $53. However, the $19 value does not include overhead, and there is no statement as to whether the $45 and $53 value includes overhead or not. The $5.00 value is the cost based on 15 minutes of user time and, therefore, does not include any operating cost.

R. R. Johnson [64], in evaluating computers, discusses the difficulty of relating the effectiveness of the computer by relating time effectiveness to input/output functions, file storage size, and calculating capability.

Helmkamp [65] presents the accounting profession approach of using a series of cost centers to collect, record,

and accumulate the costs for a reference retrieval system. It is his hypothesis that "a theoretically sound managerial cost-accounting system can be designed to meet the specific characteristics of a technical information center by revising and innovating systems utilized by other enterprises." Helmkamp did evolve a basic cost-accounting procedure designed to (1) identify the various physical operations of the system, (2) collect the various categories of appropriate costs at each of these centers, and (3) accumulate costs for each center and then aggregating these costs for the entire operation. His analysis showed that 70 percent of the cost for a reference center were fixed, largely for salaries. Among the various cost and time allocations presented, the cost of an average retrospective search recorded was $86.06.

Penner's [66] review of the literature regarding costs and charges for library information services shows that meager data exist concerning costs. He structures existing cost data into 22 items and concludes that costs are greater than charges made for information services obtained. Application of various cost values cited in his article will subsequently be used in the application chapter, Table 7.1, to be used in the total cost-value model.

Costing of specific retrieval services is available from Rogers [67,68,69,70] concerning the MEDLARS retrieval center at the University of Colorado Medical Center.

Cummings [71] presents expenditures for several aspects

of MEDLARS.

Communication from Caldwell [72], of the National Library of Medicine, presents a statement of various aspects of performance and operating costs for MEDLARS for the fiscal year of 1967.

Niland [73] establishes a rather comprehensive historical recording of a series of costs for library expenditures.

Value

Value is a rather intangible factor to define as expressed by Mueller [74], who lists the following difficulties:

1. The lack of an established market for information in the usual sense of the word,
2. The lack of a standard unit of information on which to put a price tag,
3. Information is difficult to express tangibly.

The value of the output of the system must be considered in context to the user, which implies that it must relate to his needs (assuming he has properly articulated his needs in his query to the system). Also, the output must be presented in a usable manner.

I. J. Good [75] proposes a decision-theory approach using a utility value concept. The expected value, EV (number of documents retrieved relevant, number relevant not retrieved, number retrieved not relevant) = EV (a,b,c) on

Table 2.1. It is his argument that, as the total number of relevant documents increases, the total value of the documents to the user increases, and the marginal contribution of each relevant document decreases. He suggests this may be proportional to the number of relevant retrieved documents. The loss of value in going through the nonrelevant documents retrieved is considered proportional to the number of such documents. Therefore, the value of the system can be estimated from a sample of requests as the average value,

$$\Sigma \left(\sqrt{a + c} - \sqrt{c} - \lambda b\right),$$

where λ is some positive value, and where the summation is over all the members of the sample. This principle is then extended from the concept of relevance to categories of relevance such as high relevance, low relevance, and irrelevance. The problem of application necessitates equating low relevance to high relevance in absolute units of measure.

Relevant retrieved citations can be considered as having positive values while penalties or negative values are assigned to (1) unretrieved relevant citations and to (2) retrieved nonrelevant citations. This concept is later applied in the total cost-value model developed in this paper by (1) applying a cost to evaluate all retrieved citations and (2) placing a negative value on the unretrieved relevant citations.

Other structuring of cost data is presented in an

article by Emery [76], who proposes a Bayesian approach to structure the value of information as it is applied to decision making. Mr. Emery reviews the basis of value of information and stresses the time factor relationship of value as determined by the effects of the decision with which it is concerned. Therefore, from a management viewpoint, the precision, the completeness, and the time factor of data are significant in making management decisions. However, this particular type of analysis is not largely concerned with reference retrieval systems per se.

Gotterer [77] presents the concept of using supply and demand as a means of expressing or of determing the cost and value of information. He recognizes that there are some limitations for application of this type of situation. However, he presents a procedure for the gathering of data so that a model of this kind could be implemented or at least obtain data so that the performance of an installation can be measured.

Bryant [78], in his work on document handling, goes into considerable detail in discussing procedures for determining the cost of documents. This cost was related to the effect of time required to obtain documents from the home library versus other potential sources. The procedure could be applied to ascertain the cost of all the information

needed, if a complete retrieval of all desired references is assumed.

Churchman, et al [79], suggest a procedure of weighting objectives to get the implied costs of intangible factors if total costs have the following form:

$$TC = C_1 x + [\text{other pertinent costs (known)} = f(x)]$$

where TC = total costs

C_1 = the intangible factor (value unknown)

x = the number of items of an intangible factor.

The value of the intangible cost factor, C_1, is determined by setting $d(TC)/dx = 0$, using a value of x obtained by assuming that the organization has been using an optimum policy, and solving for C_1.

Mueller [80] presents numerical values of information obtained by using three different procedures at one installation. They are

1. Report Cost

   Reports cost an average of $1200 and are used an average of 10 times. Therefore, average cost per use is $120.

2. Alternative Cost

   Saving on the cost of two people consulting on a problem, each contributing an hour of time, at $10 per hour for each person, total $20.

3. Time Saving

   Each retrieval saves 1.2 man-hours of engineering time and results in a saving of $12.

Several techniques for expressing the value of output and procedures of limited application that interrelate value and cost have been presented. An organized approach is needed to provide (1) a structure to relate benefits to the critical inputs, (2) a structure for gathering the appropriate costs as related to the critical inputs, (3) categorization of the various costs within the structure, and finally (4) an application of the first three items to evaluate a system. None of the work investigated includes all of the first three phases.

Optimization

A system can be optimized if the appropriate cost-value relationships are known in terms of the parameters that control the output levels. Subsequently, a model to describe the level of and the type of various kinds of output will be evolved. However, a procedure for optimizing this output is described in Box and Hunter's [81] article on an Evolutionary Operations procedure, which uses an analysis of variance procedure to optimize (either to maximize or minimize) the entire process in terms of the variables being studied. A knowledge of the system is inherently necessary to make the most effective use of the procedure. Evaluation of the more critical variables controlling the optimum level of output will be performed along with an analysis of the effect of variation in inputs to the optimum solution.

Carlisle [82], in his determination of the value of a

mineral deposit distinguishes between the maximization of annual profit and maximization of total profit from a given mineral deposit. Maximum annual profit is a function of the rate of extraction given reserves, and can be expressed

$$\pi_A = f(R/Q)$$

where π_A = annual profit

R = rate of extraction of reserves

Q = quantity of reserves in the deposit.

Maximum total profit from a deposit is a function of the amount of reserves given an extraction rate as shown,

$$\pi_T = f_1(Q/R)$$

where π_T = total profit from a given mineral deposit.

A direct solution for each of these models is provided. However, since this procedure does not necessarily prescribe a method of maximizing total profit from a given deposit as a function of two independent variables R and Q, the author provides a general solution of the form

$$\pi_T = f_2(R,Q).$$

This general form is expressed quantitatively in the model being developed.

The conditions for optimizing a function having two independent variables have been described by R. G. D. Allen[2], who says

---

[2]R.G.D. Allen. Mathematical Analysis For Economists. (New York: The MacMillan Company, 1968), p. 497.

In the case of two independent variables, a point where

$$f_x = f_y = 0 \tag{2.5}$$

gives a maximum value of

$$Z = f(X,Y)$$

if

$$d^2Z = f_{xx}dX^2 + f_{xy}dXdY + f_{xy}dXdY + f_{yy}dY^2$$

is negative definite, i.e., if

$$f_{xx} < 0$$

$$\begin{vmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{vmatrix} > 0.$$

It would appear that the conventional mathematical economic analysis of relating total cost to total value, in terms of profit, would be another procedure that could be investigated. However, analysis of existing work has not indicated an application of this concept in reference retrieval systems.

Conclusions Based on Review

Analysis of the work on performance of systems (both document and reference systems) shows that there are (1) examples and procedures for ascertaining causes of errors in the output of information systems, (2) algorithms for calculating the expected number of references to documents that would be recovered from a given system, and (3) models for expressing monetary expenditures.

There is no quantitative procedure for determining the

effect on the output related to the number of terms used in indexing and searching if either or both of these inputs vary from the idealistic assumption of a fixed number of terms for each, which are all used correctly.

It must be recognized that some of this theoretical work was based on document searches. In addition, Cleverdon's [83,84,85] work used artificial (manufactured) questions that were subjectively designed to be answered by the contents of some given document(s) and were the basis for the recall, or rather the lack of recall. The basis of the relevance measure was judged by retrieval of documents whose text did not answer the stated questions, that is, those that were not relevant. Lancaster's [86] evaluation of MEDLARS used the recorded case histories. But even there, the somewhat subjective opinions of the users, of necessity, had to be used. A reference retrieval system, however, presents only citations to documents, and additional steps must be performed to obtain the documents, investigate them, and verify the recall and relevance of the various citations provided for in order to fulfill a given user request or query.

Therefore, the cost considerations are of concern, and there is no procedure for relating the monetary expenditures needed to install and operate a reference retrieval system along with the costs and value to the users in a manner that includes the performance characteristics of recall and

relevance, based on some factors that can be controlled in a given reference retrieval system.

Other procedures for structuring costs can be obtained by using conventional economic models. Two types of extreme condition models are available, both for structuring inputs and for outputs separately. Input usage can have the pure competition on the pricing of inputs versus the consideration of monopsony. The output or the sale price of a product can be modeled by using a pure competition approach which implies a constant price per unit of product versus the other extreme, that of the monopoly situation. The existence of these various models implies the use of the production function to relate the inputs to the outputs, then costing the inputs in conjunction with the production function to relate them to the cost of the outputs by formulating the total cost function. The production function is a means of expressing the physical relationship between input quantities and composition employed in the production process and the output quantity yielded by the process.

Very limited work has been done in the area of describing functional relationships between the various types of work involved in operating a reference retrieval system and the quantity of output, which in the proposed model is retrieved references to documents. However, these production functions for the various phases of the surrogation system can be developed from known or accepted relationships.

CHAPTER III

PROBLEM FORMULATION AND MODEL DEVELOPMENT

Presently, a large number of reference retrieval systems are in operation. Reference retrieval systems have been defined as systems that retrieve citations to documents in response to a user's articulated query. These systems process information and data usually for a specific discipline or defined area of knowledge.

The development of relationships between the usage of terms in inputs, their errors and their relationship to total vocabulary usage are discussed. These interrelationships are formulated into a general model to relate level of inputs to the level and quality of output.

Analysis of Reference Retrieval Systems

The work done by some individuals has differentiated the errors in outputs between those caused by the inputs of indexing and those by searching. Indexing is the assignment of appropriate terms to describe the intellectual contents of a document. Searching is defined as the assignment of appropriate terms to define a user's articulated query.

In addition to the proper usage of terms to describe a document while indexing, two types of errors can be defined: an omission error, which is the lack of assignment of enough terms to properly describe the contents of a document, and

a commission error, which is the assignment of improper terms in an attempt to describe the contents of a document. As in indexing, searching can also have omission and commission errors. Therefore, there are two basic types of inputs, those derived from indexing and those from searching, each of which has its own proper and improper usage of terms. These inputs have a common vocabulary with a fixed number of terms.

It can be considered that recall in output has its corollary with use in input. The processed inputs can be specified as (1) indexed documents, (2) formulated user queries, and (3) the terms used by both the indexers and the formulators of the user queries.

The factors of recall and precision are discussed by Lancaster[3], who states

> Whereas, the recall capabilities of an index are determined by a policy decision relative to exhaustivity rather than by an intrinsic property of an index language, the precision capability of an index is entirely dependent upon the ability of the index language to describe topics precisely (i.e., upon its specificity).

Depth of indexing has been used in two contexts in the literature:

1. application of additional terms to cover more concepts (increasing the exhaustivity),
2. or to index a limited number of concepts more exactly (increasing specificity).

---

[3]F. Wilfrid Lancaster, Information Retrieval Systems (New York: John Wiley & Sons, Inc., 1968), p. 58.

Greater exhaustivity implies using a greater number of index terms. Greater precision is obtained by a greater preciseness of class definition, therefore,

recall $\approx$ exhaustivity
precision $\approx$ specificity.

Greater exhaustivity reduces precision for two reasons:

1. it includes more peripheral items,
2. in some systems there is greater probability of "false coordinations" of terms.

Greater precision is achieved by

1. better class definition,
2. fewer class definitions implying fewer "false coordinations" of terms.

A given reference retrieval system has its level of operation controlled by (1) the prescribed depth of indexing (exhaustivity) and (2) the capability to describe the contents of any document that may be indexed by its specified vocabulary (specificity). Inherent in the system is the probability of omission of relevant terms and the inclusion of nonrelevant terms in both indexing and searching. These factors interact to control the number of references to documents in the output, including the number of desired and undesired references to documents obtained and the number of desired but omitted references to documents from the output.

Index Term Usage Errors

As the indexing of a document progresses from application of the first to the last term applied, the omission

error expresses the existence of terms which are needed but have not been used at any point in indexing. Initially, none of the needed terms have been used. Therefore, as the number of terms used in indexing increases, the probability of the omission error decreases until at some level there should be no omission error. Conversely, as the number of terms used in indexing increases, the probability of using extraneous terms, or commission error, increases, Commission error can be expressed as the application of unneeded terms in an attempt to express the intellectual contents of a document. Therefore, as the number of terms used in indexing increases, the probability of committing a commission error would increase. At some finite level the omission error will be zero, and all terms being added would be essentially commission error terms.

A general outline of the indexing terms, as expressed by their need and use, is shown in Figure 3.1.

For this analysis the number of needed terms will be analyzed first.

In general, $XNU = \sum_{i=0}^{s} xn_i$, where $i = 0,1,2,---,p$

where $s \leq p$

p = number of cells or positions that are available for application of index terms,

if xn = 1, it designates that a given cell has been indexed with a needed term.

$0 \leq XNU \leq XN$.

XNU is the function which designates the intermediate accumulation of the number of terms needed to define the contents of a document, assuming no error.

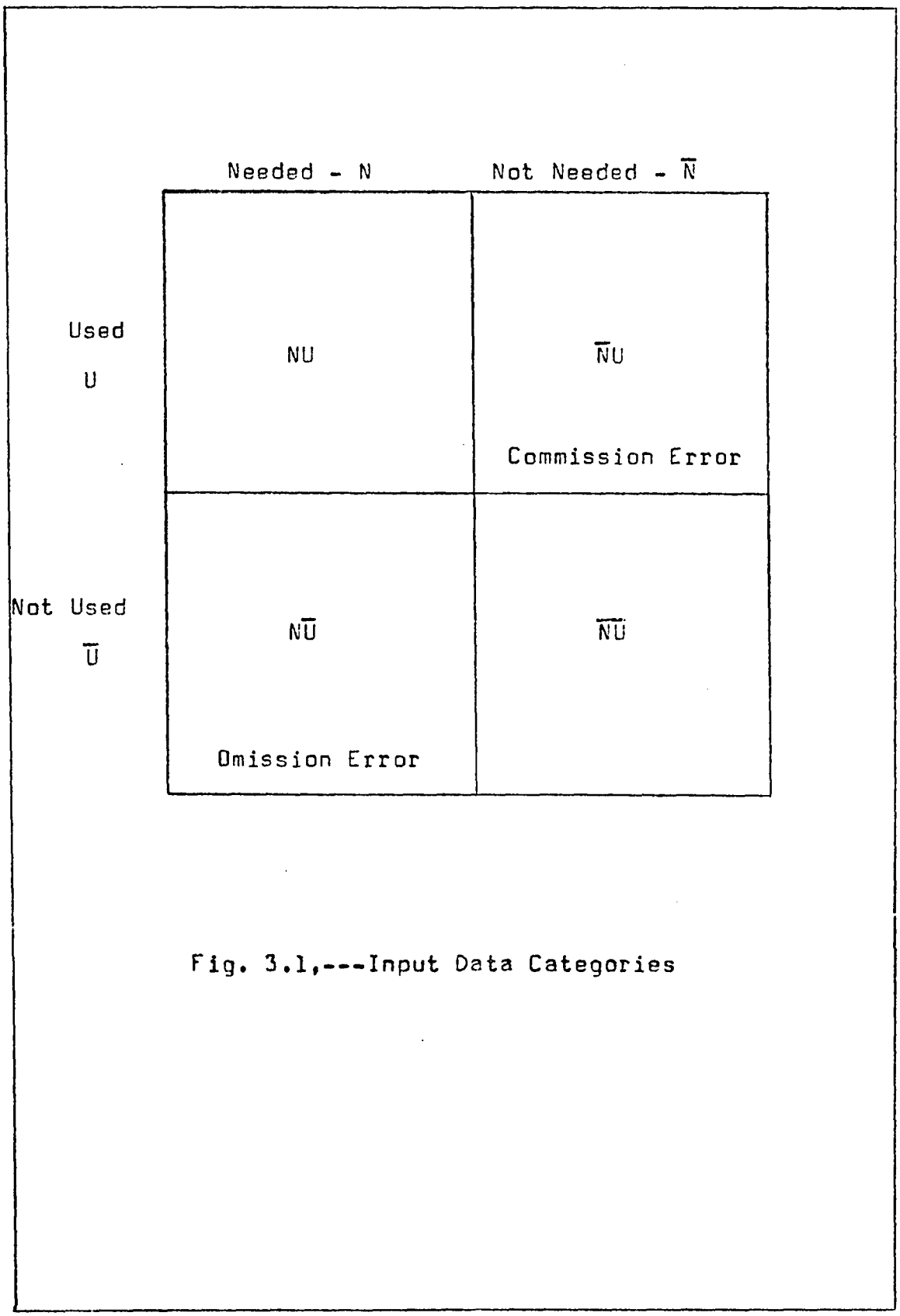


Fig. 3.1,---Input Data Categories

XN has some numerical value for the given level of exhaustivity and specificity of indexing: it is the number of terms needed to express fully all aspects of the contents of a document so that the omission error is zero, which is independent of the commission error.

Since, in general,

$$XN = XNU + XN\bar{U}, \tag{3.1}$$

and

$$XNU = XN - XN\bar{U}.$$

For example, $XN\bar{U}$ designates the number of index terms needed but not used.

The relationship of the needed index terms is shown in Figure 3.2. As shown for Figure a, XNU increases to XN, if a perfect relationship is assumed between the number of terms used and their need.

Similarly, in Figure b, XNU increases and $XN\bar{U}$ decreases.

The number of terms used to describe the contents of a document is analyzed as follows:

> XU is the variable used to designate the intermediate accumulation of the number of terms used to define the contents of a document.

XMAX = limit X MAXIMUM

$0 \leq XU \leq XMAX$,

> where XMAX is the number of terms that must actually be used to define the contents of a document so that the omission error is zero. It has some numerical value for the given level of exhaustivity and specificity of indexing.

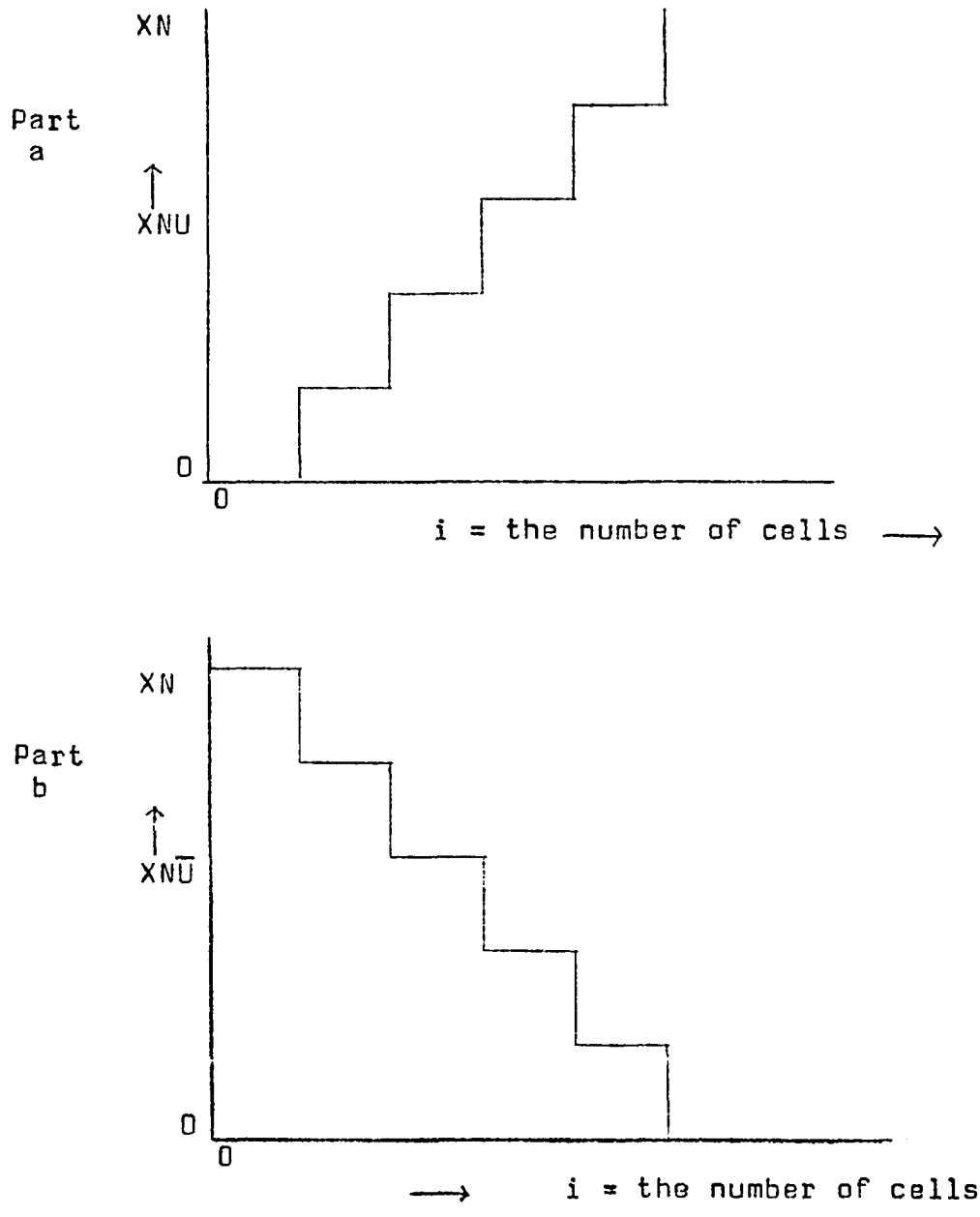


Fig. 3.2,--Ideal Relationship of Needed Index Terms

Since,

$$XU = X\bar{N}U + XNU \qquad (3.2)$$

and the expression for the number of terms not used is

$$X\bar{U} = XN\bar{U} + X\bar{N}\bar{U}$$

therefore,

$$XMAX = XU + X\bar{U}$$

$$XMAX = XNU + X\bar{N}U + XN\bar{U} + X\bar{N}\bar{U}. \qquad (3.3)$$

Similarly, with the number of needed terms,

$$XMAX = XN + X\bar{N}$$

where,

$$X\bar{N} = XN\bar{U} + X\bar{N}\bar{U}$$

and,

$$XN = XNU + XN\bar{U}, \quad \text{from equation (3.1).}$$

Therefore, if

$$XU = XMAX,$$

$$XMAX = XNU + X\bar{N}U.$$

The relationship of the used index terms is shown in Figure 3.3. Part a shows the general increase in XU to the value of XMAX. Part b shows XNU and $X\bar{N}U$ both increase until their composite effect equals XMAX.

As can be noted, a group of terms are a subset of both the needed and the used terms; therefore, both of these terms are needed and used, the XNU terms, where

$$XNU \leq XN,$$

and $$XNU \leq XU.$$

The composite effect of all of the changes in categories

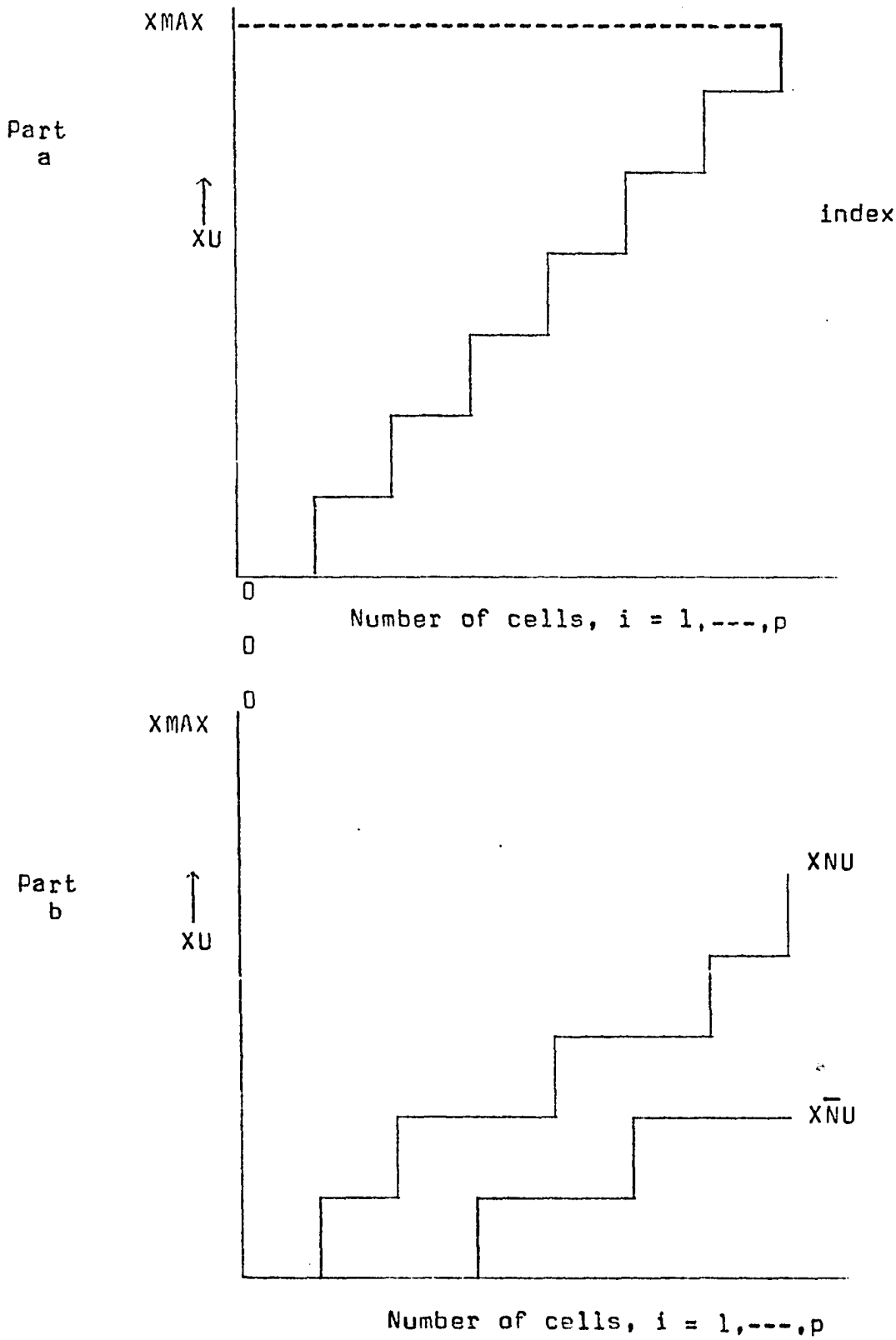


Fig. 3.3,--Relationship of Used Terms

of terms as XU increases can be viewed in Figure 3.4, which shows that, as XU, XNU, and $X\bar{N}U$ increase, $XN\bar{U}$ decreases.

Based on the premise that there is a finite number of concepts, which are expressed in a finite number of terms, there must be some value of XU for which each incremental XU will have little probability of being in the XNU category. This conclusion is based on the assumption that initially, in indexing, each term used has a high probability of being needed, and with each successive term used in indexing there is a decreasing probability of it being unique in expressing some aspect of a document. Therefore, the probability of omission error decreases, and the probability of commission error in the usage of a term increases with the number of terms used.

These results coincide with Lancaster's[4] work, who says

> These figures, of course, demonstrate the customary effect of variations in indexing exhaustivity: the more terms used, the greater will tend to be the recall but the lower the precision; the fewer, more selective the terms used, the lower will tend to be the recall and the higher the precision.

THE PREVIOUS WORK HAS INDICATED THAT THE EXISTENCE OF AND THE EXTENT OF THE OMISSION AND COMMISSION ERRORS CAN BE EXPRESSED AS A FUNCTION OF THE NUMBER OF TERMS USED IN INDEXING. The locations of some of the major points of interest have been discussed. The first is the origin XO, where no terms have been used in indexing. Therefore, the omission

[4]F. W. Lancaster, Evaluation of the MEDLARS Demand Search Service (Bethesda, Maryland, National Library of Medicine, January, 1968. Report No. PB 178-660), p.57.

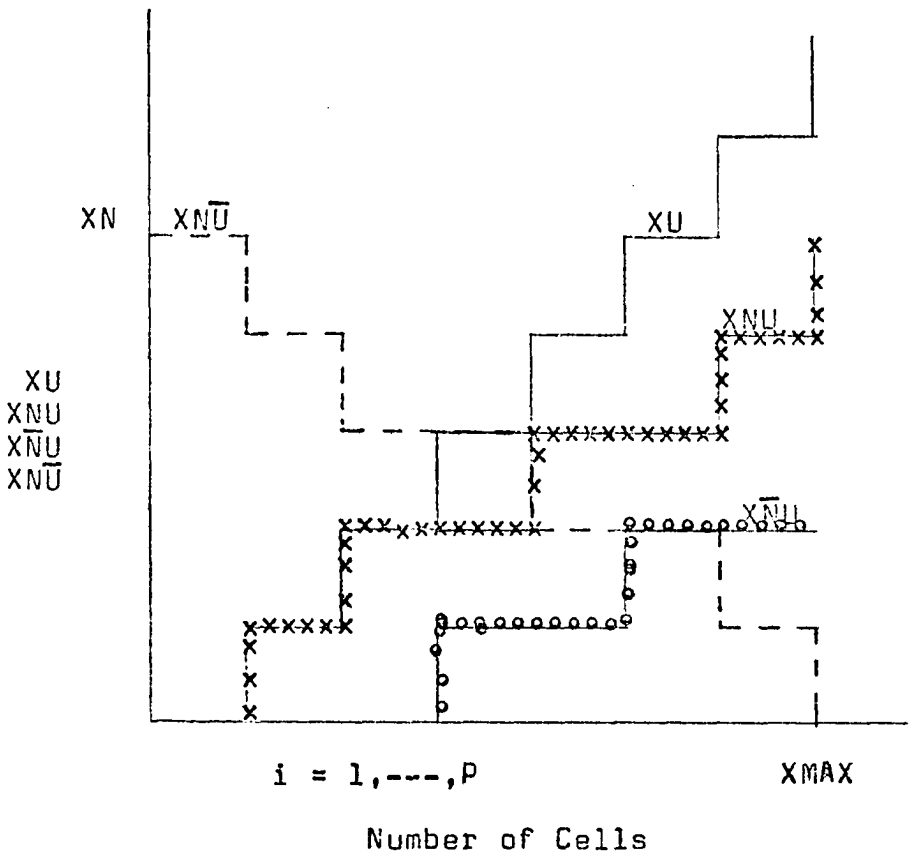
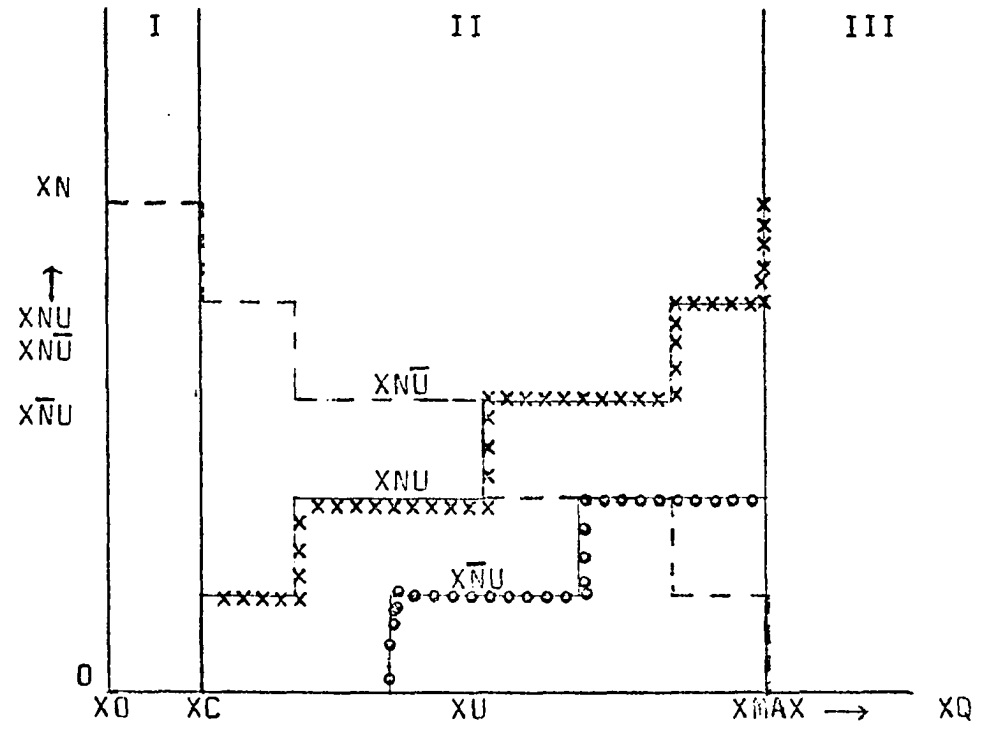


Fig. 3.4,--Combined Effect, All Categories of Terms

error is maximum, and the probability of commission error is P(c)=0. The second point is where the probability of omission error changes from P(c)=0 to P(c)>0. The third point of interest is where the probability of omission error is close to or equal to zero within some confidence limit and occurs after applying some finite number of indexing terms to describe a document. A fourth point can be defined at some distance beyond the point of "zero" omission error, where all terms being used are commission error terms.

Therefore, four points can be considered, based on the sequence of the terms used in indexing. They are shown in Figure 3.5. At the origin is the point where the probability of commission error equals 0 and is designated XO. At some point, the probability of commission error becomes greater than zero, which is designated XC. The minimum extent of this point is XO. At some finite number of indexing terms, the point where the probability of committing an omission error approaches 0 will be reached. This point is designated XMAX. The maximum extent of the third point would be the total list of all the sequential numbered terms in the vocabulary. This limiting value is designated XQ. Since four points can be defined, three areas can be defined as existing between these points. Area I is defined as that area that covers the interval in which there is probability of omission error, and commission error will be zero, and the area lies between points XO and XC. Area II has



The sequential order of the terms used in indexing a document

Fig. 3.5,--Error Regions Vs. The Number of Terms Used in Indexing a Document

probability of omission and commission error. This area lies between XC and XMAX. Area III is that area in which there is negligible probability of omission error. Therefore, the probability of commission error is $P(c) \simeq 1$. That is, for each XU term applied in indexing, $P[XU = X\bar{N}U] \simeq 1$. This region extends from XMAX to XQ. It is assumed there is in Area I and II a functional relationship between the number of terms used in indexing, XU, and the number of terms used but not needed, $X\bar{N}U$.

Index Vocabulary Usage Errors

The work in the preceding section has argued that a general relationship between index term usage and degree of error exists.

The relationship

$$XU = XNU + X\bar{N}U, \text{ from equation (3.2),}$$

gives a general numerical relationship of the number of error terms of the commission category plus the number of terms used in indexing.

Using the expression

XU = the number of terms actually used in a given system,

one can see that the total number of term uses in the entire system is the number of terms per document times the number of documents indexed, or

$$\pi_1 = XU\ D. \tag{3.4}$$

The relationship can be shown diagrammatically by the

representation in Raver's [87] work, which has been reconstructed in Figure 3.6, where

D = document numbers

J = index terms.

This diagram shows that the cumulative frequency of usage of terms must equal the number of terms per document times the number of documents which, in both cases, is the sum of the linkages between index terms and documents.

Similarly, the cumulative number of uses of all terms in the vocabulary must be equal to this amount. If it is assumed these terms are geometrically distributed, it can be expressed in current variables by using equation (2.4) as

$$V_j = \frac{(1-B)\ B^{j-1}}{(1-B^q)}\ XU\ D$$

and the cumulative distribution is expressed

$$\pi_v = \sum_{j=i}^{q} V_j = \sum_{j=1}^{q} \frac{(1-B)\ B^{j-1}}{(1-B^q)} XU\ D$$

$$\pi_v = XU\ D$$

so that from equation (3.4)

$$\pi_v = \pi_I .$$

Therefore,

$$V_J \approx \pi_v = \pi_I = XU\ D$$

and

$$XU \approx V_j .$$

It can be seen that the relationship of errors in indexing can be related to parameters of the total index-term

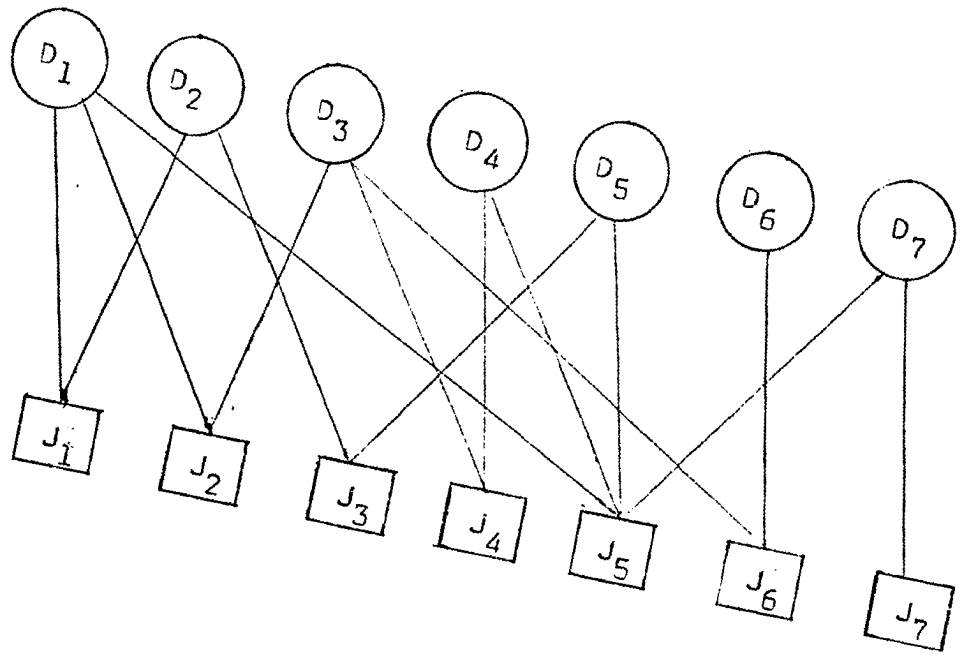


Fig. 3.6,--Linkages Between Documents, Index Terms and Their Frequency of Usage

vocabulary distribution, if one assumes that the probability of errors in indexing is applicable to each of the terms of the vocabulary. Therefore, to analyze the fre-quency of errors of the total vocabulary term uses, one can make the following assumptions:

1. The probability of error in the usage of index terms is a function of the number of times a term is used.

2. The probability of a given term being used erroneously can be either directly or inversely proportional to the frequency of usage of the term.

3. The probability of a term being used in error is a constant. This latter situation implies that collectively each term usage has the same probability of being used

   a. correctly,

   b. in an omission error situation with a constant probability of k_1,

   c. and in a commission error situation with a constant probability of k_2,

      where, k_1 is not necessarily equal to k_2.

This latter assumption implies that all terms make the same amount of information contribution in terms of gross contribution and have the same degree of error as shown in Figure 3.7. IF THIS ASSUMPTION IS NOT TRUE, THE IMPLICATIONS ARE THAT SOME TERMS ARE VERY PRONE TO BE USED IN ERROR AND THAT THE QUALITY CONTROL OF INDEX TERMS IS HIGHLY VARIABLE.

The existence of and the extent of errors of omission and commission in indexing as a function of the number of

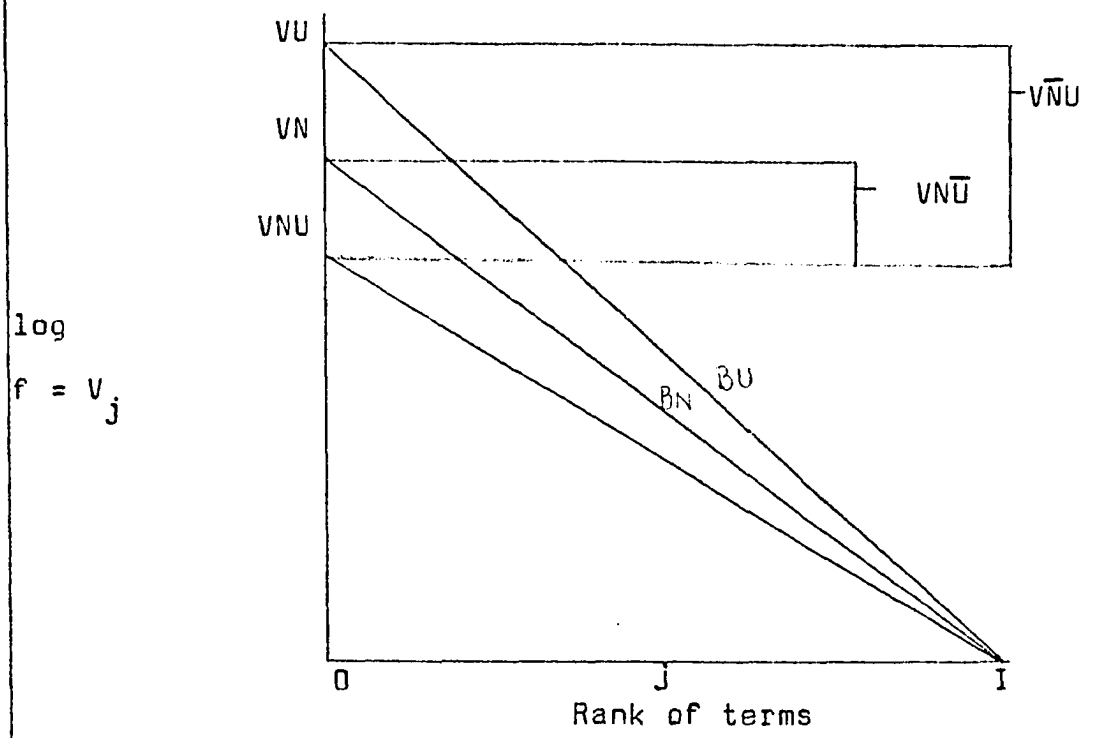


Fig. 3.7,---Vocabulary Index Term Frequency

Where VN = Constant for the needed term distribution on a frequency basis, assuming perfect indexing,

VU = Constant for the used term distribution on a frequency basis,

I = Number of terms in the indexing vocabulary,

BN = Parameter of needed term distribution,

BU = Parameter of used term distribution.

terms used, given the specified levels of indexing, has been presented as well as the ratio of error in indexing to that of total vocabulary term usage.

The concepts as developed in indexing can be directly adapted to those of searching, where there can also be omission and commission errors. The expression of terms will change from

XN, XU, XMAX and others to those of

YN, YU, YMAX, similarly.

Search Formulation

In addition to defining inputs in terms of indexing and searching and their associated errors, the input of searching can be formulated into two polar Boolean approaches, intersection and union.

Intersection Search. An intersection search is formulated with the following conditions:

1. A population of search terms E, with a series of formulated searches containing one or more elements of E, designated E_e,
2. A population of index terms I, with a series of documents indexed with 1 or more elements of I,
3. The output, ZU, equals the number of references recovered and is a function of indexing and searching formulation.

The number of elements in a search can be considered fixed for a given system and will be defined as YU. Thus, for intersection searching there are YU terms,

where $0 < YU \leq E$.

Similarly, in indexing, the number of elements used in indexing a document can be considered fixed for a given system and will be defined as XU,

where $0 < XU \leq I$, if, XU is an integer.

The subset of E, E_1 $(E_{B1}, E_{B2}, ---, E_{BN})$, is then matched with the population of I to find that series of documents that are indexed with the terms specified in the search. Only those references to documents that are indexed with all of the elements specified in the formulated search will then be retrieved and will be expressed as output, ZU.

In an intersection search only those references to documents that can be retrieved under any search are those documents indexed with more terms than those specified in the formulated search.

Therefore, $ZU \geq 0$, if $YU \leq XU$.

> ZU is proportional to the number of documents indexed with more terms than those specified in the intersection search.

or $$ZU = f(XU, YU). \qquad (3.5)$$

However, if the number of search terms is greater than the number of index terms, there will be no complete mapping of search terms into index terms and output will be zero.

If $YU > XU$,

$ZU = 0$.

Union Search. The union search is formulated by using one of the searches containing one or more elements of E_F,

$(E_{F1}, E_{F2}, ---, E_F)$.

All of the references to documents containing one or more of the elements specified in the formulated search will then be retrieved and will be expressed as output, ZF.

$$ZF \approx XU(YU).$$

Therefore, the number of index terms per document increases output. Similarly, the output increases with the number of search terms, YU.

It can be stated, therefore, that the quality and quantity of the output of a reference retrieval system is dependent on (1) the quantitative relationship between the number of index and search terms and the number of citations to documents recovered in a search which can be described as the level of performance model output and (2) the type and extent of errors in the index and search terms.

Level of Performance Model Output

A performance model is a procedure to determine the number of references to documents obtained, given the parameters of the indexing and searching aspects of the system. A simulation technique for ascertaining the number of references will be developed, which subsequently will determine the effects on output of various combinations of errors in inputs.

In the performance model developed by A. D. Little, Inc. [88], a factor to compensate for the difference between

theoretical numbers of items retrieved and the actual number of items retrieved was used. However, it must be noted that in their work although they did ascertain the frequency of terms used in indexing, they assumed that the frequency of terms in the searching vocabulary was a fixed ratio of the terms used in the indexing vocabulary. Inasmuch as the indexing and searching are independent functions, it seems unlikely that the frequency of usage of each of the search terms will be a constant proportion of their frequency of usage in indexing. An independent determination will be made of both the frequency of the number of terms used in formulating searches and in the frequency of the individual terms used in searching. This independent determination will eliminate the modification factor that is needed to correlate the theoretical model with the results obtained in actual practice.

A procedure to relate the independent distributions of the index terms and the search terms, assuming an intersection search, will now be discussed. This work by A. D. Little shows that the probability of usage of any term j in indexing, equation (2.1), could be expressed as

$$g_1(j) = \frac{(1-B)\ B^{j-1}}{(1-B^Q)}.$$

The actual number of documents to be indexed under the jth term, having a rank of j, can be expressed by using the form of equation (2.4),

$$V_j = \frac{(1-B)\ B^{j-1}\ \bar{X}\ D}{(1-B^q)}.$$

Set $f_1(j) = V_j$.

Therefore, in current variables

$$f_1(j) = \frac{(1-BI)BI^{j-1}\bar{X}\ D}{(1-BI^q)} \qquad (3.6)$$

BI = constant that specifies the slope of the index term frequency distribution.

The probability of using the kth term in formulating a search can be similarly expressed

$$g_1(k) = \frac{(1-BE)BE^{k-1}}{(1-BE^{\emptyset})}, \qquad k = 1,2,---,\emptyset$$

where g(k) = probability of using the kth term in formulating a search,

BE = constant that specifies the slope of the search term frequency distribution,

k = rank of search terms.

The expected number of references to documents to be retrieved by using one term will be the number of references to documents times the probability of selecting the kth search term summed over all k terms in the search vocabulary. This procedure assumes that the rank of the search terms is the same as the rank of the index terms.

$$\bar{Z} = \sum_{j=1}^{q} \sum_{k=1}^{\emptyset} f(j)\ g(k)\ \bar{X}\ D.$$

Where, $\bar{Z}$ = the expected number of references to documents to be retrieved.

However, it is proposed by the author that the functional expressions for the index and search relationships reflect different constants for the slope of the distribution of terms. Therefore the following expression is derived.

$$\bar{Z}_i = \sum_{j=1}^{q} \sum_{k=1}^{\emptyset} \frac{(1-BI)BI^{j-1}}{(1-BI^{\emptyset})} \frac{(1-BE)BE^{k-1}}{(1-BE^{q})} \bar{X} D.$$

Since the number of terms in the indexing vocabulary, q, is equal to the number of terms in the searching vocabulary, $\emptyset$, the equation can be written as follows:

$$\bar{Z}_i = \sum_{j=1}^{q} \sum_{k=1}^{q} \frac{(1-BI)(1-BE)BI^{j-1}BE^{k-1}}{(1-BI^{q})(1-BE^{q})} \bar{X} D.$$

As the number of index and search terms changes, this formulation becomes rather complex. Since q can range up to 10,000 in value, a two term problem such as the one treated here can give computational problems. A three term problem would be intractible.

A mathematical model has been generally developed to express the output of a reference retrieval system on the basis of the expected number of references to documents by using an intersection search. The limitations on values are that the number of search terms must be less than the number of index terms. This model uses independent determination of terms, assuming an intersection of search terms. However, this model does not adequately handle all input error effects. Therefore, it is proposed to design a simulation model that will use the data as prescribed for the

proposed level of operation (including the errors). The output, on the basis of the number of references to documents of various categories, will be determined.

Proposed Model

Of primary concern is a means with which to determine the performance, cost, and value of a reference system so that its performance can be optimized. Performance is measured in terms of the number of cited documents retrieved from a system in response to formulated user queries interacting with the indexes of cited documents. Total cost is based on the necessary monetary expenditures incurred to produce output and the cost of the user's effort to initiate the search and evaluate the retrieved output. Total value is obtained by determining a per-unit price for usable references, then relating the price to the total number of usable citations. By its description this will be a system that operates with terms to describe the subject matter of a document. These terms can be identified and the number counted. This type of system is typified by the use of coordinate indexes. Much of the theory previously reported was based on document retrieval systems. It is recognized that the users of any information system must, at some stage, have access to the contents of documents (in some form). However, the concern here is only with reference retrieval systems. The value of the reference retrieval system must

be considered in terms of the larger aspect of document identification retrieval. The main concern of the system is with identifying the relevant documents, not in obtaining them.

Therefore, the proposed model consists of two phases, the reference retrieval model and the total cost-value model.

Reference Retrieval Model

The reference retrieval model consists of three stages (1) the error-determination technique, which was presented earlier in this chapter, for measuring "omission" and "commission" error in indexing and searching, given the number of terms and their categories, (2) a performance model for calculating the expected number of references to documents in each category, and (3) the output evaluation procedure.

The second phase develops a cost and value structure to evaluate policy changes in indexing and searching and to provide a procedure for optimizing the system, given the constraints.

Error Determination Technique. The technique for developing and measuring omission and commission error in indexing and searching was presented previous to the "Proposed Model" section in this chapter.

Performance Model. The use of a mathematical approach to develop a simulation model will treat the performance of an actual reference retrieval system as a "black box." The quantification of inputs and their usage in the model will

allow analogizing of an existing system as searching operates on the references to indexed documents to produce output of citations to documents. The output of the performance model will be the expected number of references (citations to documents) of the relevant-recalled, the nonrelevant-recalled, and the relevant-nonrecalled documents. The first of these outputs, the relevant-recalled, is desired; and the latter two outputs represent errors of the system. The performance is based on the following parameters and factors of the system to be determined as in Table 3.1.

Table 3.1. Parameters and Factors of the System

Endogeneous Variables

Inputs

XU = Number of terms actually used to index a document.

YU = Number of terms actually used to formulate a search.

Each of the above two inputs may exist in the two categories; needed-used, not needed-used; the needed not-used category is omitted.

Outputs

ZN = Number of references to documents obtained as output with perfect indexing and searching.

ZU = Number of references to documents actually obtained as output.

ZNU = Number of references to documents needed and obtained as output.

$Z\bar{N}\bar{U}$ = Number of references to documents needed, and obtained as output.

$\bar{Z}\bar{N}U$ = Number of references to documents not needed, but obtained as output.

As with the inputs, the output may be two of three categories, NU, $\bar{N}U$. Also, the output is the result of a process expressed by two independent variables. Its output will be a variable with the associated categories of NU, $N\bar{U}$, and $\bar{N}U$.

Exogeneous Constants

D Number of indexed documents in the reference file.

S Number of user queries to be formulated annually.

I Number of terms in the indexing vocabulary.

E Number of terms in the searching vocabulary.

T Number of searching installations.

R Number of new documents indexed annually. (replacements)

A Number of search files repreduced annually per installation as related to the number of new documents indexed annually.

Endogeneous Constants

XMAX Total number of terms needed to be used in indexing a document to avoid any omission error.

YMAX Total number of terms needed to be used in formulating a search to avoid any omission error.

XN Number of terms needed to index a document, error-free.

YN Number of terms needed to formulate a search, error-free.

The performance model is designed to show how the exogeneous and endogeneous constants interrelate and how these variables interact. Their end result or output are the retrieved completed user queries. Parallel structuring of the model, based on intersection or union searches, are necessary to accommodate two approaches to formulating searches. Previous work by A. D. Little [89] will be expanded and modified to facilitate a simulation approach of a performance model.

The functional relationships, pertaining to the performance model, are expressed as follows:

1. Number of terms per indexed document

   $X = f(I), \qquad 0 < X < \underline{I},$

2. Number of terms per formulated user query

   $Y = f(E), \qquad 0 < Y \leq E,$

3. Number of citations to documents per retrieved user query

   $Z = h(X,Y)$, but, it is also

   $Z = f(I,E/D),$

4. Total term usage in indexing for all documents

   $\pi_I = F(X/D,I),$

5. Total term usage in searching

   $\pi_E = F(Y/S,E),$

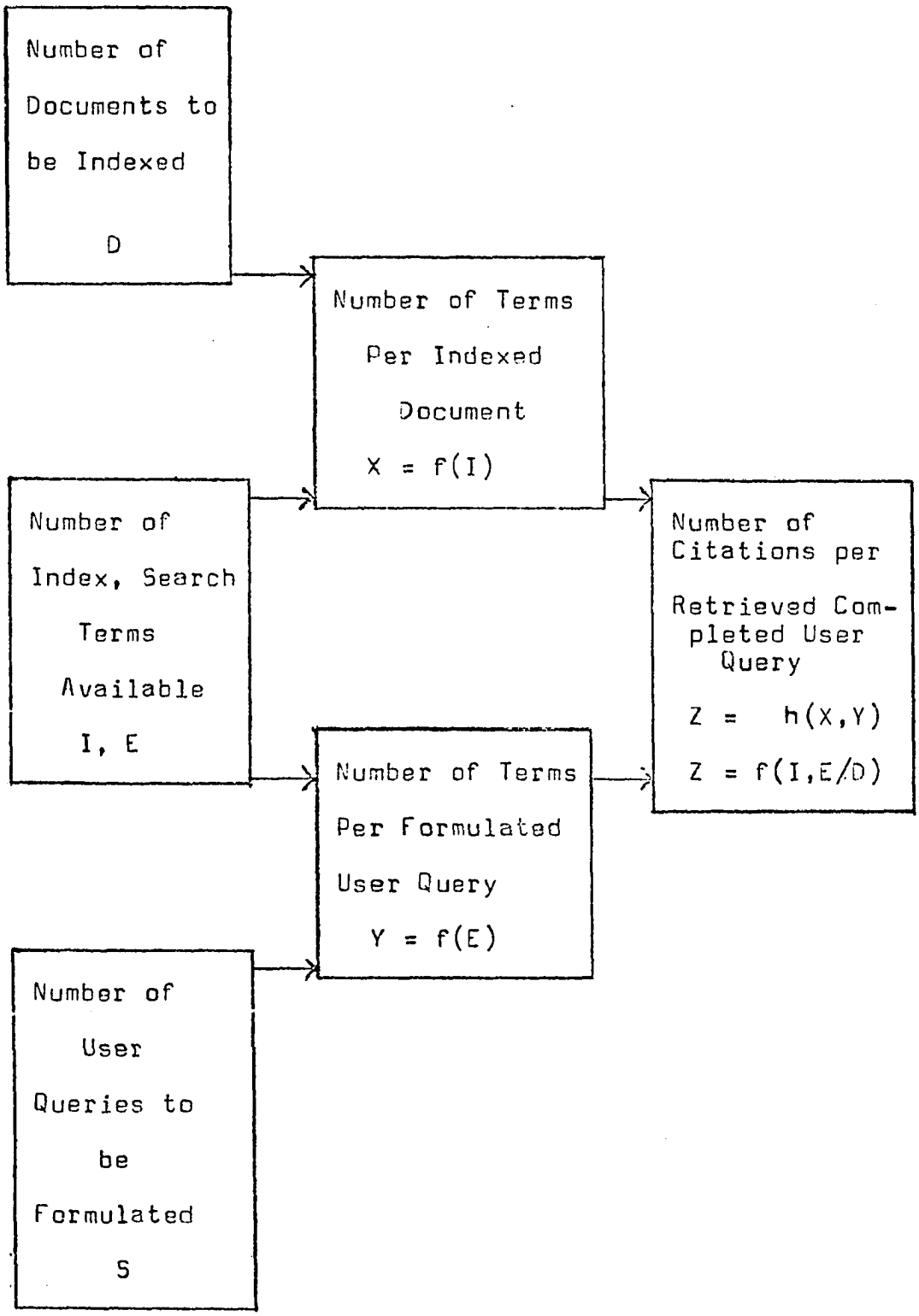


Fig. 3.8,--Performance Model

6. Total number of retrieved citations to documents in all retrieved searches

$$\pi_Z = F(Z),$$

$$\pi_Z = F(X,Y),$$

$$\pi_Z = F(I,E/D,S),$$

$$\pi_Z = \Sigma 2.$$

Output Evaluation. The output evaluation procedures utilize the output from the performance model and interrelate it with the inputs. Since there are two inputs of index and search terms, there is no direct mathematical way to relate directly the level of output back to the indexing or searching variables. Therefore, alternative indirect procedures must be devised.

Total Cost-Value

Knowledge of the cost structure of the system is needed. The costs must be related to the inputs so that changes in the inputs of indexing and searching and their associated cost changes are properly reflected in (1) the number and categories of output of citations to documents, and (2) the total costs of the system.

The general format is the development of production functions to relate the independent input variables to a stage of output but expressed at the level of usage.

A production function is a means of expressing the

physical relationship between the input quantities and the composition employed in the production process and the output quantity yielded by this process. Application of relevant cost data for the various inputs will allow determination of total costs. The cost and value model will be developed in a series of sequential stages, each of which represents one phase in the operation of a reference retrieval system.

The variables of inputs and outputs, X, Y, and Z, will be subscripted to identify the various phases of intermediate output. Application of relevant cost data to each of these phases will allow formation of a set of independent cost functions. Summing up these various cost functions will yield the total cost function for the entire reference retrieval system. A similar development of value functions is derived.

Total System Production Functions. The production functions for the system will be subdivided into two phases for total cost of facilities and one phase for the total output benefits.

The two phases for the total cost of facilities are

1. Total initial facilities investment and indexing facilities

   These include the facilities for obtaining, indexing and storing the documents, index data preparation and production of accessible search files

and equipment costs. A series of functions measuring the increase in labor, services, material, and equipment used as the level of index terms used, XU, increases as related to prescribed activities will be developed. This scope of activities will be independent of the number of search terms, YU.

These functions will be formulated so that when the cost of the factors are introduced, it will be in the form of an equivalent annual cost.

2. Operating functions

A series of production functions will be generated for indexing and searching separately. These functions will relate the quantity of labor, services, and material that correspond to the feasible range of index and search terms being considered.

The series of functions for indexing will relate to (1) primary indexing and (2) data input and constants to the system. Since this is presented on an annual basis, the number of documents indexed annually must be considered.

The functions for searching can be grouped as follows: (1) those directly relating to the search formulation and (2) those relating to the file-search operation. The annual number of searches, S, and the number of search installations, T, must be considered.

Total Value of Output Benefits. Value in the real system is obtained by retrieving references to needed documents in response to a specific request.

A positive value can be assigned to each desired reference retrieved and a penalty assigned to each needed document not retrieved.

There is a negative value or cost based on the premise that there is a cost of evaluating all of the items listed

on a completed user query. This cost would include both relevant and nonrelevant items. Therefore, the cost of evaluating the completed query is in direct relationship to the number of items listed. The values and costs for each of the phases are then combined to produce the total cost model.

The model being designed will ascertain the number of documents of the specified categories of output as a function of the number of index and search terms.

Optimization. Optimization can be viewed in the economic sense of

$$\text{Profit} = \text{TVOB} - \text{TCF}, \tag{3.7}$$

where TVOB is designed to include all benefits derived from retrieved citations to documents and the costs to the user for preparing and analyzing the output. Cost, TCF, includes all monetary expenditures, whether they are investment or operating expense. The system, therefore, includes all costs based on the user's effort and the monetary expenditures necessary to develop and operate the system.

The output level of maximum profit is expressed in the economic sense of maximum positive difference between TVOB - TCF, where the value and cost surfaces are based on the independent variables prescribed by the number of terms used in indexing and searching, X and Y.

CHAPTER IV

REFERENCE RETRIEVAL MODEL

The real reference retrieval file consists of references to documents identified by "flags" or index terms. These files are formed by the aggregate of all the documents indexed, along with their index terms. The searching consists of expressing the user's request in terms chosen from the same vocabulary used by the indexers, then interrogating the aggregate index file and recovering those documents that are "flagged" by the search terms. Whether or not the references to documents and their "flags" are expressed in prose or in some compatible notation is of no consequence here. However, if different expressions are used, these expressions must be consistent and compatible. The output of the real system consists of a listing of citations obtained as described above and includes the information needed to identify the documents listed. The expressions used may include factors such as author, title, source, and other data; and they are at the option of the individual system designer.

A model to evaluate the stages of a reference retrieval system must relate the inputs to the system, simulate their interaction, and provide quantified output. The output can then be expressed to relate the quantified output to the level and type of inputs.

The reference retrieval model consists of three aspects,

1. the categorization of the input and associated errors,

2. the ascertainment of the output as prescribed by the level of indexing, and

3. the relation of the level of operation to the level and type of input, since this is an analytic phase. This relationship will be developed in subsequent work.

Input Error Categorization Technique

The model will be developed as related specifically to indexing. The technique consists of the intellectual phase, and the analytic and development phase.

Intellectual Phase

The intellectual phase consists of the concepts involved in

1. designation of discipline of interest, which includes,

   a. describing the defined discipline, and

   b. the rules for choices of indexing vocabulary terms of a reference retrieval system,

2. the document content determination concepts, which include

   a. the operating level of indexing,

   b. indexing rules,

   c. factors of indexing evaluation.

Discipline Designation. The discipline designation includes (1) the definition of the subject and the prescribed area to be covered by a reference retrieval system, and (2) the rules for choosing the indexing vocabulary terms

consistent with (1).

The relationship of the number of terms needed as a function of the document collection size has been analyzed in historical performance of operating systems as shown in previous work. In A. D. Little's [90] work, equation (2.2) showed that

$$I = 18\sqrt{D}$$

and in Houston and Wall's [91] work, equation (2.3) showed

$$I \approx \sqrt{\bar{X}\, D}.$$

Therefore, the number of terms needed in the indexing vocabulary is a function of the document collection size and can be determined so that an adequate number of terms are available to uniquely identify the content of any and all documents in a collection.

The subject of the discipline of interest must be identified and defined. The concepts to be included must be expressed, and the vocabulary terms chosen must be consistent with discipline subject. The definition of terms to be used must be in accord also. The following items must be considered in developing this vocabulary.

1. The terms must be mutually exclusive.
2. All aspects of the discipline being considered are to be included.
3. All terms are to possess or have equal value for expressing information.
4. The entire population of the discipline of interest is to be expressed by an appropriate number of terms.

Document Determination Concepts. The document determination concepts include factors concerned with the operating level of indexers, the indexing rules, and indexing evaluation considerations.

The application procedure consists of conceptual consideration of exhaustivity and specificity. Exhaustivity is expressed as the decision of the management of the system to define the number of concepts that exist in a document that are to be recognized and indexed, which is the operating level of the system. Specificity is the restriction placed on the indexer by the operating level of the system that defines his range of the number of terms to be used to describe the contents of a document.

Since, for purposes of control there must be a procedure for evaluation, this procedure is also included. Therefore, three phases are defined.

1. The number of concepts that are to be recognized and expressed in evaluating a document for indexing are defined. This definition is prescribed by management and is the level at which the system is operated, where each document is indexed by some number of terms.

2. The operating procedures for indexers are

    a. to identify the concepts of the document,

    b. to ascertain which concepts are pertinent to the subject area,

    c. to arrange the concepts in descending order of significance,

    d. to express these concepts in terms of the vocabulary, that is, specificity.

Indexing is an attempt to describe the intellectual contents of the documents and is not a quality control of the contents of the documents themselves.

3. Operating procedures for indexing evaluators are

   a. to make an independent determination or re-indexing of any document being considered by using the vocabulary as given,

   b. to then ascertain the degree of adherence to the operating rules for indexers and determine the type and degree of disagreement.

Analytic and Development Phase

The analytic and development phase, which defines a procedure for quantifying the inputs of indexing and searching and their errors, is divided into three parts. They are as follows:

1. the functional relationships,
2. testing procedure and experiment conduct,
3. parameter determination.

Functional Relationships. Functional relationships express the category of an index or search term with the level of indexing or searching. The prescribed conditions are (1) a specified number of concepts given, (2) a set of vocabulary terms to express these concepts, and (3) the presence of "omission" and "commission" errors. Therefore, the number of terms needed will be equal to or greater than the number of concepts given. The requirements are (1) a procedure to determine the number of terms needed to express all aspects of the given concepts, assuming perfect indexing,

(2) accepting the probability of errors that exist in the indexing phase as applied, and (3) devising a procedure for determining the number of terms actually needed to define fully all aspects of the concepts given the existence of errors. Previous work has shown that there are, theoretically, three areas of knowledge which can be defined, assuming that there are an adequate number of terms in the vocabulary to express all aspects of the concepts, which are designated areas I, II, and III. Area I is that region where there is no probability of omission errors. Area II is where there is probability of omission and commission errors. Area III is the region where there is no probability of omission errors, but the probability of commission errors is equal to 1.

The indexing can be considered as a group of m cells in which the terms to define concepts will be the number of cells used. Indexing is, essentially, a sampling of vocabulary of terms; however no one term can be used more than once. Therefore it is sampling without replacement. Thus various levels of indexing are not independent. Therefore, a factorial analysis experiment, as such, can not be used in evaluating the outcome of an experiment on this system.

Given that a term is to be applied to a document, the applicability of that term as related to the concept can be considered on a 0,1 basis. Either it is applicable or it is not applicable. However if several documents are considered or a document is reindexed several times, indexing

a document at the m^{th} cell is sampling with replacement, and the $(P_j(m)) = P_1$ is constant, where P_j is the probability of a term being used properly. Therefore if the number of concepts and other variables are held constant, the probability of P should decrease as the number of terms used in describing a document increases.

Testing Procedure and Experiment Conduct. The testing procedure consists of having a representative number of documents indexed by using enough applicable indexing terms so that the appropriate categories can be defined. These indexed documents must then be analyzed, using the same indexing rules, to ascertain the applicability of the terms used. The fixed elements of indexing are (1) documents, (2) indexing terms (indexing vocabulary), (3) indexers, and (4) qualified evaluators.

The other three factors are (1) the choice of documents to be indexed, (2) the choice as to which documents a given indexer will index, and (3) the decision as to which documents and indexers the evaluators will review. The above three factors can be sampled or randomized to minimize the effect of interaction. The documents to be indexed would be chosen at random from the population of documents. The choice of the indexers versus the documents would be randomized. Similarly the choice of the evaluators versus the documents would be randomized.

The experiment itself consists of having randomly

selected individuals index documents picked at random. The choice of terms used in indexing is based on the rules prescribed under indexing, at the prescribed level of exhaustivity and specificity. The number of terms that must be applied having been previously determined, the identify of each index term and its sequential order number must be recorded for each document in order to ascertain the applicability of the index terms to describe the various aspects of the concepts of a document that are of interest to the given retrieval system. The evaluators must record, in the sequential order of the indexed terms, the applicability or the nonapplicability of all the terms per document. Upon completion of the evaluation, a summary of terms will be available which can be presented as shown in Table 4.1.

Parameter Determination. The first step in parameter determination will be to evaluate the data in Table 4.1. These data will be summarized, and the value of $\bar{P}$ for each cell will be determined as follows by assigning a value of 1 for a success and 0 for a failure. Following termination of the indexing and evaluation, the number of terms for each cell (j) are recorded, and the estimator of $\bar{P}$ is determined

$$\text{where } \bar{P} = \frac{x}{n} \text{ for each cell.} \qquad (4.1)$$

Where $\bar{P}$ = probability of success,

x = number of successes,

n = number of elements.

These data are shown on the bottom line in Table 4.1 and are now ready for usage in the performance model which is subsequently developed in this chapter.

Performance Model

Errors of omission and commission, in both indexing and searching independently, have been discussed previously. Also, a technique for expressing these errors for all necessary categories in a quantified manner have been developed. The procedure for utilizing these quantified values to generate output is the function of the simulation.

The output of the real system, which consists of a listing of citations to documents, will be simulated by a procedure giving the number of documents to be recovered under specified index and search conditions. Since it is possible to have all the necessary categories of inputs quantified, the output of the model will be the number of references to documents. Three categories of output will exist, of which one category will be the desired references and the other two categories will be error output.

It will be possible to ascertain the effect on output of the real system as the indexer and/or the search formulator; each independently commit omission and/or commission errors. The simulation model is constructed so that the effect on output of no errors, or of various combinations of no errors and/or with errors in input, can be measured.

TABLE 4.1, Summary of Evaluated Experimental Data

No. of Documents	Sequential No. of Cells - m						
	j = 1	j = 2	j = 3	j = 4		j = m-1	j = m
i = 1	1	0	1	0		0	0
i = 2	1	1	0	1		1	0
i = 3	1	0	1	0		0	0
i = 4	1	1	1	1		0	0
i = 5	0	1	1	0		0	0
i = 6	1	1	0	0		0	0
i = 7	1	0	0	1		0	0
• • •							
i = n-1	1	0	0	0		0	0
i = n	0	1	0	0		0	0
X	7	5	4	3		1	0
$\bar{P}$	7/9	5/9	4/9	3/9		1/9	0

Where,

j = 1,2,---,m = No. of the cell,

i = 1,2,3,---,n = No. of documents per cell.

This error effect is expressed in the simulation model by the appropriate categories and numbers of terms in each category. The errors of output, which in the real system consist of omissions of citations to relevant documents and inclusions of citations to nonrelevant documents, are determined by comparing the document numbers retrieved for each desired combination of inputs with those document numbers retrieved with an ideal set of inputs. The total number of documents of each category that corresponds to each combination of inputs can then be expressed numerically.

The simulation model describes the inputs to the model and their errors followed by a description of the initial operational stages of the simulation model and the output preparation.

Input Description

The inputs to the model are shown on Table 4.2, which includes data, parameters, and variables for directing the simulation model. Numerical values are shown in the relevant locations that will be utilized to describe the operation of the simulation model.

The errors in the inputs which are carried through the simulation process and used to depict the output are shown in Figure 4.1. These errors are the same as those depicted in Figure 3.1 but are extended further to facilitate computer calculations, including expressing the various categories

Table 4.2. Performance Model

Summary of Input

TOTAL NUMBER OF INDEXED DOCUMENTS IN SYSTEM 15.

NUMBER OF TIMES TO RUN THE SIMULATION WITH THE SAME PROBABILITY OUTCOME IS 1 .

NUMBER OF TIMES TO RUN THE SIMULATION, CHANGING THE (PROBABILITY) OUTCOME EACH TIME IS 1 .

NUMBER OF INDEX TERMS IS JI.

NUMBER OF SEARCH TERMS IS KE.

NUMBER OF TERMS ACTUALLY USED TO INDEX A DOCUMENT IS 10, MAXIMUM 15.

NUMBER OF TERMS ACTUALLY USED TO FORMULATE A SEARCH IS 5, MAXIMUM 9.

PROBABILITY DISTRIBUTION OF INDEX TERMS, (XP(I), I) = .99, .95, .90, .85, .80, .75, .70, .65, .50, .35, .25, .15 .10, .05, .01

PROBABILITY DISTRIBUTION OF SEARCH TERMS (YP(I), I) = .95, .85, .65, .50, .40, .30, .20, .10, .05.

CONSTANT FOR THE PROBABILITY DISTRIBUTION OF SEARCH VOCABULARY TERMS USED IS BEU.

CONSTANT FOR THE PROBABILITY DISTRIBUTION OF INDEX TERMS USED IS BIU.

THIS OUTPUT ASSUMES THAT A SLAB OF 15 IS REPRESENTATIVE OF THE TOTAL POPULATION.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">$N \leq XMAX$</td></tr>
<tr><td colspan="2">N</td><td>$\overline{N}$</td></tr>
<tr><td rowspan="3">$U \leq XMAX$</td><td>U</td><td colspan="2" rowspan="2">$NU = 9$</td><td>$\overline{N}U = 11$</td></tr>
<tr><td>$\overline{N}UR = 5$</td><td>$\overline{N}UT = 6$</td></tr>
<tr><td>$\overline{U}$</td><td>$N\overline{U} = 7$</td><td>$N\overline{U}R = 3$
$N\overline{U}S = 4$</td><td>$\overline{NU} = 0$</td></tr>
</table>

Fig. 4.1,---Input/Output Classification Scheme

Input term nonmenclature would be expressed as follows:

NU = Number of terms needed and used,

$\overline{N}U$ = Number of terms not needed but used,

$\overline{N}UR$ = Number of terms not needed but used, where the total number of terms is less than N,

$\overline{N}UT$ = Number of terms not needed but used in excess of the needed terms, N,

$N\overline{U}$ = Number of terms needed, not used

$N\overline{U}R$ = Number of terms needed, not used that would replace the $\overline{N}UR$ terms erroneously used,

$N\overline{U}S$ = Number of terms needed, not used, that are in excess of the $N\overline{U}R$ terms and equals the remaining $N\overline{U}$ terms.

Note: $\overline{N}UT = 0$
or $N\overline{U}S = 0$

in a numerical form.

Determination of the subcategories can be facilitated using equation (3.3),

$$MAX = NU + \bar{N}U + N\bar{U} + \bar{N}\bar{U}.$$

Since,

$$N \gtrless U$$

where, from equation (3.1),

$$N = NU + N\bar{U}$$

and from equation (3.2) if

$$U = NU + \bar{N}U$$

so is

$$N\bar{U} \gtrless \bar{N}U.$$

Let

$$N\bar{U} = N\bar{U}R + N\bar{U}S$$

and

$$\bar{N}U = \bar{N}UR + \bar{N}UT \qquad (4.2)$$

and let

$$N\bar{U}R = \bar{N}UR. \qquad (4.3)$$

Therefore, if

$$N\bar{U} > \bar{N}U \qquad (4.4)$$

$$\bar{N}UR = \bar{N}U \qquad (4.5)$$

and

$$N\bar{U}S = N\bar{U} - N\bar{U}R \qquad (4.6)$$

if

$$\bar{N}U > N\bar{U}$$

$$N\bar{U}R = N\bar{U}$$

and from equation (4.2)

$$\bar{N}UT = \bar{N}U - \bar{N}UR.$$

Therefore, a means of determining the actual values of the various subcategories has been developed for use in operational stages.

Initial Operational Stages

The initial operational stages simulate the indexing of documents with their inherent errors and placing them in a file which is represented by an array. This operation of indexing is divided into three phases: error determination of the index and search terms, search term designation and formulation of array of indexed documents of interest, and the category designation of the document index terms.

Error Determination. The categories of error and/or non-error in both searching and indexing are determined, based on their probability distribution in conjunction with their position in the sequence of terms used to formulate a search or to index a document. This effort will be demonstrated for both the search and the index terms. Therefore, the probability distribution of search terms will be converted to a frequency distribution for the search terms for the categories of interest.

The number of search terms of each category, YNU, $Y\bar{N}U$ and $YN\bar{U}$ and $Y\bar{N}\bar{U}$, are determined by using the values of YMAX

and the number of actual terms, YU, along with parameters of the binomial distribution of each of the cells of the YMAX terms. The equation (3.3) translated into search terms shows that

$$YMAX = YNU + Y\bar{N}U + YN\bar{U} + Y\bar{N}\bar{U}.$$

A numerical example will be used to demonstrate the procedure for determining the number of search terms of each category.

Given the following values

$$YU = 5$$

$$YMAX = 9, \qquad I = 1,2,---,YMAX.$$

YP(I) is the probability that a particular search term would be needed. YP(I) values are shown on Table 4.3. Using values from the random number generator and comparing this with the value of $\bar{P}$ for each of the cells from cell numbers 1 to 9, the determination is made for each of these 9 cells as to whether or not the terms actually would have been needed in formulating a search. The hits are depicted as having a value of 1 and are equivalent to being needed, and those not needed are depicted with 0's. For example, in cell number 1 the probability of a hit or of the term being needed is 0.95. In this example a random number is generated; if it is less than 0.95 the model assumes or depicts that this particular term would have been hit. Reference to Table 4.4 shows that this term was needed as shown in column s since the outcome category is designated by the presence of

Table 4.3. Search Term Input Values

Cell No. = I	$\bar{P}$ = YP(I)
1	.95
2	.85
3	.65
4	.50
5	.40
6	.30
7	.20
8	.10
9 = YMAX	.05
	4.00 = E(YN)

a 1. This procedure followed through with the other cells in this particular example. Cell no. 2 has a probability 0.85 of being needed. In this particular example it was shown that the term was not needed and is designated 0. Similarly, with cell no. 3 and cell no. 4, and they are shown as hits. Cell no. 5 was not a hit. Cells 6, 7, and 8 were hits, and 9 was not a hit. Summation of all the hits

Table 4.4. Search Term Output

Cell No. = I	Outcome Category	Total No. of Terms Needed YN = YNU + YN$\overline{U}$	Search Term Category and Designation for Simulation Purposes
1	1	1	YNU = 9
2	0	1	Y$\overline{N}$UR = 5
3	1	2	YNU = 9
4	1	3	YNU = 9
5=YU	0	3=YNU	Y$\overline{N}$UR = 5
6	1	4	YN$\overline{U}$R = 3
7	1	5	YN$\overline{U}$R = 3
8	1	6	YN$\overline{U}$S = 4
9=YMAX	0	6=YN	Y$\overline{NU}$ = 0

shown in column 3 produces a total of 6 hits, which equals the number of search terms needed, YN. Therefore, the number of terms not needed is 3, which is the remainder.

The second stage consists of relating the number of terms used to the total, YMAX. As noted, column 3 is entitled

$$YN = YNU + YN\overline{U}.$$

Reference to equation (3.1) and (3.2) shows that YNU is a subset of both the needed and the used set. Therefore, of the terms used, those that were hits are in the YNU category and obtain a 9 designation. The terms that were used and not needed are in the YNU category, have a 0 designation, and are located in cells 2 and 5. It will be subsequently determined for these terms whether they are in categories 5 or 6. Going on to the terms that were not used, that is, cells 6, 7, 8, and 9, one notes that cells 6, 7, and 8 have hits or a 1 designation, which implies they are needed. Cell 9 has a 0 designation which indicates it was not needed nor used. It is noted that the number of needed terms exceeded the number of used terms in this example.

Based on previous work, equation (4.4) is applicable. Since,

$$YN > YU$$

$$(6 > 5),$$

and from equation (3.1)

$$YN\bar{U} = YN - YNU$$

$$YN\bar{U} = 6 - 3$$

$$YN\bar{U} = 3$$

and from equation (3.2)

$$Y\bar{N}U = YU - YNU$$

$$Y\bar{N}U = 5 - 3$$

$$Y\bar{N}U = 2.$$

Therefore, from equation (4.5)

$$Y\bar{N}UR = Y\bar{N}U$$

$$Y\bar{N}UR = 2$$

and from equation (4.3)

$$YN\bar{U}R = Y\bar{N}UR = 2$$

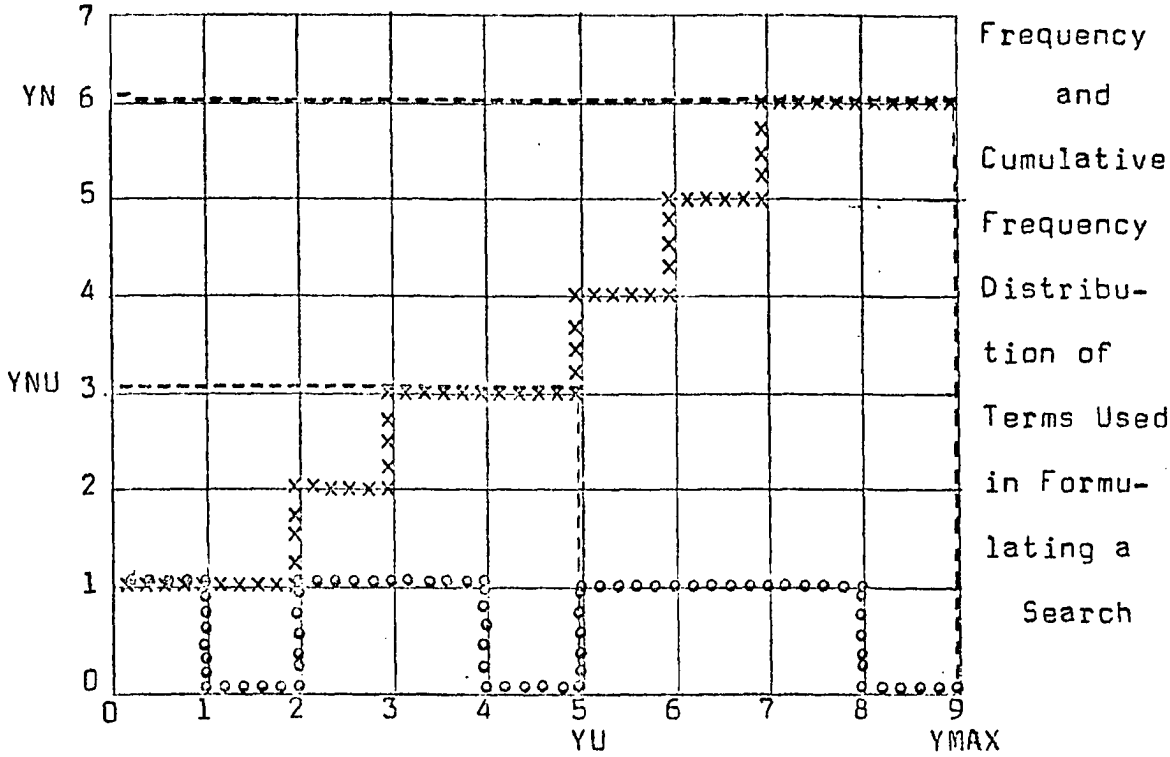
and from equation (4.4) and (4.6)

$$YN\bar{U}S = YN\bar{U} - YN\bar{U}R$$

$$YN\bar{U}S = 3 - 2$$

$$YN\bar{U}S = 1.$$

Therefore, there are three terms in the NU category, two terms in the $N\bar{U}R$ category (3), two terms in the $\bar{N}UR$ category (5), one term in the $N\bar{U}S$ category (4), and one term in the $\bar{N}\bar{U}$ category. These results are shown in column 4, "Search Term Category and Designation for Simulation Purposes," and is depicted by array SEARCH in the program.



Position (cell) of Each Term Used in Formulating a Search

oo = hits
xx = YNU + YNŪ

Fig. 4.2,---Search Term Category and Frequency Distribution

Therefore, a perfect search is designated

$$YN \approx (1,3,4,6,7,8), \text{ and}$$

an actual search is

$$YU \approx (1,2,3,4,5).$$

The above data and the data from Table 4.4 are plotted in Figure 4.2.

The mechanics of indexing for simulation purposes have two aspects: (1) formulation of the array of index terms of interest along with the appropriate document number identification for each term and (2) determination of the number of terms in each category and in which cells they are located. This determination must be made for each formulated search.

The steps involved in the latter of ascertaining the categories of index terms and the number of terms in each category are shown below. The former aspect will be covered in a later section.

Given:

XMAX = 15

XU = 10

and the remainder of the data are given in the following Tables 4.5 and 4.6.

Determination of the number of terms of each category in the simulation is the same process as shown under searching.

Table 4.5. Index Term Input and Output Values

Cell No. = I	$\bar{P}$=XP(I)	Outcome Category	Total No. of Terms Needed XN = XNU+XN$\bar{U}$	Index Term Category and Designation for Simulation Purposes
1	.99	1	1	XNU = 9
2	.95	1	2	XNU = 9
3	.90	0	2	X$\bar{N}$UR = 5
4	.85	1	3	XNU = 9
5	.80	1	4	XNU = 9
6	.75	1	5	XNU = 9
7	.70	0	5	X$\bar{N}$UR = 5
8	.65	1	6	XNU = 9
9	.50	0	6	X$\bar{N}$UT = 6
10=XU	.35	1	7	XNU = 9
11	.25	0	7	X$\bar{N}$UT = 0
12	.15	1	8	XN$\bar{U}$R = 3
13	.10	0	8	X$\overline{NU}$ = 0
14	.05	1	9	XN$\bar{U}$ = 3
15=XMAX	.01	0	9=XN	X$\overline{NU}$ = 0
	8.00=E(XN)			

The cells of the terms for each category and subcategory of index terms are as follows:

Perfect indexing, XN $\approx$ (1,2,4,5,6,8,10,12,15) and

Actual indexing, XU $\approx$ (1,2,3,4,5,6,7,8,9,10).

Table 4.6. Index Term Categories and Cells

Name of Index Term Category	Index Term Category for Simulation Purposes	Cell No. of Index Terms
XNU	9	= 1,2,4,5,6,8,10
$XN\bar{U}R$	3	= 12,15
$XN\bar{U}S$	4	= nonexistent
$X\bar{N}UR$	5	= 3,7
$X\bar{N}UT$	6	= 9
$XN\bar{U}$	0	= 11,13,14

. . Note cells 11, 13, and 14, which are in category 0, have no further use.

A 15-position array with a numerical value depicting each of the index terms has now been prepared. This array is the population which will be randomly sampled to ascertain the category of the terms used in indexing and is depicted as the array INDEX in the program.

Designation of Search Term Rank. The file of indexed documents in the model is limited to those documents that will be investigated in a given search. Therefore, the search must be formulated into the appropriate terms, which are then mapped into the index terms of interest.

The rank of each search term is determined by using a random number generator, where the probability distribution of the terms is geometric as shown in equation (2.1).

$$g_1(k) = \frac{(1-BE)BE^{k-1}}{(1-BEU^{KE})}.$$

Since the frequency of error terms is a ratio of the terms actually used, the probability of obtaining the maximum number of terms, YMAX, can be shown as having the same probability.

$$g_2(k) = \frac{(1-BEU)BEU^{k-1}}{(1-BEU^{KE})} = g_1(k).$$

This function is transposed and solved for k and expressed in the program and $g_2(k)$ is replaced by XY.

In the format of program terminology, this function is expressed

$$NRANK(I) = k = \frac{\ln\left(\frac{(1-BEU^{KE})\ XY}{(1-BEU)}\right)}{\ln\ BEU} + 1$$

I = 1,2,---,YMAX,

NRANK(I) = rank of a search term,

where XY = random number from a linear random number generator.

Also, the ranks must be determined so that

$$k_1 \neq k_2 \neq,---,K_E.$$

The ranks of the search terms are now generated for the same relative ratio for any particular search term. Therefore, it can be readily shown that this function can be expressed for use in the model, using equation (3.6),

$$g_2(j) = \frac{(1-BIU)BIU^{k-1}}{(1-BIU^{JI})} XMAX$$

for all terms used in indexing = XMAX.

Let this be expressed in the program as

$$STRIKE = \frac{(1-BIU)(BIU^{NRANK-1}}{(1-BIU^{JI})} JXMAX.$$

Since the rank of the terms of interest has been generated in the search, the documents that are indexed by these terms are now determined. A uniformly distributed random number generator is then used on a 0,1 basis to determine if the particular document was indexed with terms in question. Because the number of documents in the entire file is large, it is desirable to sample this population. The sample size desired is designated in the input to the program. Storage and computation limitations of the computer dictate that in turn this sample size be processed in a series of segments

of 100 documents. The data are aggregated by sample size before being compiled for intermediate output determination. This action results in a file where the documents indexed by the terms of interest are depicted by 1's and the non-indexed documents are shown as 0's, as shown in Table 4.7, which is represented by the array "BLOCK" in the program.

Since the location of the strikes is random and the volume of material being generated is large, a reduction of the real data to the useful set is desirable. Output specifications indicate that all the output has a common search input subset of the needed used (9) category. Since searching is the first step to investigate, it is apparent there must be index term strikes for each of the search terms of category 9. These search terms are located in the first three columns of Table 4.7. For example, documents 6, 8, and 11 would not be investigated and would be removed from the array as shown in Table 4.8.

Category Designation of Document Index Terms. The determination of which term has been used to index each document has been made for the search terms of interest and is shown in Table 4.8. The procedure for designating categories of search terms will be described. Document 1 in Table 4.8 is shown to have been indexed with six terms out of a possible eight search terms of interest and are shown in columns 1, 2, 3, 5, 6, and 7. Randomly picking one of the fifteen terms from the index terms category array shows that the first term

Table 4.7. Document Identification Number Versus Search Terms

ID = Document Identification Number	Search Term Categories							
	YNU = 9			YNŪ = 7			YN̄U = 11	
	Search Term Rank			YNUR = 3		YNUS = 4	YNUR = 5	
	1075	935	452	851	600	540	1340	470 Search Term Rank
15	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1
12	1	1	1	1	0	1	1	1
11	0	1	0	1	1	1	1	0
10	1	1	1	1	0	0	1	1
9	1	1	1	0	1	1	1	1
8	0	1	1	0	1	1	1	0
7	1	1	1	1	1	1	1	1
6	1	0	0	1	1	0	1	1
5	1	1	1	1	1	0	1	1
4	1	1	1	1	1	1	1	0
3	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1	0

Documents identified are not printed out or included in array, "BLOCK".

Table 4.8. Document Index Term Categories

ID = Document Identification Number	Search Terms of Interest								
	1	2	3	4	5	6	7	7	Cell Numbers
15	9	9	9	9	9	9	9	9	
14	9	9	3	9	3	9	9	3	
13	9	6	9	9	9	6	9	5	
12	3	9		9		9	5	3	
10	5	5	9	3			6	9	
9	9	5	5		5	6	9	9	
7	9	3	9	9	9	3	9	3	
5	9	9	9	9	9		9	3	
4	9	5	6	6	9	9	6		
3	9	9	9	9	3	9	9	5	
2	9	9	5	9	5	9	3	9	
1	9	5	0		9	9	9		
Index Term Category	9,9,9	9,9,9	9,	3,	3,	5,5,6	0,	0,0	

is in category 9, the second term is in category 5, the third term is in category 0, and on across the columns. Columns 4 and 8 would not be assigned any categories of index terms since they start with a designation of 0, indicating nonindexing of this document by the terms represented by those columns.

The results of these operations of identifying the categories of index terms of interest are shown in Table 4.9, which is a representation of array "BLOCK" in the program.

Output Preparation

The preparation of the output consists of the index-search term interaction, the formulation of intermediate output, and the preparation of the final output into its prescribed format.

Interaction. The interaction consists of expressing the internal operation of matching formulated searches versus the file of indexed documents. A graphical representation of a file of indexed documents is shown in Table 4.9.

Given the chosen output combination of index and search terms, it is now necessary to form a new series of arrays, one for each combination of index and search terms. Only those output that correspond to the search terms of the prescribed combinations are to be included in an array of indexed documents for each run of the simulation. In addition, only the document identification number need be presented under each search.

For example, if the documents listed under the desired

TABLE 4.9. Documents Indexed Under Search Terms

Document Identification Number	Search Term Categories								
	YNU = 9			YNU = 7			YNU = 11		
				YNUR = 3		YNUS =4	YNUR = 5		
	1	2	3	4	5	6	7	8	Cell No.
15	9	9	9	9	9	9	9	9	
14	9	9	3	9	3	9	9	3	
13	9	6	9	9	9	6	9	5	
12	3	9		9		9	5	3	
10	5	5	9	3			6	9	
9	9	5	5		5	6	9	9	
7	9	3	9	9	9	3	9	3	
5	9	9	9	9	9		9	3	
4	9	5	6	6	9	9	6		
3	9	9	9	9	3	9	9	5	
2	9	9	5	9	5	9	3	9	
1	9	5			9	9	9		

Index Term Categories

index and search categories are to be recalled by using the following input categories,

Index categories XN = 9,3,4, and

Search categories YN = 9,3,4.

Referring to Table 4.9 it is shown that references to documents listed under search terms in cells 1, 2, 3, 4, 5, of search term categories 9, 3, and 4, would be used with the index categories 9, 3, and 4. The rest of the terms and documents would be omitted. References to documents 3, 7, 14, and 15 would be included.

Each search term must be searched to verify the presence of the document identification number listed under it. Inspecting the file will ascertain which documents are indexed with all the prescribed terms and only those documents that are indexed under all terms are retrieved. The results of the operation are expressed in the simulation by a listing of the document numbers as shown in Table 4.10, which is a partial representation of array, REF.

A review of the actual situation of indexing and searching is depicted by comparing the output in REF(4,4) with that in REF(1,1). The ideal situation, REF(1,1), shows that, under perfect indexing and searching, document no.'s 3, 7, 14, and 15 would be retrieved, ZNU = 4, as shown in Table 4.11, ZNU (1,1,1) = 4. Since there is no error in the output,

$\overline{ZNU}(1,1,2) = 0$ and

$\overline{ZNU}(1,1,3) = 0$.

Table 4.10. Document Identification No's Recovered Vs. Input Categories

Index \ Search		Needed Terms Correct Number	Needed Terms, Inadequate in Number	Used Terms, $\leq$ Number of Needed Terms	Used Terms $\geq$ Number of Needed Terms
		YN = YNU + YN$\bar{U}$ 9 + 4 + 3	YNU + YN$\bar{U}$R 9 + 3	YNU + Y$\bar{N}$UR 9 + 5	YU = YNU + Y$\bar{N}$U 9 + 5 + 6
Needed Terms Correct Number	XN = XNU+XN$\bar{U}$ 9+4,3	15 14 7 3 **1,1**	15 14 7 5 3 **1,2**	15 14 7 5 **1,3**	15 14 7 5 **1,4**
Needed Terms Inadequate in Number	XNU+XN$\bar{U}$R 9+3	15 14 7 3 **2,1**	15 14 7 5 3 **2,2**	15 14 7 5 **2,3**	15 14 7 5 **2,4**
Used Terms $\leq$ Number of Needed Terms	XNU+X$\bar{N}$UR 9+5	15 2 **3,1**	15 5 2 **3,2**	15 9 3 **3,3**	15 9 3 **3,4**
Used Terms $\geq$ Number of Needed Terms	XU = XNU+X$\bar{N}$U 9+5,6	15 13 4 2 **4,1**	15 13 5 4 2 **4,2**	15 13 10 9 3 **4,3**	15 13 10 9 3 **4,4**

TABLE 4.11. Retrieved References to Number of Documents of Each Category

Index Term Categories		Search Term Categories			
		YN = YNU+YN$\bar{U}$ 9 + 3, 4	YNU + YN$\bar{U}$R 9 + 3	YNU + Y$\bar{N}$UR 9 + 5	YU = YNU + Y$\bar{N}$U 9 + 5, 6
	XN = XNU + XN$\bar{U}$ 9 + 3, 4	ZNU(1,1,1) = 4 ZN$\bar{U}$(1,1,2) = 0 Z$\bar{N}$U(1,1,3) = 0	ZNU(1,2,1) = 4 ZN$\bar{U}$(1,2,2) = 0 Z$\bar{N}$U(1,2,3) = 1	ZNU(1,3,1) = 3 ZN$\bar{U}$(1,3,2) = 1 Z$\bar{N}$U(1,3,3) = 1	ZNU(1,4,1) = 3 ZN$\bar{U}$(1,4,2) = 1 Z$\bar{N}$U(1,4,3) = 1
	XNU + XN$\bar{U}$R 9 + 3	ZNU(2,1,1) = 4 ZN$\bar{U}$(2,1,2) = 0 Z$\bar{N}$U(2,1,3) = 0	ZNU(2,2,1) = 4 ZN$\bar{U}$(2,2,2) = 0 Z$\bar{N}$U(2,2,3) = 1	ZNU(2,3,1) = 3 ZN$\bar{U}$(2,3,2) = 1 Z$\bar{N}$U(2,3,3) = 1	ZNU(2,4,1) = 3 ZN$\bar{U}$(2,4,2) = 1 Z$\bar{N}$U(2,4,3) = 1
	XNU + X$\bar{N}$UR 9 + 5	ZNU(3,1,1) = 1 ZN$\bar{U}$(3,1,2) = 3 Z$\bar{N}$U(3,1,3) = 1	ZNU(3,2,1) = 1 ZN$\bar{U}$(3,2,2) = 3 Z$\bar{N}$U(3,2,3) = 2	ZNU(3,3,1) = 2 ZN$\bar{U}$(3,3,2) = 2 Z$\bar{N}$U(3,3,3) = 1	ZNU(3,4,1) = 2 ZN$\bar{U}$(3,4,2) = 2 Z$\bar{N}$U(3,4,3) = 1
	XU = XNU + X$\bar{N}$U 9 + 5, 6	ZNU(4,1,1) = 1 ZN$\bar{U}$(4,1,2) = 3 Z$\bar{N}$U(4,1,3) = 3	ZNU(4,2,1) = 1 ZN$\bar{U}$(4,2,2) = 3 Z$\bar{N}$U(4,2,3) = 4	ZNU(4,3,1) = 2 ZN$\bar{U}$(4,3,2) = 2 Z$\bar{N}$U(4,3,3) = 3	ZNU(4,4,1) = 2 ZN$\bar{U}$(4,4,2) = 2 Z$\bar{N}$U(4,4,3) = 3

Intermediate Output. Following completion of the construction of the arrays of document numbers, the presence or absence of specific document numbers that correspond to various categories of index and search terms must be tested. This testing would then provide the document numbers that are the intermediate output for each type of input. The output corresponding to each of the specified combinations of inputs is determined by recording the document identification numbers for each combination of inputs. The effectiveness of each of these combinations of inputs in retrieving the desired references to documents is achieved by comparing the list of document identification numbers of each combination of inputs against those in the ideal combination, REF(1,1,1). Therefore, REF(1,1,1) is the desired state of perfect indexing and searching (no omission or commission error). This phase also includes comparison with the actual combination of terms used, REF(4,4,1), Table 4.11.

Comparing the documents listed in REF(4,4,1) with those in REF(1,1,1) shows that of the documents that should be recovered, 3, 7, 14, and 15, documents 3 and 15 are recovered. Therefore, ZNU(4,4,1) = 2. Since these are two needed documents not recovered,

$$Z\bar{N}\bar{U} = REF(4,4,2) = 2.$$

Of the five documents recovered, two were needed, three were not needed, and

$$Z\bar{N}U = REF(4,4,3) = 3.$$

Final Output Preparation. The preparation of the final output consists of a physical and a mathematical phase. The physical phase relates to the operations of collecting, aggregating, calculating various categories of data, retrieving this phase and printing it in the desired form. The mathematical phase necessitates defining the types of output desired, then formulating means for its determination.

The desired output consists of the three categories of the end product of the simulation model, that is the ZNU, $ZN\bar{U}$, and $Z\bar{N}U$. In addition, because this is a simulation procedure, the model is interacted a number of times. Also some measures of statistical variation are desired, and the standard deviation of category of output is determined. Therefore, the intermediate output, as illustrated in Table 4.11, shows three categories of output for sixteen combinations of index and search terms. To facilitate calculations, accumulators for the linear and square values are set up on arrays OUTSMF and OUTSQF in a manner that can be depicted similar to that in Table 4.11.

At the termination of the simulation, the appropriate values of the mean and standard deviation are calculated, where,

$$\text{Mean} = \frac{\text{OUTSMF}}{\text{No. of runs of simulation}}$$

$$\text{Std Dev.} = \sqrt{\frac{(\text{OUTSQF}) - \frac{(\text{OUTSMF})^2}{\text{ISRT}}}{\text{ISRT} - 1}}$$

where, ISRT = No. of times simulation has been run.

OUTSMF Three-dimensional array with accumulators for the linear values of the number of references to documents for each of the three categories of output obtained for each of the sixteen combinations of index and search terms defined.

OUTSQF Three-dimensional array with accumulators for the corresponding squared values of OUTSMF.

Therefore, the comparison of the list of document numbers determines a lack of existence of these numbers, which is the same as the presence of the categories of output, recalled-relevant, nonrecalled-relevant, and recalled-nonrelevant. Having these categories of output for the various combinations in inp ts provides a means of relating errors in output to errors in input. These data are placed in the array similar to Table 4.11 for the types of inputs expressed containing both the mean and standard deviation. Therefore, there is a means of identifying the output, both that desired and undesired, with the inputs of indexing and searching. Although output is a function of both indexing and searching, the model is developed so that the effect of errors in both inputs can be distinguished as to whether they are independently identifiable with each input or there is an interaction effect of both inputs that produces the errors.

The total population of documents was sampled to

determine the relative number of references to documents of each category. Calculation of the final output necessitates relating these sample values to the population values. The total number of documents in the file of index documents TOTDOC, when a sample size, DOCMAX, is taken. Therefore, the calculated values of output which is the number of references to documents must be increased by the ratio of the size of the population to size of the sample.

CHAPTER V

REFERENCE RETRIEVAL SYSTEM APPLICATION

A three-stage procedure for simulating and evaluating the level of operation of a reference retrieval system, given certain fixed factors and relevant decision variables, has been developed. This section will demonstrate the feasibility of application of the model and its evaluation.

Objective

The objective of this section is to relate the quantity of output of citations to documents to the level of input of index and search terms. Inferences about the functional form of the relationship can then be made. The results of this stage will subsequently be used to optimize the entire retrieval system based on cost of operating and value of benefits obtained. The following four conditions must be met when one is hypothesizing the form of the surface to be generated.

1. The feasible region must be defined.
2. The regression model generated must be statistically significant.
3. The model must be economically feasible, i.e., all positive output.
4. Conditions for economic optimization must be considered, i.e., second derivatives.

However, prior to formulating an experimental design to test the hypothesis that the performance surface has a

particular form, some preliminary analysis of the basic input data and the stability of the performance model seems in order. Also the effect on output of the indexing and searching terms independently is investigated.

Model Stability

Stability of the performance model will be investigated by allowing two factors of the system to vary independently over a range of values and measuring the variation of the indicated level of output. The variables investigated are the number of times the data are iterated or looped through the model and variation of the sample size of the number of documents queried out of the total population of indexed documents in the collection.

Input Errors

The evaluation of errors in input data (that of the inability of index and search terms, as they are used, to represent concepts) is explored. Previous work has shown decreasing returns to scale which approach zero and become asymptotic at the value of XMAX, YMAX for the index and search terms respectively. The general form of relating the number of needed, used index terms, XNU, versus the total number of index terms used, XU, will be expanded, as will the relationship of YNU to YU.

Output Dependence on Inputs

Investigation of the effect of the level of indexing and searching is performed to aid in formulating the performance output surface as a function of the number of index and search terms and their associated errors. Experiments are conducted at various levels of XU and YU, and the results are evaluated. Specifically the form investigated is taken from equation (3.5),

$$ZU = f(XU, YU).$$

The particular form of the output to be investigated is the desired output, ZNU, where

$$ZNU = f_1(XU,YU).$$

Fixing each of the variables yields the following functions

$$ZNU = f_2(XU/YU),$$

and

$$ZNU = f_3(YU/XU).$$

An attempt to evaluate the level of the desired output on level of indexing, XU, with a measure of the error in indexing, XNU, will be investigated. It is recognized that a dependency exists between XNU and XU. The form postulated is

$$ZNU = f(XU,XNU/YU,YNU). \qquad (5.1)$$

Similarly, for searching

$$ZNU = f(YU,YNU/XU,XNU). \qquad (5.2)$$

The values investigated are shown on Figure 5.1. A small-scale version of the reference retrieval system being

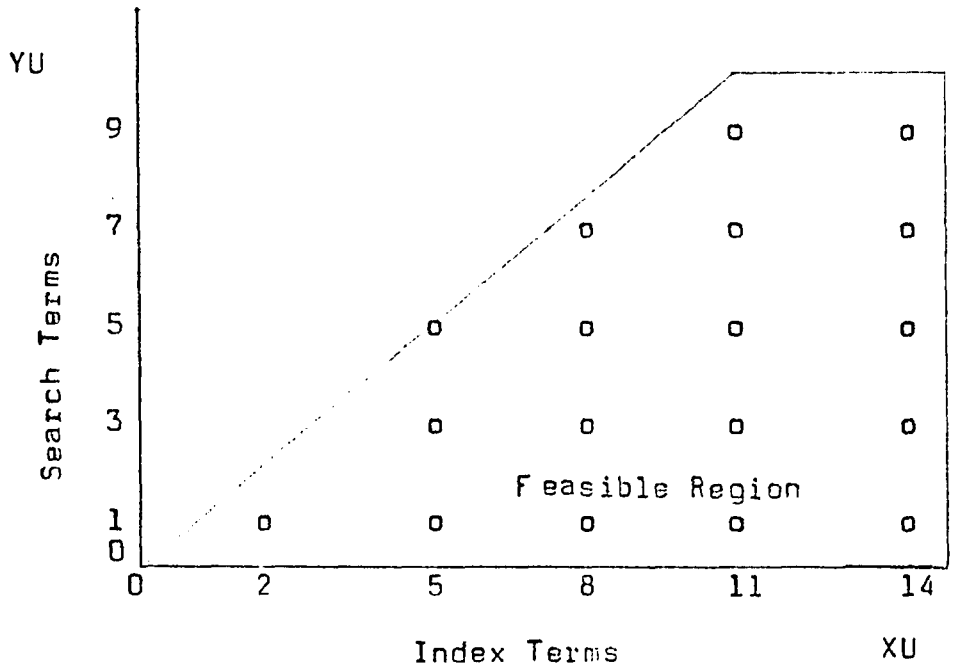


Fig. 5.1,---Feasible Region

simulated will be used in the interest of economy of operation.

Since intersection searches only are being investigated, the feasible region includes the boundaries of XU and YU, along with the area included with the limits of

$$XU \geq YU,$$

and

$$1 \leq XU \leq 15$$

$$1 \leq YU \leq 9.$$

Although the detail of boundary conditions will not be investigated in this project, they will be examined so that

the influence on future work is foreseen.

In an attempt to quantify the effect of error in input on output, other forms of expressing error in the inputs were investigated. A specific form reviewed was

$$P_n = k - n(\theta)$$

where P_n = probability of being needed,

k = constant

θ = variable

n = number of terms used.

$$1 \leq n \leq XMAX.$$

Let

$$n = XMAX$$

and

$$P = 0$$

so,

$$0 = k - XMAX(\theta)$$

or

$$\theta = \frac{k}{XMAX},$$

let

$$k = 1$$

$$\theta = \frac{1}{XMAX}.$$

Since $$XNU = \sum_{n=1}^{XU} P_n \cdot 1$$

$$XNU = \sum_{n=1}^{XU} (k-n(\theta))$$

$$XNU = XU - \frac{n(n+1)}{2}\left(\frac{1}{XMAX}\right).$$

This straight-line approach for showing the variation in the amount of error in the basic inputs will be used in addition to the OC curve type of error previously discussed.

The relationship between output from the model to the level of index and search terms discussed above is needed to investigate the interactive effects of both indexing and searching for total system optimization. Optimizing output necessitates defining the quantified relationship between the inputs with their associated error and the quantity of output over the feasible operation range. The exploratory work previously discussed will indicate specific directions for the specific form of the relationship to be generated.

Experimental Design for Performance Surface

Exploratory work to determine the operating characteristics of the performance model are followed by a designed experiment to ascertain the feasibility of using the performance model within the general goal of the objectives of the system application.

Application of the performance model has two objectives: that of relating the output to the input and its errors and that of measuring the output for the purpose of relating value to the cost of the system. Therefore, the experiment must be conducted to achieve these ends. Generation of a surface representing output as a function of the levels of inputs XU and YU and their associated errors is structured in terms of the objective and the specific hypothesis to be tested.

Objective. The specific objectives of the experiment are (1) to relate the output of the number of desired

citations to documents in the inputs and (2) to determine the error categories of output as a function of the number of index and search terms. Specifically it is desired to define the output as

$$ZNU = f(\ln XU, \ln YU). \qquad (5.3)$$

In general ZNU will be correlated positively with XU and inversely proportional to YU.

In addition there will be a decreasing marginal effect as XU or YU independently increase. Also the ZNU output must be defined as a function of the same variables.

Regression models for performance-model operation are developed from available data.

Hypothesis. The hypothesis to be tested is whether output is a function of the inputs as represented. Therefore the coefficient of the multiple determination will be tested. This procedure will test whether the partial regression coefficients are equal to zero. This test is accomplished by making an F test, following use of the regression effort to determine the coefficients of the proposed model.

Work on a sample size determination is not readily applicable to this type of expression. Previous experimental work indicates that an orderly and evenly spaced series of points to be evaluated is desirable. Therefore, it is proposed to evaluate index terms, XU, at intervals of three and the search terms, YU, at intervals of three terms. In addition, the feasible region of the system exists at values

where

$$XU \geq YU.$$

The desirable number of replications at each level of term usage is indicated to be at least 5. Although a larger number of replications may be desirable, costs of running the simulation provide a realistic constraint for using a higher number.

The specific form hypothesized is from equation (5.3).

$$ZNU = f(\ln XU, \ln YU).$$

In addition, the quantity relationships for the undesired output is desired. The proposed function for relating the $Z\bar{N}U$ citations will be formulated as

$$Z\bar{N}U = f(XU, \ln XU, YU, \ln YU). \qquad (5.4)$$

Data for this experiment will be the same as that generated for the previous experiment.

Data Sources

The general scope of the system being analyzed is based on data available from several sources. The interest of time and resource limitation necessitate using assumptions for the determination of several of the constants and variables necessary for implementation of the model.

Error Determination

The lack of any significant usable data in the literature

precludes any attempt at implementation of the errors in index and search terms. Therefore, output data for the error-determination stage of the model are assumed and are formulated so that they are consistent with their intended use in the performance model.

These data are generated, and the form of the discrete values is similar to an operating characteristic curve, S shaped, where the ideal situation would be represented by a rectangular distribution.

Performance Model

The performance model uses several parameters that were obtained from MEDLARS; and other sources, which again in the interest of limited resources, caused several parameters to be estimated.

Table 5.1 shows the values used and their source. In addition some considerations inherent in the model preclude direct comparison of the level of output of the performance model with the level of output with MEDLARS. The application test of the performance model uses independent determinations of index and search terms, whereas MEDLARS have a series of categories of terms. All terms in MEDLARS cannot be used independently because there is a series of mutually exclusive-situations in the application of these terms. In addition the searches formulated in the application test are intersection searches. MEDLARS searches can be formulated

Table 5.1. Numerical Values for Parameters and Factors of the System

Exogeneous Constants	Definition	Value	Sources Used for Derivation
D	Number of indexed documents in the reference file	700,000	MEDLARS
S	Number of user queries to be formulated annually	10,000	Estimate
I	Number of terms in the indexing vocabulary	7,000	MEDLARS
E	Number of terms in the searching vocabulary	7,000	MEDLARS
T	Number of searching installations	1	Estimate
R	Number of new documents indexed annually (replacements)	175,000	Costello MEDLARS
A	Number of search files reproduced annually per installation as related to the number of new documents indexed annually.	2	Estimate

Endogeneous Constants

XMAX	Total number of terms needed to be used in indexing a document to avoid any omission error	15	Review of literature
YMAX	Total number of terms needed to be used in formulating a search to avoid any omission error	9	Review of literature
XN	Number of terms needed to index a documsnt, error-free	8.0	Estimate
YN	Number of terms needed to formulate a search, error-free	4.75	Estimate

XP(I) = .99,.95,.90,.85,.80,.75,.70,.65,.50,.35,.25,
.15,.10,.05,.01.

I = 1,2,---,15.

YP(I) = .99,.90,.80,.70,.60,.40,.25,.10,.01.

I = 1,2,---,9.

by using either intersection or union of terms, along with various combinations of both the above. In addition, the searches can include negations.

Results of Experiments

The results of the experiments conducted relate to performance model stability, relationships of system variables, and finally the functional form of outputs as they are dependent on the independent variables.

Model Stability

The evaluation of the stability of the performance model can be based on the number of times the file of documents is searched or the number of simulation runs and the effect of the size of sample of indexed documents analyzed.

Number of Simulation Runs. The number of times to run the simulation are evaluated by running the system utilizing a small-scale version of inputs to minimize run time, at a given level of indexing and searching. The values of the parameters and variables at which the simulation was conducted are as shown in Table 5.2.

The number of references to be recovered under perfect searching and indexing is a constant for each search, however, there is a variation between searches. Therefore, the objective was to determine if changes in the number of times the file was searched caused any variation in the number of references recovered.

Table 5.2. Experimental Data

No. of indexed documents	TOTDOC	700
Sample size of documents	DOCMAX	100
No. of times to search the file of documents (iterate the model)	ISRT	= 20,40,60,80,100
No. of index, search vocabulary terms	JI = KE	70
No. of search terms used	YMAX YU	9 4
No. of index terms used	XMAX XU	15 8
No. of independent replications		5

The simulation was run using eight terms in indexing and four terms in searching. Then a sample was created which consisted of 100 in total population of 700 documents. A total of 20 searches was conducted and the output obtained: ZNU, ZN$\bar{U}$, and ZN. The number of references to citations, ZNU, ZN$\bar{U}$, and ZN, were determined and expressed in averages based on the individual runs which were then extrapolated to the total population of 700 documents.

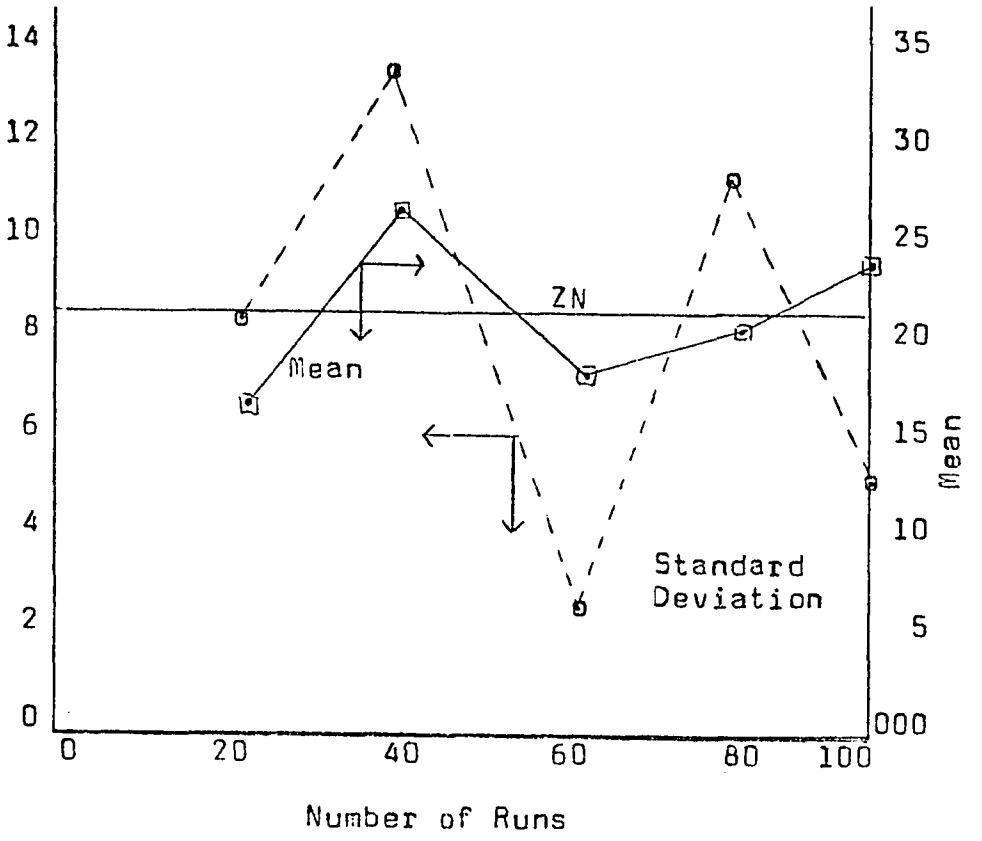
A case consisting of 5 separate determinations of application, using a sample size of 100 documents, and consisting of 20 separate searches of the file of documents (iterating the model) was made. The composite means for these 5 separate replications, which ranged from 9.8 to

28.35, was calculated to be 17.36 and is plotted, along with values obtained using an increasing number of searches, in Figure 5.2a. Subsequently, four additional cases each consisting of a series of 5 replications having a sample size of 100 documents were made with an increasing number of searches; 40, 60, 80, and 100. The values of the composite means, which is an estimator of the parameter ZN, was calculated for each of the remaining 4 cases of runs and plotted in Figure 5.2a. As shown, ZN for each of the 5 cases ranges from a value of 17.36 references to 26.74, which yield an aggregate mean of 21.28. The number of searches did not significantly change this value. Statistical analysis showed that the null hypothesis for equality of means could not be rejected at a 5 percent significance level. The standard deviation for each of the 5 cases, yielding values of 4.69 to 13.36, was determined and is also plotted in Figure 5.2a.

Sample Size. Using the same data discussed above, a group of runs were made at various sample sizes of the population of indexed documents to ascertain the effect of sample size. Each document was indexed with 8 terms and each search formulated with 4 terms. The number of searches of the file of indexed documents (iterations of the model) was held constant at 40. As with the previous analysis a group consisting of a series of 5 replications for each level of input was made.

Standard Deviation

Fig. 5.2a



Standard Deviation

Fig. 5.2b

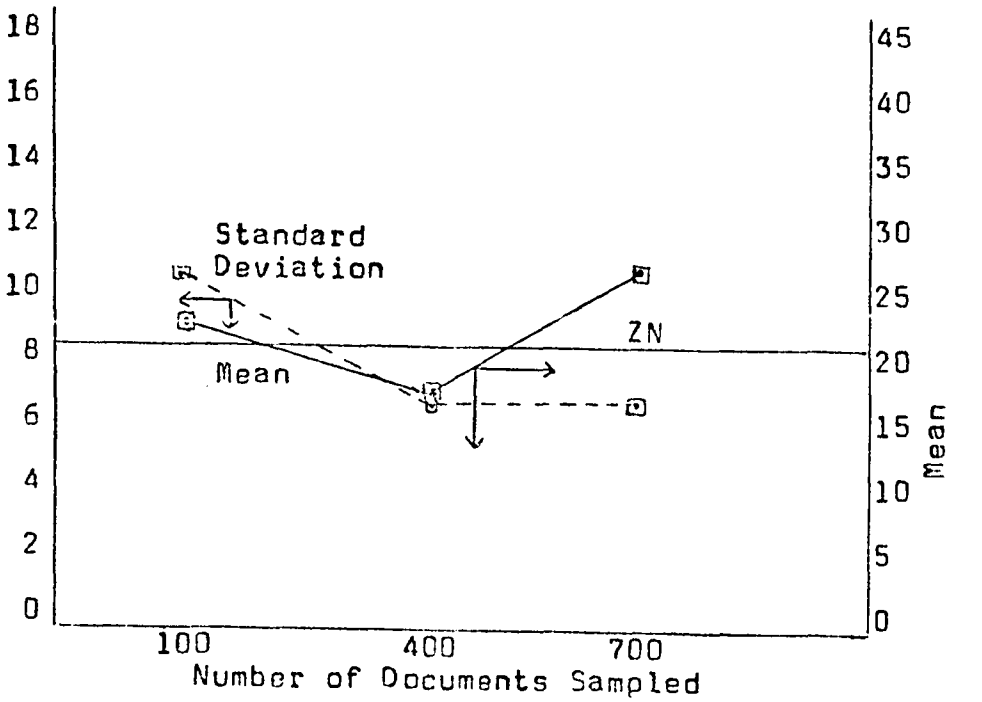


Fig. 5.2,---Mean and Standard Deviation Vs. Variables

For the first group investigated, the entire population of 700 documents was searched 40 times and this action was replicated 5 times. The value of ZN obtained for each of these 5 replications, which ranged from 17.10 to 34.22, yielded a composite mean of 26.12 with the standard deviation of 6.74 and is plotted in Figure 5.2b, along with the results obtained by changing the sample size of the documents investigated. Subsequently 2 additional groups, each consisting of 5 replications, searching the file of indexed documents 40 times were made. The file of indexed documents was reduced to sample sizes of 400 and 100 documents, respectively. This means that document population to sample size was on a 7 to 4 basis when the sample size was 400 and on a 7 to 1 basis when the sample size was 100. The numerical results obtained using these groups of searches were extrapolated to the population of 700 documents. The average number of references to citations, ZN, at the various sample sizes of 100, 400, and 700 documents, ranged from 16.77 to 26.12 with an aggregate mean value of 22.01 as shown in Figure 5.2b. The value of the standard deviation for each of the 3 series of runs ranged from 6.62 to 10.52. Statistical analysis showed that the null hypothesis for equality of means could not be rejected at a 5 percent significance level. Therefore, the variation in the sample size of the number of documents does not seem to affect the value of ZN.

Relationship of Variables

The relationship of variables is defined in the functional relationship of inputs to the amount of error and the relationship of output to inputs and their errors.

Input Errors. The relationship of the number of index terms to their needed aspect was plotted in Figure 5.3a. The curve became asympototic at

$$XNU = 8.00 = XN.$$

As can be seen this is a generally increasing function consisting of two areas. The first part has a constant slope which decreases to a second range of shape, and the curve became asympotic at

$$XU = XMAX.$$

A similar situation with respect to the search terms is shown in Figure 5.3b.

Output Dependence. The level of and the variation of the number of desired references retrieved as it is dependent on the number of index and search terms was investigated as previously described. The most ideal situations were based on the maximum number of data points. The general data are shown in Table 5.3, where three replications were run with 40 iterations of the data, using a sample size of 200 documents.

The general form of the equation, from equation (5.1)

Fig. 5.3a

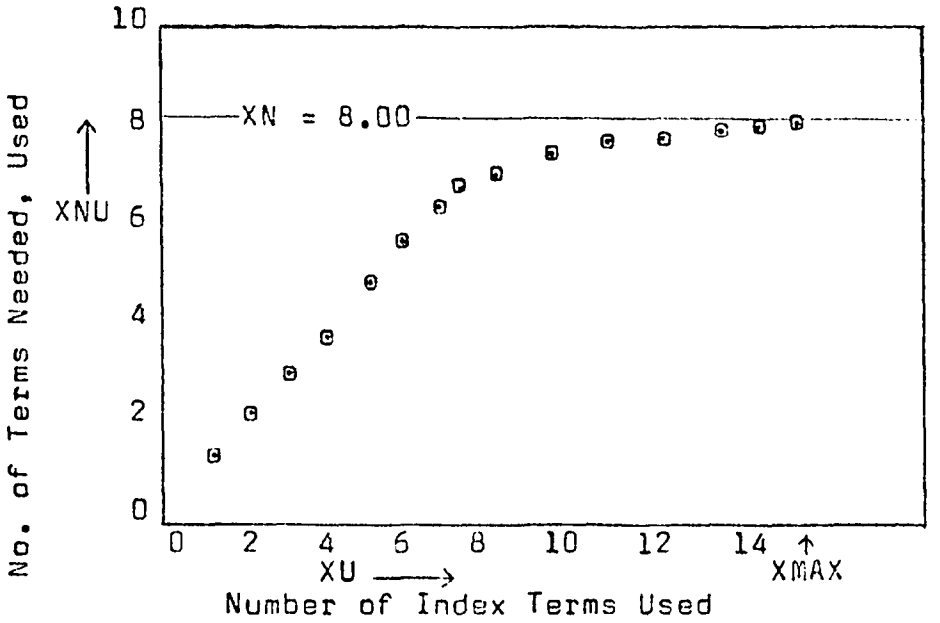


Fig. 5.3b

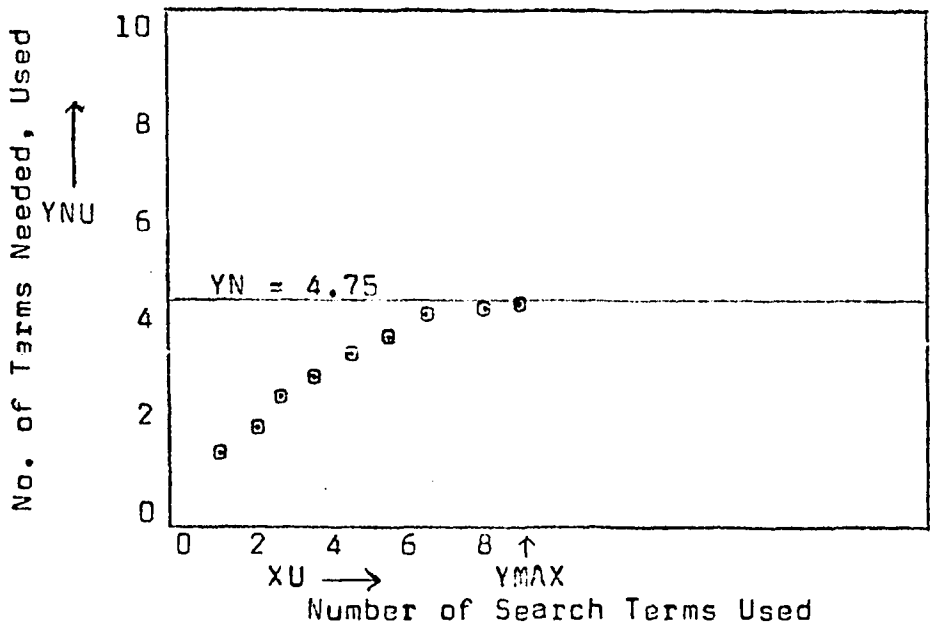


Fig. 5.3,---Number of Terms Vs. Error

Table 5.3. Test Data For Output Dependence

No. of Indexed Documents,	TOTDOC	7000
Sample size of documents,	DOCMAX	200
No. of times to run iterate model,	ISRT	40
No. of index, search vocabulary terms,	JI = KE	70
No. of search terms,	YMAX	9
	YU	1,3,5,7,9
No. of index terms used,	XMAX	15
	XU	2,5,8,11,14
No. of independent runs, replications		3

was

$$ZNU = f(XU, XNU/YU, YNU).$$

Using data from Table 5.3,

$$YU = 9,$$

and the equation generated was

$$ZNU = 21.3769 + 2.283\ XU + 4.0112\ XNU,$$

and

$$R^2 = 0.7084.$$

As the number of index terms decreases (considering the feasible region), the validity of the functions defining the relationships deteriorate. The first problem concerns the reduction in data points and consequently less statistically significant results. The second factor is the

physical relationship in that the upper portion of the curve gets successively truncated as the number of terms is reduced. This is a complex curve and the combined effects of these actions are that, as the number of index terms is reduced the successive results are (1) negative constants, (2) negative coefficients, and, finally, (3) both.

The specific function generated using the maximum number of searches from the general equation, (5.2), is as follows:

$$ZNU = f(YU, YNU/XU, XNU)$$

$$YU = 1,3,5,7,9$$

$$XU = 14.$$

These data generate the specific equation:

$$ZNU = 94.0971 - 10.5702\ YU + 1.0457\ YNU,$$

$$R^2 = 0.6612.$$

It appears that the physical condition that caused the deterioration of the indexing does not hold for the searching terms. Since the results for a similar situation where

$$XU = 11,$$

are

$$ZNU = 93.4609 - 9.9704\ YU - 2.3140\ YNU$$

$$R^2 = 0.8487.$$

It is to be noted that the coefficient of the second variable, that of the number of needed and used search terms, YNU, has changed sign. However, the multiple correlation

coefficient has increased from 0.6612 to 0.8487.

Statistical tests of the coefficients in both equations show that the coefficient of YU is significant, while that of YNU is not significant. However the general form of the relationship does not deteriorate until the statistical constraints of too small a sample size of the number of search terms affect the results. The results of the straight-line error approach versus those of the previously discussed OC curve concept are illustrated in Figure 5.4, where the ratio of error of $Z\bar{N}U$ of SL to OC is plotted versus the number of index and search terms.

As indicated the SL curve approach shows a higher level of retrieved $Z\bar{N}U$ output than does the OC form.

Performance Surface

The performance surface for the two categories of output being quantified were generated by using the performance model to generate the prescribed values at the chosen levels of indexing and searching, then testing the hypothesized relationship of the output to independent variables. The data used are shown in Table 5.1. The levels of indexing chosen were 2, 5, 8, 11, and 14; and the levels of searching were 1, 4, and 7. In addition the effects of boundary conditions was considered by obtaining values at XU = YU: 1, 2, 5, 8, and 9. Thus a total 17 points were determined with 5 replications of 40 iterations at each point. Thus, a total of 85

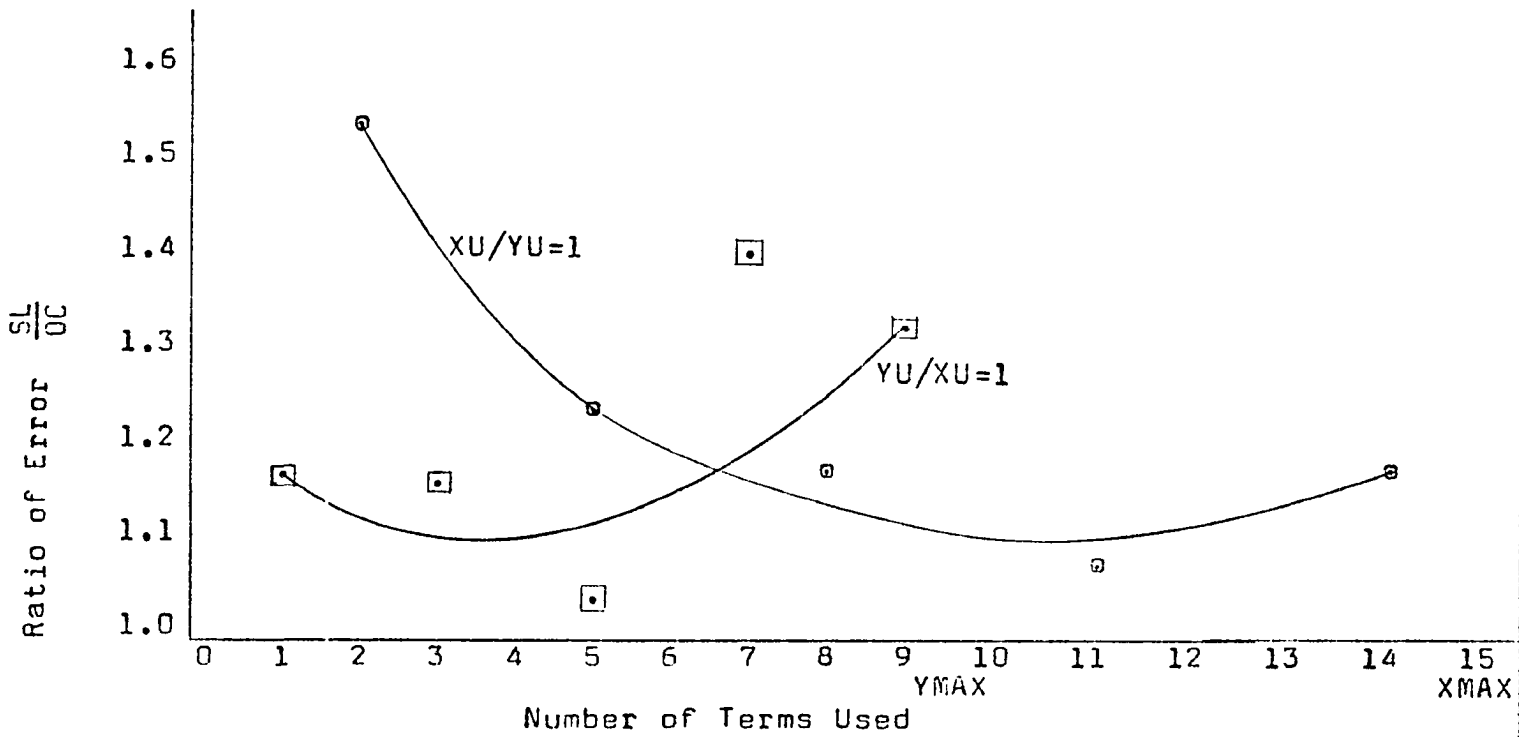


Fig. 5.4,---Error in Output Related to Level of Input ($Z\bar{N}U$)

values of output consisting of 40 queries to the model were utilized. The sample size of the document was 1000, and the results were then expanded to the population value of 700,000 documents in the system. After the requisite number of runs from the performance model were obtained, the assembled data were used in regression models of the form previously specified. Investigation of the feasible region as defined above specifies an inclosed area as shown in Figure 5.5.

Reference to the ZNU output from the simulation model indicates that some of the points within the theoretically feasible region show that two or more values obtained are equal to zero. These are also shown in the Figure 5.5b. The results of the equation are also shown were the calculated value of other output, ZNU $\geq$ 0. The illustration does indicate that the results of the simulation produce a limiting condition that is a close approximation to the boundary conditions imposed by the mathematical feasibility and the data from the regression model.

The desired output expressed from the general form (5.3) yields an equation of the form

$$ZNU = 673.4 + 1268.0\ \ln XU - 1675.4\ \ln YU \qquad (5.5)$$

$$t = \qquad 8.7554 \qquad 13.2539$$

$$R^2 = 0.6961$$

$$F = 93.9269.$$

Analysis of the results shows that the null hypothesis, the value of the individual coefficients is zero, is rejected

$\overline{ZNU}$ = 8922.4

+ 473.0 XU + 103,437.3 lnXU

+ 31,862.5 YU

- 240,755.31 lnYU

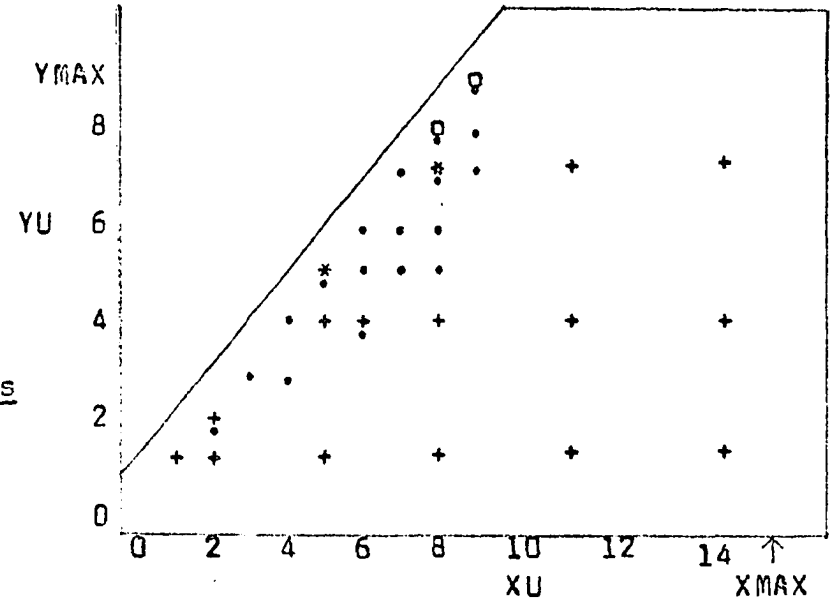


Fig. 5.5a

Simulation Results

+ = Data ZNU > 0

* = Data ZNU ≥ 0

o = Data ZNU = 0

Regression Results

• = Output = 0

ZNU = 673.4

+ 1268.0 lnXU

- 1675.4 lnYU

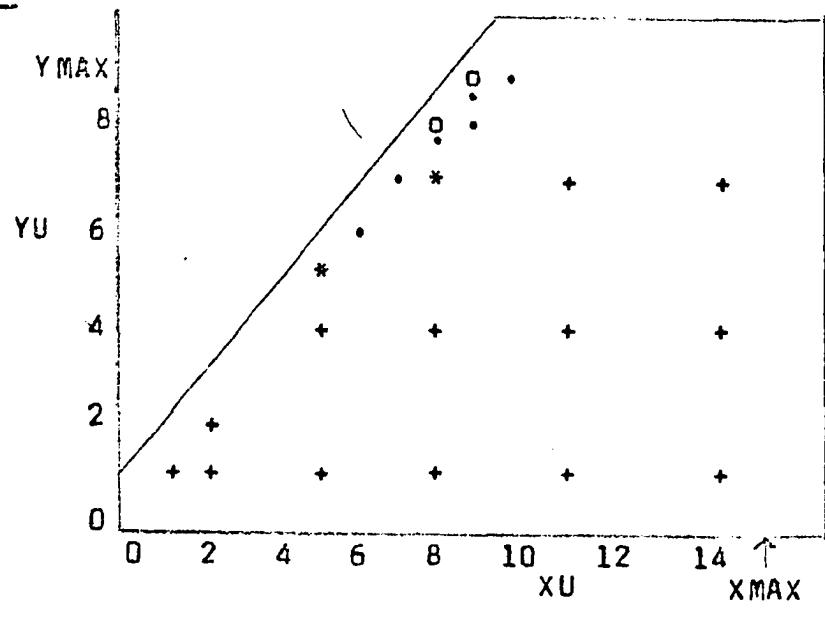


Fig. 5.5b

Fig. 5.5,---Feasible Region of Output (Based on Simulation and Regression Results)

by t tests. In addition the null hypothesis, the coefficient of multiple determination being equal to zero, is evaluated. The tabulated value of $F_{.05,(2,82)}$ is 4.88, which is less than the calculated value of F = 93.9269. Therefore the null hypothesis is rejected. The multiple correlation coefficient shows that 70 percent of the variation in the data has been handled in the derived equation.

An investigation of equation (5.5) for the indexing input above shows that the marginal contribution of indexing is 1268.0 XU^{-1}. Therefore within the area defined, output will increase with the number of indexing terms.

The limits imposed by the searching are an inverse relationship and indicate the effect of decreasing the level of desired output as the number of search terms increases.

The undesired output, from the general form of equation (5.4), is expressed in the function

$$\bar{ZNU} = 8922.4 + 473.0\ XU + 103{,}437.3\ \ln XU + 31{,}862.5\ YU - 240{,}755.3\ \ln YU \qquad (5.6)$$

$$t = 0.1100 \quad 4.1858 \quad 3.9484 \quad 9.2356$$

$$R^2 = 0.8332$$

$$F = 99.8897.$$

The t tests to determine if the individual coefficients will have a value of zero shows that the null hypothesis would be rejected at the 0.05 level, except the XU value. However in the application, all of the coefficients will be utilized. The test of the null hypothesis, the coefficient of multiple

determination being equal to zero, is rejected, since the calculated value, F = 99.8897, is greater than the tabulated value $F_{.05,(2,82)}$. The coefficient of multiple determination indicates that 83.3 percent of the variation of the data about the regression phase has been expressed in the equation depicted.

Conclusions

Conclusions based on usage of the reference retrieval model relate to model stability, numerical results, and error source identification.

Performance Model Stability

Analysis of the data indicate that the performance model is stable even though some of the probabilistic models used in calculation of output have a wide range of values. The results of the preliminary runs showed limited change in the output as the sample size to population of documents varied over a ratio of 1:7, to 7:7, and the number of iterations was modified over a range of 20 to 100, as shown in Figure 5.2.

The range of values of output ranged considerably from search to search at a given level of indexing. This action simulates a real reference retrieval system very well because the number of references to documents retrieved in actual systems will vary significantly. However, the levels of output do follow realistic trends.

Error Sources

Results indicate that the amount of the undesired output increases by the positive value of search terms and by negative logarithm of the number of search terms. However, the magnitude of the coefficients is such that the logarithmic value overrides the linear values for the interval of interest.

Since the desired output increases with the logarithmic value of index terms, the values of the coefficients indicate that the total ratios of undesired output to desired increases with the number of index terms. Reference to the search terms used shows that each additional term reduces the desired output, ZNU, approximately one and one third times as much as the index terms increase output.

The results of the performance model subsequently utilized in the regression model generate numerical results depicting a real-life reference retrieval model. The results of the regression model suggest that, because of the limited range of the independent variables, XU, YU, their relationship to the dependent variables ZNU can be expressed by more than one relationship, which is approximated by a logarithmic function.

Analysis of the results obtained indecates three major conclusions.

1. The model is stable,
2. Economic feasibility and conditions for optimization

are present,

a. the numerical values obtained fit the feasible region,

b. positive output was obtained for the feasible region,

c. the output of the ZNU model is positively correlated with the number of index terms using a logarithmic function. The output is inversely related to the level of searching as shown by the negative coefficient of the logarithmic expression for the number of search terms,

d. the output of the $Z\bar{N}U$ model shows a similar relationship except that linear terms have been introduced and the coefficients for both the indexing and searching are positive. However the coefficients of the logarithmic elements are positive and negative for the indexing and searching terms respectively and their magnitude is such that they override the linear relationship except at the (1,1) position.

3. Error sources can be identified and quantified in the inputs and their effect on output determined. The observed errors in indexing and searching stem from imperfections in language and inconsistencies in usage. The results of this model do not, however, provide a basis for discriminating between these two error sources.

CHAPTER VI

TOTAL COST OF FACILITY-TOTAL VALUE OF OUTPUT BENEFITS MODEL

The total cost of facility-total value of output benefits model is a structure for computing system costs, user costs, and the value of the output to the user. These costs are expressed so that they can be related to the decision variables (the number of index and search terms, X and Y); and the optimum or most profitable level of operation for the entire system will be determined.

The model will be structured by identifying all aspects of the systems independently, then relating them to the appropriate outputs by use of production functions. The decision variables in this model are the number of index and search terms. Therefore, all of the "intermediate outputs" from the various production functions will be related to X or Y, or both variables.

The production functions are aggregated for the intermediate phases, and a composite production function for the entire reference retrieval system is included. Production functions for both the cost and value phases are prepared. Cost and value will be expressed by applying the appropriate price per unit of input for each unit of each type of input used and are generalized as follows, from equation (3.7).

$$\text{Profit} = TVOB - TCF.$$

The total cost of facilities includes all expenditures and costs of installing and operating the reference retrieval system up to the level of producing output. The total value of output benefits includes the benefits (retrieved references to citations) and the associated costs of retrieving and evaluating the output of references. The rationale of allocating costs into these groups rests on two points. The first point of consideration is that this approach categorizes cost by source of data, a deterministic determination of the TCF and use of simulation to generate the magnitude of TVOB. The second point is that in deriving the net benefits, the loss associated with the unrecalled desired references (which have a negative value) are mathematically added to the value of the desired output of the system, the desired references. Thus it seems rational to add the other determination of TVOB, the cost of retrieving all output from the computer, along with the cost to the user of evaluating the output to the value of output producing a package of costs derived by simulation.

The entire reference retrieval system will be modeled by using a number of independent production functions. The aggregation of these functions will include all phases of the operations in the system. The output of each production function can be considered as a step toward the preparation of the final output which is retrieved references to indexed documents. The production functions will all be formulated

by using the basic inputs related to the index terms, X, and search terms, Y, or final output, Z, to provide consistent variables for interrelating the various phases of the system. Each successive step in the operation can be considered as making a marginal contribution toward the final output, Z, starting with X and/or Y. The output quantity of the intermediate production functions will not be recorded as such. The entire operation must be concluded before any quantity of output, Z, is measured.

The basic production function is of the form

$$Z_{ij} = h_{ij}(X,Y) = g_{ij}(r_{1j},s_{1j},t_{1j},a_{ij}/D,S,I,E,T,R,A).$$

The basic total cost function is of the form

$$C_{ij}(Z) = P_{ij}[f_{ij}(r_{1j},s_{1j},t_{1j},a_{ij}/D,S,I,E,T,R,A)],$$

where r_{1j}, s_{1j}, and t_{1j} are the independent variables of the various production functions, and the a_{ij}'s are the coefficients of the functions.

X_{ij}, Y_{ij} = intermediate outputs of the system indexed formulated searches

Z_{ij} = final output = number of documents retrieved

C_{ij} = total cost phases of the system

The independent variables are as follows:

D = Number of indexed documents in the reference file,

S = Number of user queries to be formulated annually,

I = Number of terms in the indexing vocabulary,

E = Number of terms in the searching vocabulary,

T = Number of searching installations,

R = Number of new documents indexed annually (replacements),

A = Number of search files replaced annually per installation as related to the number of new documents indexed annually,

and where,

H = Number of times a search file is replaced annually,

G = The ratio of the number of new documents indexed annually to the number of indexed documents included in a replacement reference file,

$$\frac{R}{D} \leq G \leq \frac{R}{R} = 1$$

$$A = \frac{HR}{G} = \frac{HR}{\frac{R}{D}} = HD$$

or $$A = \frac{HR}{1} = HR.$$

The detailed development of the production functions and costs are demonstrated for the cost aspect of the system, then followed by the value functions. The basis of the mathematical form of the functions is determined from (1) known derived relationships, and (2) previously cited relationships.

Previous work in indexing has shown a curvilinear relationship between the independent variable of the time needed to index a document and the number of terms used to index a document. Previous work has shown that computer usage is distributed on a linear time basis. Also many clerical functions are treated similarly. Therefore, unless specific data are available to suggest other forms of the relationships

between output and input for any phase of the system, straight-line relationships are presented.

Total System Production Functions

Three major phases of the operation of a reference retrieval system can be readily identified: capital equipment and development, operations, and measure of output.

The first two phases constitute the total facilities, and the last is the total output benefits. The discussion presented under the succeeding three subheadings is identical to the format presented in Chapter III. It is also consistent with the data presented in Table 6.1.

Total Facilities

This phase dexcribes the physical equipment and facilities needed to implement a reference retrieval system.

Capital Equipment and Development. This subheading is equivalent to phase level I of Table 6.1 and would include all of the necessary installations, equipment, material, and indexing effort to establish the reference retrieval system as a full-capacity operating entity. The development will be the establishment of production functions, followed by the construction of cost functions. The production functions for initial investment and document acquisition have the general form

$$X_j = g_j(r_{1j}, a_{1j}, b_{1j}/D, I, T, R, A)$$

where r_{1j} = the independent variable under consideration,

a_{1j} and b_{1j} = coefficient showing the dependency relationship of the function,

and X_j = intermediate level of output.

Two phases of capital are considered; (a) the development which is indexing and document acquisition and services needed to transform raw documents into retrievable indexed references to documents, and (b) the capital equipment and facilities acquisition.

A. The production functions for the development phase includes the initial indexing and the document acquisition which can be grouped as follows:

1. Initial Indexing consists of indexing and data input preparation.

a. Indexing--The relationship between time expended in indexing a document and the output of index terms per document is shown. This relationship is based on the work of Cleverdon (page 26) and Costello (92), who presented results showing decreasing returns which are approximated by a linear relationship. See Figure 6.1.

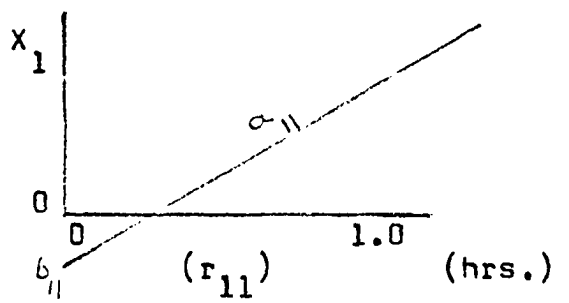


Fig. 6.1,--- Indexing Vs. Time

$$X_1 = b_{11} + a_{11}(r_{11}) \qquad (6.1)$$

where X_1 = stage of output in number of index terms,

r_{11} = time in hours,

a_{11} and b_{11} = constants for the function.

Subsequently, in the total cost function, the inverse relationship is desired and will be developed at that time.

b. Data-Input Preparation--Data-Input Preparation can be expressed as a linear function, related to the number of terms per document, given the number of indexed documents and the number of index terms.

$$X_2 = a_{12}(r_{12}),$$

r_{12} = the time in hours,

a_{12} = the coefficient.

2. Index Input consists of two parts, data insertion and file preparation of indexed documents.

a. Data Insertion consists of the operations necessary to update files and is, therefore, a function of the time involved related to the number of index terms, X, and the number of documents, D.

$$X_3 = a_{13}(r_{13}),$$

where r_{13} = the time needed to insert data into the appropriate records.

b. File Preparation is the formulation of index citations records for the distribution to the appropriate search centers. These records may be replaced throughout the year in operations, but only one record per installation, I, is required initially. The volume of throughput is a function of the number of documents, D, and the number of index terms, X.

$$X_4 = a_{14}(r_{14}),$$ where r_{14} = the time needed to reproduce records.

3. Document Acquisition Expense includes the expenditures for document acquisition and time involved in their processing.

a. Document Acquisition

$$X_5 = a_{15}(D)$$ for D, documents.

b. Document Processing Operations

$$X_6 = a_{16}(D) \quad \text{for D, documents.}$$

8. The category of Equipment and Facilities includes the items which have some salvage value.

1. Equipment includes all computer and necessary attached hardware.

$$Z_1 = a_{17}(r_{17})$$

where r_{17} = measure of computer capability.

2. Facilities include space for operation, auxiliary equipment, and related items.

$$Z_2 = a_{18}(r_{18})$$

where r_{18} = measure of computer associated facilities.

Operating Functions. This subheading is equivalent to phase level II of Table 6.1. Operating functions can be subdivided into several phases after the independent aspects of indexing and searching are considered. The indexing functions that have been previously developed under the initial investment phase are shown on phase level C of Table 6.1. For this phase the appropriate variables used are the same as under level I, but the values of the constants may be changed. The annual replacement of references to documents and the frequency of replacing of retrieval files per year is T x A. There is no acquisition of fixed factors.

The searching production functions include all phases of activity from preparation of a request by a user to interrogating a retrieval file including request formulation.

Table 6.1. Total System Production Functions

Part a. Total Facilities

Phase Level	Phase Definition	Production Function Dependant Variable Phase Value	Independant Variable of Production Function	Production Function	Number of Production Functions in System
	TOTAL FACILITIES				
I.	Capital Equipment & Development				
1.	Initial Indexing				
a.	Indexing	X_1	r_{11}	$X_1 = b_{11} + a_{11}\ (r_{11})$	D
b.	Data-Input Preparation	X_2	r_{12}	$X_2 = a_{12}(r_{12})$	D
2.	Index Input				
a.	Data Insertion	X_3	r_{13}	$X_3 = a_{13}(r_{13})$	D
b.	File Preparation	X_4	r_{14}	$X_4 = a_{14}(r_{14})$	D·T
3.	Document Acquisition Expense				
a.	Document Acquisition	X_5	D	$X_5 = a_{15}(D)$	1
b.	Document Processing	X_6	D	$X_6 = a_{16}(D)$	1
A.	Indexing & Document Acquisition	$X_o = D[X_1 + X_2 + X_3 + T(X_4)] + X_5 + X_6$ $X_o = D[b_{11} + a_{11}\ (r_{11}) + a_{12}(r_{12}) + a_{13}(r_{13}) + T(a_{14}(r_{14})) + a_{15} + a_{16}]$			
1.	Equipment	Z_1	r_{17}	$Z_1 = a_{17}(r_{17})$	1
2.	Facilities	Z_2	r_{18}	$Z_2 = a_{18}(r_{18})$	1
B.	Capital Equipment and Facilities Acquisition	$Z_o = Z_1 + Z_2$ $Z_o = a_{17}(r_{17}) + a_{18}(r_{18})$			

Table 6.1. Part e. (Continued) Total Facilities

II.	Operating Functions				
1.	Primary Indexing				
a.	Indexing	X_{11}	r_{11}	$X_{11} = b_{11} + a_{11}\ (r_{11})$	R
b.	Data-Input Preparation	X_{12}	r_{12}	$X_{12} = a_{12}(r_{12})$	R
2.	Data Input				
a.	Data Insertion	X_{13}	r_{13}	$X_{13} = a_{13}(r_{13})$	R
b.	File Preparation	X_{14}	r_{14}	$X_{14} = a_{14}(r_{14})$	T·A
3.	Document Acquisition Expense				
a.	Document Acquisition	X_{15}	R	$X_{15} = a_{15}(R)$	1
b.	Document Processing	X_{16}	R	$X_{16} = a_{16}(R)$	1
C.	Operating Indexing & Document Acquisition	$X_{20} = R[X_{11} + X_{12} + X_{13}] + T \cdot A(X_{14}) + X_{15} + X_{16}$ $X_{20} = R[b_{11} + a_{11}\ (r_{11}) + a_{12}(r_{12}) + a_{13}(r_{13}) + a_{15} + a_{16}] + T \cdot A[a_{14}(r_{14})]$			
1.	Search Preparation				
a.	User Request Preparation	Y_{21}	s_{11}	$Y_{21} = a_{21}(s_{11})$	S
b.	Search Formulation	Y_{22}	s_{12}	$Y_{22} = b_{12} + a_{22}\ (s_{12})$	S
c.	Data Input Preparation	Y_{23}	s_{13}	$Y_{23} = a_{23}(s_{13})$	S
2.	Retrieval Installation Operation				
a.	Basic Operation	Y_{24}	s_{14}	$Y_{24} = a_{24}(s_{14})$	T
b.	File-Searching	Y_{25}	s_{15}	$Y_{25} = a_{25}(s_{15})$	S
D.	Searching =	$Y_{30} = S[Y_{21} + Y_{22} + Y_{23} + Y_{25}] + T[Y_{24}]$ $Y_{30} = S[a_{21}(s_{11}) + b_{12} + a_{22}\ (s_{12}) + a_{23}(s_{13}) + a_{25}(s_{15})] + T[a_{24}(s_{14})]$			

TOTAL FACILITIES = TF = $X_o + Z_o + X_{20} + Y_{30}$

$$TF = [D + R][b_{11} + a_{11}(r_{11}) + a_{12}(r_{12}) + a_{13}(r_{13}) + a_{15} + a_{16}] + S[a_{21}(s_{11}) + b_{12} + a_{22}(s_{12}) + a_{23}(s_{13}) + a_{25}(s_{15})] + T[(A + D)(a_{14}(r_{14})) + (a_{24}(s_{14}))] + a_{17}(r_{17}) + a_{18}(r_{18})$$

Table 6.1. Part b. Total Output Benefits

Phase Level	Phase Definition	Production Function Dependent Variable Phase Value	Independent Variable of Production Function	Production Function	Number of Production Functions in System
	TOTAL OUTPUT BENEFITS				
1.	Output Preparation	Z_{41}	ZU	$Z_{41} = a_{41}(ZU)$	S
2.	Output Data Evaluation	Z_{42}	ZU	$Z_{42} = a_{42}(ZU)$	S
A.	Output Preparation and Evaluation	$Z_{40} = S[Z_{41} + Z_{42}]$ $Z_{40} = S[(a_{41} + a_{42})(ZU)]$ $Z_{40} = S[(a_{41} + a_{42})(ZNU + Z\bar{N}U)]$			
1.	Value Needed Used Output	Z_{43}	ZNU	$a_{43}(ZNU)$	S
2.	Cost of Needed Not Used Output	Z_{44}	$Z\bar{N}U$	$a_{44}(Z\bar{N}U)$	S
B.	Value of Output	$Z_{45} = S[Z_{43} + Z_{44}]$ $Z_{45} = S[a_{43}(ZNU) + a_{44}(Z\bar{N}U)]$ $Z_{45} = S[(a_{43} + a_{44})(ZNU) + (a_{44})(ZN)]$			

TOTAL OUTPUT BENEFITS = $TOB = Z_{40} + Z_{45}$

$$TOB = S[(a_{41} + a_{42})(ZNU + Z\bar{N}U) + (a_{43} + a_{44})(ZNU) + a_{44}(ZN)]$$

$$= S[(a_{41} + a_{42} + a_{43} + a_{44})(ZNU) + (a_{41} + a_{42})(Z\bar{N}U) + (a_{44})(ZN)]$$

All functions are either related to the number of search terms, Y, or are constant. These are shown in phase level D of Table 6.1. These functions are subdivided into those attributable to a search formulation and those related to the retrieval installation operation. Searching functions can, in general, be expressed as

$$Y_{2j} = g_{2j}(s_{1j}, a_{2j}/D, S, E, T),$$

where s_{1j} = the dependent variable under consideration, and

a_{2j} = the dependent relationship coefficient,

Y_{2j} = intermediate processing output.

1. Search Preparation. This includes all functions from user-time cost to search-data preparation attributed to the number of search terms used.

   a. User Request Preparation includes the user's time in preparing a request.

   $$Y_{21} = a_{21}(s_{11})$$

   where a_{21} = coefficient

   s_{11} = user-time to prepare a request in hours

   b. Search Formulation is the use of the search intermediary's effort to translate a user's request into terms compatible with the search vocabulary of the system so as to interdict with the reference retrieval file. Since this has many attributes of indexing, the same type of function is used.

   $$Y_{22} = b_{12} + a_{22}\ (s_{12})$$

   where s_{12} = time in hours

a_{22} = constant

b_{12} = constant

c. Data Input Preparation, which is a linear function of the number of terms used, is

$$Y_{23} = a_{23}(s_{13}),$$

where s_{13} = time in hours.

2. Retrieval Installation Operation. These are the functions involved in interrogating the reference retrieval file for the documents of interest at each installation, A.

a. Basic Operation, materials, utilities per installation, T, are

$$Y_{24} = a_{24}(s_{14}),$$

where s_{14} = a measure of materials quantity.

b. File-Searching costs associated with accessing the files at each of the T installations, which is proportional to the number of terms per each formulated search of S searches, are

$$Y_{25} = a_{25}(s_{15}),$$

where s_{15} = time in hours to search.

Total Output Benefits Functions

Output functions relate the preparation of the number of references to documents and their evaluation along with their value or benefits which are related to output quantity, Z. The relation of Z to the number of index terms, X, and search terms, Y, is obtained by use of the performance model which is actually a point profuction function. Therefore, for a given level of output, Z, its relationship to the independent variables X and Y in known. Thus, the number of

references obtained is related to X and Y. The general form is

$$Z_{4j} = g_{4j}(t_{1j}/D,S,I,E,T,R,A).$$

A. Output Preparation and Evaluation include two phases.

1. Output Preparation equals basic output preparation and is proportional to the number of references to be cited.

$$Z_{41} = a_{41}(ZU)$$

where ZU = number of references to be cited.

2. Output Data Evaluation function, which is the effort (or time used) by the user evaluating the output to ascertain if the references obtained are consistent with his request, is

$$Z_{42} = a_{42}(ZU).$$

B. Value of Output includes the positive value of retrieved output and a penalty for unretrieved needed citations.

1. Value of Needed Used Output

$$Z_{43} = a_{43}(ZNU)$$

2. Cost of Needed Not Used Output

$$Z_{44} = a_{44}(Z\bar{N}U)$$

At this point some combination of these functions using known relationships can be accomplished. Equation (3.2) can be converted to output and the result is

$$ZU = ZNU + Z\bar{N}U.$$

Therefore, a new expression can be derived, where

$$Z_{40} = S(a_{41} + a_{42})(ZU),$$

$$Z_{40} = S(a_{41} + a_{42})(ZNU + Z\bar{N}U).$$

A similar relationship for equation (3.1) is available.

$$ZN = ZNU + ZN\bar{U}.$$

Transposing,

$$ZN\bar{U} = ZN - ZNU.$$

This can be written as shown in Table 6.1.

$$Z_{45} = S[a_{43}(ZNU) + a_{44}(ZN\bar{U})]$$
$$Z_{45} = S[a_{43}(ZNU) + a_{44}(ZN) - a_{44}(ZNU)]$$
$$Z_{45} = S[(a_{43}-a_{44})(ZNU) + (a_{44})(ZN)].$$

Therefore, the combined equation for TOB can be written, where

$$TOB = Z_{40} + Z_{45}$$
$$TOB = S[(a_{41}+a_{42})(ZNU + Z\bar{N}U)]$$
$$+ S[(a_{43}-a_{44})(ZNU + a_{44}(ZN)].$$

Rearranging,

$$TOB = S[(+a_{41}+a_{42}+a_{43}-a_{44})(ZNU)$$
$$+ (a_{41}+a_{42})(Z\bar{N}U) + (a_{44})(ZN)].$$

It is to be noted that all of these values are derived as part of the output from the performance model. In addition ZN is a constant; therefore, surfaces for only two functions need be generated, those for ZNU and for $Z\bar{N}U$.

Total System Cost Functions

Since total costs are a function of the number of units of input used, they are a product of the price per unit of input and the number of units used. It is desired to relate these costs to the level of usage of the index and/or search terms and the number of units of output.

The inverse function between inputs and outputs is desired as shown.

$$C_{ij}(X,Y,Z) = P_{ij}[f_{ij}(r_{1j},s_{1j},t_{1j},a_{ij}/D,S,I,E,T,R,A)]$$

Total costs are generated by applying the price per unit of each input in conjunction with the production functions. A total cost function is shown for a given case and the general situation. Since benefits, costs, and expenditures are matched timewise in the operating and output phase, they can be related directly by using production functions to formulate the cost functions. However, the inverse relationship of variables is desired because costs are related to the independent variable of the production function.

Total Cost of Facilities

Reference to the production functions under initial indexing showed the equation (6.1)

$$X_1 = b_{11} + a_{11}(r_{11}),$$

where X_1 is the output or dependent variable and while total cost is related to the output level, it has a fixed relationship to the level of input, r_{11}.

$$r_{11} = \left(\frac{X_1+b_{11}}{a_{11}}\right), \text{ since } b_{11} \text{ has a negative value.}$$

Since there are D documents the total cost equals

$$C_1(X) = D\ P_{11}\left(\frac{X_1+b_{11}}{a_{11}}\right)$$

P_{11} = cost of indexing, $/hr.,

and is shown graphically in Figure 6.2.

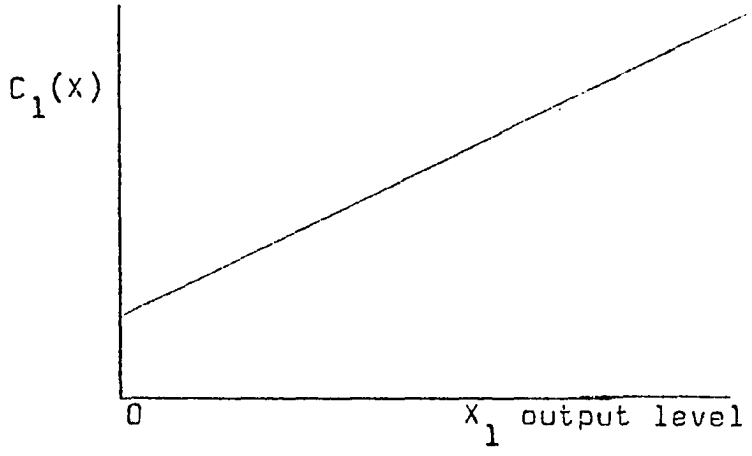


Fig. 6.2,---Cost of Indexing

For data input preparation, D documents must be inserted; therefore, their number must be included; and the specific form is

$$C_2(X) = D\ X_2 = D\ P_{12}(X_2/a_{12}).$$

The specific form of the production function for data insertion is

$$X_3 = a_{13}(r_{13}).$$

In the above case the total cost function for this step is

$$C_3(X) = D\ P_{13}(X_3/a_{13}).$$

In the initial phase, however, cost adjustments must be made. The production functions are constant for the system; however, since the benefits of the initial phase are obtained over a period of time, the costs will be allocated so that they match the time period of benefits. However, the investment expenditures are made initially; to obtain consistent comparisons the concept of time value of money will be used to put cost on an annual basis.

Therefore, the monetary expenditure cost function is related to the equivalent annual cost function. These functions are presented separately for the fixed and variable components of the initial phase. The initial investment and indexing facilities contain factors that are fixed for the given installation and those that vary with the number of index terms, X. This procedure provides a natural division of factors, and therefore the production and cost functions are similarly grouped. Therefore, there will be an equivalent annual cost for each of the two phases.

Total cost and equivalent annual cost are obtained from the format of Barish [93].

$$EAC = \left[\frac{m(1+m)^{o}}{(1+m)^{o}-1}\right]$$

EAC = the equivalent annual cost.

Given:

o = the life of the capital equipment and facilities,

L = the ratio of salvage value to initial investment (straight-line depreciation),

m = the annual interest rate,

CR = the capital recovery factor.

Therefore, the expenditure cost function for the development phase is

$$C_o(X) = D \times \left[\frac{P_{12}}{a_{12}} + \frac{P_{13}}{a_{13}} + \frac{P_{14}}{a_{14}}\right] + D\left[P_{11}\frac{X+b_{11}}{a_{11}} + P_{15}+P_{16}\right].$$

The The equivalent annual cost function for the development or indexing and document acquisition is

$$C_{50}(X) = C_o(X) \left[\binom{m}{o}^{CR}\right].$$

This equation does not provide for any salvage because there are no assets that can be considered as having tangible resale value.

The cost function for the capital equipment and facilities acquisition phase is

$$C_{10}(Z) = P_{17} + P_{18}.$$

The equivalent annual cost function for the equipment and facilities phase is

$$C_{60}(Z) = C_{10}(Z)\left[(1-L)\binom{m}{o}^{CR} + L\ m\right].$$

The equivalent annual cost functions for the initial phases of the model are

$$C_{50}(X) + C_{60}(Z) = \left[D\ X\left(\frac{P_{12}}{a_{12}} + \frac{P_{13}}{a_{13}} + \frac{P_{14}}{a_{14}}\right) + D\right.$$
$$\left.\left(P_{11}\binom{\frac{X+b_{11}}{a_{11}}}{} + P_{15} + P_{16}\right)\right]\left[\binom{m}{o}^{CR}\right] + \left[P_{17} + P_{18}\right]$$
$$\left[(1-L)\binom{m}{o}^{CR} + L\ m\right].$$

Similar utilization of the previously developed production functions to produce the appropriate total cost functions for the various phases of the inputs are developed and are shown on Table 6.2.

The final aggregate cost of facilities can be expressed as depicted and is

Table 6.2. TOTAL SYSTEM COST FUNCTION AND DATA FOR MODEL APPLICATION

Part a. Total Cost of Facilities

Phase Level	Phase Definition	Total Cost Function General Expression	Price per Unit of Independent Variable and Function	Specific Total Cost Function Used to Demonstrate the Model
	TOTAL COST OF FACILITIES			
I.	Capital Equipment & Development			
A.	Indexing and Document Acquisition	$C_o(X)$		
1.	Initial Indexing			
a.	Indexing	$C_1(X) = D \cdot P_{11}\left(\frac{X_1 - b_{11}}{a_{11}}\right)$	P_{11} = \$10.00/hr.	$C_1(X) = 700{,}000\left[10.00 \cdot \left(\frac{X + 4.00}{53.33}\right)\right]$
b.	Data Input Preparation	$C_2(X) = D \cdot P_{12}(X_2/a_{12})$	P_{12} = \$ 7.00/hr	$C_2(X) = 700{,}000[0.133\ X]$
2.	Index Input			
a.	Data Insertion	$C_3(X) = D \cdot P_{13}(X_3/a_{13})$	$C_3(X) + C_4(X) = \$1900$	$C_3(X) + C_4(X) = 1800$
b.	File Preparation	$C_4(X) = D \cdot T \cdot P_{14}(X_4/a_{14})$		
3.	Document Acquisition Expense			
a.	Document Acquisition	$C_5(X) = D \cdot P_{15}/a_{15}$	P_{15} = \$.28/document	$C_5(X) = 700{,}000(.28)$
b.	Document Processing	$C_6(X) = D \cdot P_{16}/a_{16}$	P_{16} = \$.25/document	$C_6(X) = 700{,}000(.25)$

Indexing & Document Acquisition = $C_o(X) = C_1(X) + C_2(X) + C_3(X) + C_4(X) + C_5(X) + C_6(X)$

$$C_o(X) = D\left\{X\left[\frac{P_{12}}{a_{12}} + \frac{P_{13}}{a_{13}} + T \cdot \frac{P_{14}}{a_{14}}\right] + \left[P_{11} \cdot \left(\frac{X + b_{11}}{a_{11}}\right) + \frac{P_{15}}{a_{15}} + \frac{P_{16}}{a_{16}}\right]\right\}$$

$$C_o(X) = 700{,}000\left[10.00 \cdot \left(\frac{X + 4.00}{53.33}\right) + 0.133X\right] + 372{,}800$$

Equivalent Annual Cost $= C_{50}(X) = C_o(X)\left[\binom{CR}{m}_{o}\right]$

$$C_{50}(X) = \left\{700{,}000\left[10.00 \cdot \left(\frac{X + 4.00}{53.33}\right) + 0.133X\right] + 372{,}800\right\}\left[\binom{CR}{m=.10}_{o=8}\right]$$

$$C_{50}(X) = \left\{700{,}000\left[10.00 \cdot \left(\frac{X + 4.00}{53.33}\right) + 0.133X\right] + 372{,}800\right\}[.18744]$$

B.	Equipment and Facilities Acquisition	$C_{10}(Z)$		
1.	Equipment	$C_1(Z) = P_{17}(Z_1/a_{17})$	$C_1(Z)$=\$35,000/total equipment	$C_1(Z) = 35{,}000$
2.	Facilities	$C_2(Z) = P_{18}(Z_2/a_{18})$	$C_2(Z)$=\$20,000/total facility	$C_2(Z) = 20{,}000$

Equipment and Facilities Acquisition Cost $C_{10}(Z) = C_1(Z) + C_2(Z)$

$$C_{10}(Z) = P_{17}(Z/a_{17}) + P_{18}(Z/a_{18})$$

$$C_{10}(Z) = 55{,}000$$

Equivalent Annual Cost $C_{60}(Z) = C_{10}(Z)\left[(1-L)\binom{CR}{m}_{o} + (L \cdot m)\right]$

$$C_{60}(Z) = [55{,}000]\left[(1-L)\binom{CR}{m=.10}_{o=8} + (L \cdot m)\right]$$

$$C_{60}(Z) = [55{,}000][(1-.33333)(.18744) + (.33333)(.10)]$$

$$C_{60}(Z) = 8706$$

Table 6.2. Part a. (Continued) Total Cost of Facilities

II.	Operating Functions			
C.	Operating Indexing & Document Acquisition			
1.	Primary Indexing			
a.	Indexing Documents	$C_{11}(X) = R \cdot P_{11} \left(\frac{X_{11}+b_{11}}{a_{11}}\right)$	P_{11} = \$10.00/hr.	$C_{11}(X) = 175{,}000\left[10.00 \cdot \left(\frac{X+4.00}{53.33}\right)\right]$
b.	Data Input Preparation	$C_{12}(X) = R \cdot P_{12}(X_{12}/a_{12})$	P_{12} = \$ 7.00/hr.	$C_{12}(X) = 175{,}000[0.133X]$
2.	Data Input			
a.	Data Insertion	$C_{13}(X) = R \cdot P_{13}(X_{13}/a_{13})$	$C_{13}(X)+C_{14}(X)$ = \$3600	$C_{13}(X) + C_{14}(X) = 360$
b.	File Preparation	$C_{14}(X) = T \cdot A\ P_{14}(X_{14}/a_{14})$		
3.	Document			
a.	Document Acquisition	$C_{15}(X) = R \cdot P_{15}/a_{15}$	P_{15} = \$.28/document	$C_{15}(X) = 175{,}000(.28)$
b.	Document Processing	$C_{16}(X) = R \cdot P_{16}/a_{16}$	P_{16} = \$.25/document	$C_{16}(X) = 175{,}000(.25)$

Operating Indexing and Document Acquisition = $C_{20}(X) = C_{11}(X) + C_{12}(X) + C_{13}(X) + C_{14}(X) + C_{15}(X) + C_{16}(X)$

$$C_{20}(X) = R \cdot X\left[\frac{P_{12}}{a_{12}} + \frac{P_{13}}{a_{13}}\right] + \left[P_{11}\left(\frac{X+b_{11}}{a_{11}}\right) + \frac{P_{15}}{a_{15}} + \frac{P_{16}}{a_{16}}\right] + T \cdot A \cdot X\left[\frac{P_{14}}{a_{14}}\right]$$

$$C_{20}(X) = 175{,}000\left[10.000 \cdot \left(\frac{X+4.00}{53.33}\right) + 0.133X\right] + 95{,}350$$

D.	Searching			
1.	Search Formulation			
a.	User Request Preparation	$C_{21}(Y) = S \cdot P_{21}(Y_{21}/a_{21})$	P_{21} = \$10.00/user request	$C_{21}(Y) = 500\ [10.00] = 5{,}000$
b.	Search Formulation	$C_{22}(Y) = S \cdot P_{22}\left(\frac{Y+b_{12}}{a_{22}}\right)$	P_{22} = \$10.00/hr.	$C_{22}(Y) = 500\left[10.00 \cdot \left(\frac{Y+4.00}{53.33}\right)\right]$
c.	Data Input Preparation	$C_{23}(Y) = S \cdot P_{23}(Y_{23}/a_{23})$	P_{23} = \$ 7.00/hr.	$C_{23}(Y) = 500\ [0.133 \cdot Y]$
2.	Retrieval Installation Operation			
a.	Basic Operation	$C_{24}(Y) = T \cdot P_{24}(Y_{24}/a_{24})$	$P_{24}\frac{Y_{24}}{a_{24}}$ = \$6,000/search facility/year	$C_{24}(Y) = 1(6{,}000) = 6{,}000$
b.	File Searching	$C_{25}(Y) = S \cdot P_{25}(Y_{25}/a_{25})$	P_{25} = \$159.33/hr.	$C_{25}(Y) = 500\ (2.41Y)$

Searching Cost = $C_{30}(Y) = C_{21}(Y) + C_{22}(Y) + C_{23}(Y) + C_{24}(Y) + C_{25}(Y)$

$$C_{30}(Y) = S\left(Y\left[\frac{P_{21}}{a_{21}} + \frac{P_{23}}{a_{23}} + \frac{P_{25}}{a_{24}}\right] + \left[P_{22}\left(\frac{Y+b_{12}}{a_{22}}\right)\right] + T \cdot Y\left(\frac{P_{24}}{a_{24}}\right)\right)$$

$$C_{30}(Y) = 500\ \left[10.00 \cdot \left(\frac{Y+4.00}{53.33}\right) + 2.543 \cdot Y\right] + 11{,}000$$

TOTAL COST OF FACILITIES = TCF = $C_{50}(X) + C_{60}(Z) + C_{20}(X) + C_{30}(Y)$

$$TCF = D\left\{X\left[\frac{P_{12}}{a_{12}} + \frac{P_{13}}{a_{13}} + \frac{P_{14}}{a_{14}} + T \cdot \frac{P_{14}}{a_{14}}\right] + \left[P_{11}\left(\frac{X+b_{11}}{a_{11}}\right) + \frac{P_{15}}{a_{15}} + \frac{P_{16}}{a_{16}}\right]\right\}\left\{\binom{CR}{m}_{o}\right\} + \left\{\left[\frac{P_{17}}{a_{17}} + \frac{P_{18}}{a_{18}}\right]\left[(1-L)\binom{CR}{m}_{o} + (L \cdot m)\right]\right\}$$

$$+ R\left\{X\left[\frac{P_{12}}{a_{12}} + \frac{P_{13}}{a_{13}}\right] + \left[P_{11}\left(\frac{X+b_{11}}{a_{11}}\right) + \frac{P_{15}}{a_{15}} + \frac{P_{16}}{a_{16}}\right]\right\} + S\left\{Y\left[\frac{P_{21}}{a_{21}} + \frac{P_{23}}{a_{23}} + \frac{P_{25}}{a_{25}}\right] + \left[P_{22}\left(\frac{Y+b_{12}}{a_{12}}\right)\right]\right\} + \left\{T\left[XA\left(\frac{P_{14}}{a_{14}}\right) + Y\frac{P_{24}}{a_{24}}\right]\right\}$$

$$TCF = \left\{700{,}000\left[10.00 \cdot \left(\frac{X+4.00}{53.33}\right) + 0.133X\right] + [372{,}600]\right\}\{.18744\} + \left\{175{,}000\left[10.00\left(\frac{X+4.00}{53.33}\right) + 0.133X\right]\right\}$$

$$+ \left\{500\ \left[10.00\left(\frac{Y+4.00}{53.33}\right) + 2.543Y\right]\right\} + \{116{,}056\}$$

$$TCF = 415{,}590 + 98{,}140\ X + 500\ (0.75 + 2.7305\ Y).$$

Table 6.2. TOTAL SYSTEM COST FUNCTION AND DATA MODEL APPLICATION

Part b. Total Value of Output Benefits

Phase Level	Phase Definition	Total Cost Function General Expression	Price Per Unit of Independent Variable & Function	Specific Total Cost Function Used to Demonstrate the Model
	TOTAL VALUE OF OUTPUT BENEFITS			
1.	Output Preparation	$C_{41}(Z) = S(-P_{41})(\frac{ZU}{a_{41}})$	P_{41} = \$.01/citation	$C_{41}(Z) =$ 500 (-.01) ZU
2.	Output Evaluation	$C_{42}(Z) = S(-P_{42})(\frac{ZU}{a_{42}})$	P_{42} = \$10.00 hr.	$C_{42}(Z) =$ 500 (-0.10) ZU
A.	Output Preparation & Evaluation Cost	$= C_{40}(Z) = C_{41}(Z) + C_{42}(Z)$		$C_{40}(Z) =$ 500[(-0.11) ZU]
		$C_{40}(Z) = S(\frac{-P_{41}}{a_{41}} - \frac{P_{42}}{a_{42}})(ZU)$		$C_{40}(Z) =$ 500[(-0.11) (ZNU + ZN̄U)]
		$= S(\frac{-P_{41}}{a_{41}} - \frac{P_{42}}{a_{42}})(ZNU + Z\bar{N}U)$		
1.	Value Needed & Used Output	$C_{43}(Z) = S(P_{43})(\frac{ZNU}{a_{43}})$	P_{43}	$C_{43}(Z) =$ 500 [(10.00) ZNU]
2.	Value Needed Not Used Output	$C_{44}(Z) = S(-P_{44})(\frac{ZN\bar{U}}{a_{44}})$	P_{44}	$C_{43}(Z) =$ 500 [(-1.00) ZN̄U]
B.	Value of Output	$= C_{45}(Z) = C_{43}(Z) + C_{44}(Z)$		
		$C_{45}(Z) = S\left[\frac{P_{43}}{a_{43}}(ZNU) - \frac{P_{44}}{a_{44}}(ZN\bar{U})\right]$		$C_{45}(Z) =$ 500 [(10.00)(ZNU) - (1.00) (ZN̄U)]
		$C_{45}(Z) = S\left[(\frac{P_{43}}{a_{43}} + \frac{P_{44}}{a_{44}})(ZNU) - \frac{P_{44}}{a_{44}}(Z\bar{N})\right]$		

TOTAL VALUE OF OUTPUT BENEFITS = TVOB = $C_{40}(Z) + C_{45}(Z)$

$$TVOB = S\left[(\frac{-P_{41}}{a_{41}} - \frac{P_{42}}{a_{42}} + \frac{P_{43}}{a_{43}} + \frac{P_{44}}{a_{44}})(ZNU) - (\frac{P_{41}}{a_{41}} + \frac{P_{43}}{a_{42}})(Z\bar{N}U) - (\frac{P_{44}}{a_{44}})(ZN)\right]$$

$$TVOB = 500[+10.89 \quad (ZNU) - 0.11(Z\bar{N}U) - 1.00(ZN)]$$

$$TCF = 415{,}590 + 98{,}140\ XU + 500(0.75 + 2.7305\ YU). \qquad (6.2)$$

The cost is expressed in linear functions of the number of index terms and search terms.

Total Value Output Benefits

Value measures or expresses the cost of preparing and evaluating the final output and the benefits derived from having the output. The output is expressed in quantitative units as the number of needed references to documents obtained. Also, there is a penalty for not obtaining all desired output which is the unrecalled needed references. The TVOB can be expressed related to three subphases of output: preparation and evaluation, value of needed used output, and a penalty for needed but not used (unretrieved) output. In addition there are, S, searches applicable to these individual query costs, output preparation and evaluation cost.

The value of the output is defined in relation to itself; converting the production functions to total cost functions is readily accomplished.

The relationship for TOB can be explicitly defined as taken from Table 6.1.

$$TOB = S[(a_{41} + a_{42} + a_{43} + a_{44})(ZNU) + (a_{41} + a_{42})(Z\bar{N}U) + (a_{44})(ZN)]$$

In general,

$$TVOB = P_{4j}[TOB].$$

So from Table 6.2

$$TVOB = S[(-\frac{P_{41}}{a} - \frac{P_{42}}{a} + \frac{P_{43}}{a} + \frac{P_{44}}{a})(ZNU) - (\frac{P_{41}}{a_{41}} + \frac{P_{42}}{a_{42}})(Z\bar{N}U) - (\frac{P_{44}}{a_{44}})(ZN)]. \qquad (6.3)$$

Where

- $-\frac{P_{41}}{a_{41}}$ = cost per retrieved citation of output preparation,
- $-\frac{P_{42}}{a_{42}}$ = cost per retrieved citation to the user for output evaluation,
- $\frac{P_{43}}{a_{43}}$ = value of each retrieved needed citation,
- $-\frac{P_{44}}{a_{44}}$ = loss in value of each unretrieved citation.

Three of the unit value parameters (P_{41}, P_{42}, and P_{44}) have negative signs since they are a cost or penalty factor.

CHAPTER VII

APPLICATION OF TOTAL COST-VALUE MODEL

Application of the total cost-value model depicted in the previous section necessitates use of the specific total cost of the facilities model developed for the deterministic phase and use of the total value model. The value in turn incorporates the probabilistically derived performance surfaces for the ZNU and $Z\bar{N}U$ output discussed in Chapter V. Additional cost or value data to generate the total value and total cost surfaces and to complete the model are needed. Optimization of the model can then be achieved by determining the point of maximum profit in terms of the variables of the system.

Data Sources

The data available for use in the total cost-value model are subject to the limitations discussed in Chapter II. Review of the literature, discussion with knowledgable individuals, and correspondence with various organizations have yielded reasonable data for utilization of the various relationships expressed in the entire model. All available pertinent data were used, and where possible the relevant relationships were established, followed by the application of cost data. Absence of data to implement the production functions necessitated the use of cost data directly related to

the decision variables of the number of index and/or search terms. Aggregation of these data allows for the determination of the most profitable level of operation of the decision variables X, Y of the system. These data are presented in Table 7.1, along with their sources.

Profit Maximization

Profit is the measure of optimization and is the difference between total value and total cost, where both factors are expressed in terms of output, which is represented by a surface in XU and YU. The maximum profit or optimum level of operation will, therefore, be where the two surfaces have the greatest difference. Normally, this relationship can be solved by taking the derivative of each equation, setting them equal to zero, and solving for the appropriate level of operation. However, this model has used two independent variables from equation (3.7).

$$\text{Profit} = \text{TVOB} - \text{TCF}.$$

Let profit be expressed as ZP.

From equations (6.3) and (6.2) the values of TVOB and TCF can be substituted into (3.7). This equation can now be expressed

$$ZP = S\left[\left(-\frac{P_{41}}{a_{41}} - \frac{P_{42}}{a_{42}} + \frac{P_{43}}{a_{43}} + \frac{P_{44}}{a_{44}}\right) ZNU - \left(\frac{P_{41}}{a_{41}} + \frac{P_{42}}{a_{42}}\right) ZNU - \left(\frac{P_{44}}{a_{44}}\right) ZN\right] - 415{,}590 - 98{,}140\ XU - S[0.75 + (2.7305\ YU)]. \qquad (7.1)$$

Table 7.1. Cost or Value of Various Items Needed for Application of TVOB-TCF System

Designation of Cost-Value Item	Definition	Cost or Value of Item	Data Sources
P_{11}	Indexing Cost		Montague*
P_{12}	Input Preparation Cost		Kuney
$C_3(X)+C_4(X)$	Index Input Cost		Kuney
P_{15}	Document Acquisition Cost		Library Journal
P_{16}	Document Processing Cost		Mueller
$C_1(Z)$ + $C_2(Z)$	Equipment Cost Facilities Cost		Personal estimate in conjunction with computer staff
P_{21}	User Request Preparation Cost		Mueller
P_{22}	Search Formulation Cost		Use same as indexing; Montague*
P_{23}	Data Input Preparation Cost		Use same as in indexing; Kuney

$Y \frac{p_{24}}{a_{24}}$	Basic Operation Cost		Personal estimate in conjunction with computer staff
p_{25}	File Searching		Rogers, U of Colo. Med. Center
$\frac{p_{41}}{a_{41}}$	Cost per retrieved citation of output preparation		Personal estimate in conjunction with computer staff
$\frac{p_{42}}{a_{42}}$	Cost per retrieved citation to the user for output evaluation		Mueller
$\frac{p_{43}}{a_{43}}$	Value of each retrieved needed citation		Mueller
$\frac{p_{44}}{a_{44}}$	Loss in value of each unretrieved citation		Estimate
m	Interest rate on initial investment	10%	Estimate
o	Expected life of system	8 yrs.	Review of iterature
L	Ratio of salvage value to initial investment expenditures	.33333	Personal estimate in conjunction with computer staff

* Also reported in Penner

Substituting the functional forms of ZNU and $Z\bar{N}U$ from equations (5.5) and (5.6) into (7.1), along with the value data, produce

$$\begin{aligned} ZP = 500[&(10.89)(673.4 + 1{,}268.0\ \ln XU - 1{,}675.4\ \ln YU) \\ &- (0.11)(8{,}992.4 + 473.0\ XU + 103{,}437.3\ \ln XU \\ &+ 31{,}8625\ YU - 240{,}755.3\ \ln YU) \\ &- (1.00)(6{,}279.2)] - 415{,}590 \\ &- 98{,}140\ XU - 500(0.75 + 2.7305\ YU). \end{aligned} \qquad (7.2)$$

The conditions for optimization were reviewed previously by Allen, page 49. The condition of the first partial derivations needed from equation (2.5) stated in the current variables, is

$$\frac{\partial ZP}{\partial XU} = \frac{\partial ZP}{\partial YU} = 0,$$

and will be expressed independently for the XU and YU varibles. Therefore, from equation (7.2),

$$\frac{\partial ZP}{\partial YU} = 500(2{,}430.43\ XU^{-1} - 248.31) \qquad (7.3)$$

has the value of zero when XU = 9.788.

Similarly,

$$\frac{\partial ZP}{\partial YU} = 500(8237.98\ YU^{-1} - 3507.61) \qquad (7.4)$$

has the value of zero when YU = 2.349.

The values of XU = 9.788 and YU = 2.349 were rounded to the closest integer point, XU = 10, YU = 2. The validity of this point to give maximum profit was verified by direct enumeration of the 135 feasible solution points. This solution is in the feasible range

$1 \leq XU \leq 15, \qquad XU = 10$

$1 \leq YU \leq 9, \qquad YU = 2$

$XU \geq YU, \quad 10 > 2.$

Analysis of Results

The objective of this phase of the work is to determine the optimum level of usage of the decision variables which are the number of index and search terms. The benefits are the the number of needed citations to documents retrieved. Therefore, the analysis is devoted to relating the output level and cost to the number of index and search terms. The total number of citations retrieved annually is a function of the number of searches conducted, which is held constant, and the number of citations retrieved per search. This analysis will present value, cost, and profit on a (1) total basis and (2) on a per needed retrieved citation. The second approach is generated by dividing the applicable numbers from the first approach by the number of needed retrieved citations obtained.

$ZNU \cdot S.$

Analysis of the results of model application is related to three facets: (1) an overall review of the points where various functions such as output values, cost, and profit are maximized or minimized, (2) the impact of the number of index and search terms, and (3) a detailed analysis of the region surrounding the most profitable level of operation.

Overall Review

An overall review of optimum positions of maximum value, minimum cost, and maximum profit are dependent on (1) the number of references to citations (ZNU and $Z\bar{N}U$) recovered, and (2) the per-unit value, cost, and profit which are related to the number of references to desired citations, ZNU. These factors in turn produce the total value of output benefits, total cost of facilities, and profits. Therefore an examination of the physical output will be followed by a review of the optimum points.

A review of physical output shows that ZNU is maximized at XU = 15, YU = 1 for a total of 4,107 needed references retrieved per search. Similarly the maximum value of $Z\bar{N}U$ is at the same point, and a value of 327,993 undesired references is provided per search. These data suggest that ZNU increases directly with XU and inversely with YU. Obviously this is not a desirable level of operation because of the high level of unneeded retrieved references. In addition, the impact of TCF shows that this position has a negative profit. The total value of output benefits (TVOB) is maximized at XU = 15, YU = 2 for a total value of $2,287,142, and profit is positive at this point. However, maximum TVOB per needed retrieved citation is at the level of XU = 5, YU = 3, which provides a value of $2.586 per retrieved needed citation. The profit per needed retrieved citation at this level is $0.52, which is close to the maximum profit

per citation of $0.58 per unit. The cost of preparing inputs increases with the number of index and search terms used, as shown in equation (6.2) and are, therefore, a minimum at XU = 1, YU = 1. The total cost, TCF, at this level is $515,470. However, minimum cost per needed retrieved citation is at XU = 4, YU = 1, where the cost is $0.66 per citation.

The feasible region is shown in Figure 7.1, where

$$1 \leq XU \leq 15$$

$$1 \leq Y \leq 9.$$

The profitable region is shown to lie in a limited part of the feasible region where

$$4 \leq XU \leq 15$$

$$2 \leq YU \leq 3.$$

Impact of Indexing and Search Terms

The results of the calculated output are provided in Figure 7.2 and 7.3. These figures present a section which runs through the points where the values of XU and YU (in integer values) show maximum profit. In Figure 7.2 the number of index terms varies, holding the number of index terms constant at YU = 2, and Part a shows the general increase in total value, cost, and profit with the number of index terms. The TVOB increases with the number of index terms as does TCF. Total profit is shown to be at a maximum at XU = 10. Part b is a graphical representation of the average value,

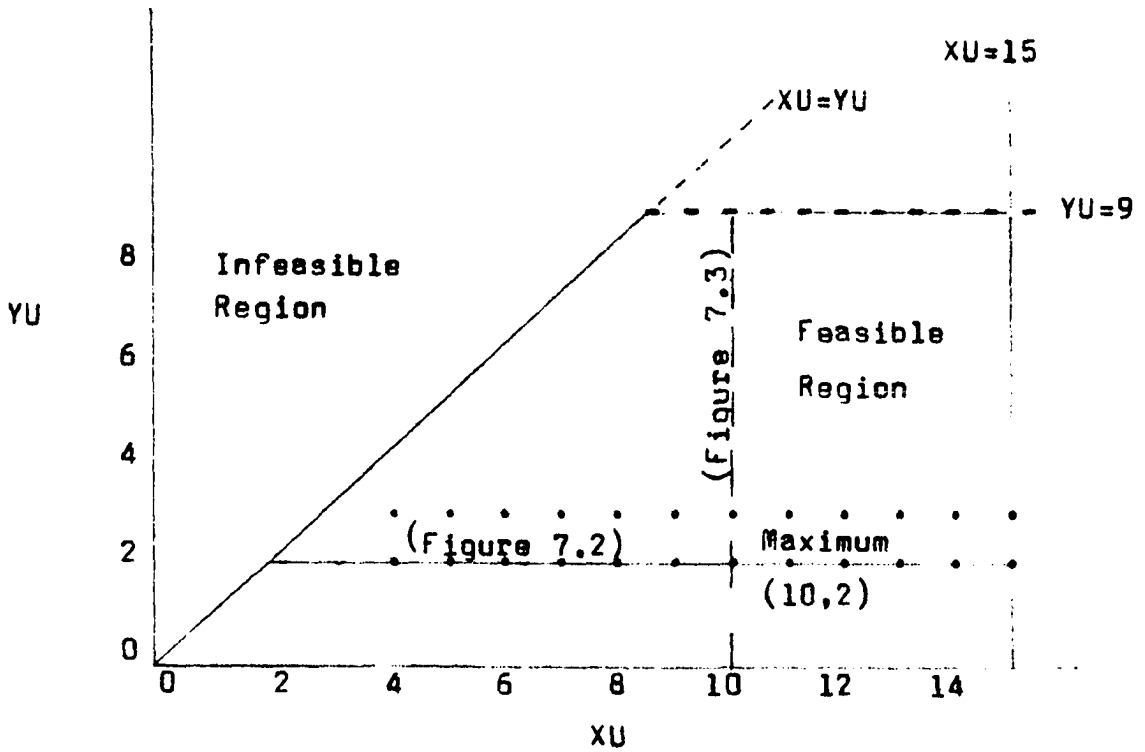


Fig. 7.1,---Plan View of Solution Profiles

Note: (·) Points of Positive Profit,
Maximum at (10,2)

---

cost, and profit per needed citation that is retrieved. These average results are plotted versus the number of index terms as in part a. As indicated, TCF/(ZNU·S) is a minimum at XU = 7 while TVOB/(TCF·S) is a maximum at XU = 8 and the average profit is maximum at XU = 8.

Reference to Figure 7.3 shows the relationships of value, costs, and profit to the number of search terms, holding the number of index terms fixed at XU = 2. Part a shows that

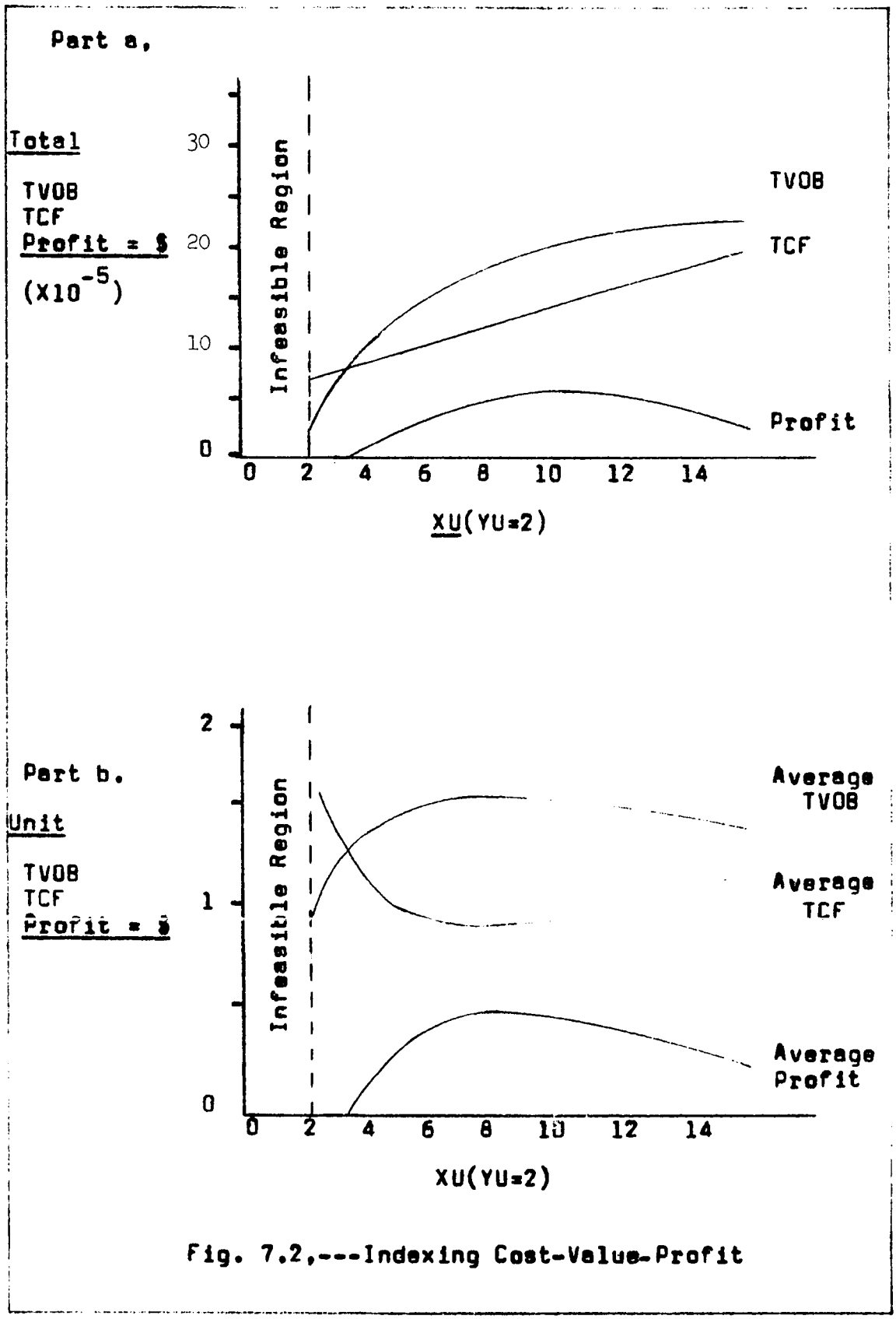


Fig. 7.2,---Indexing Cost-Value-Profit

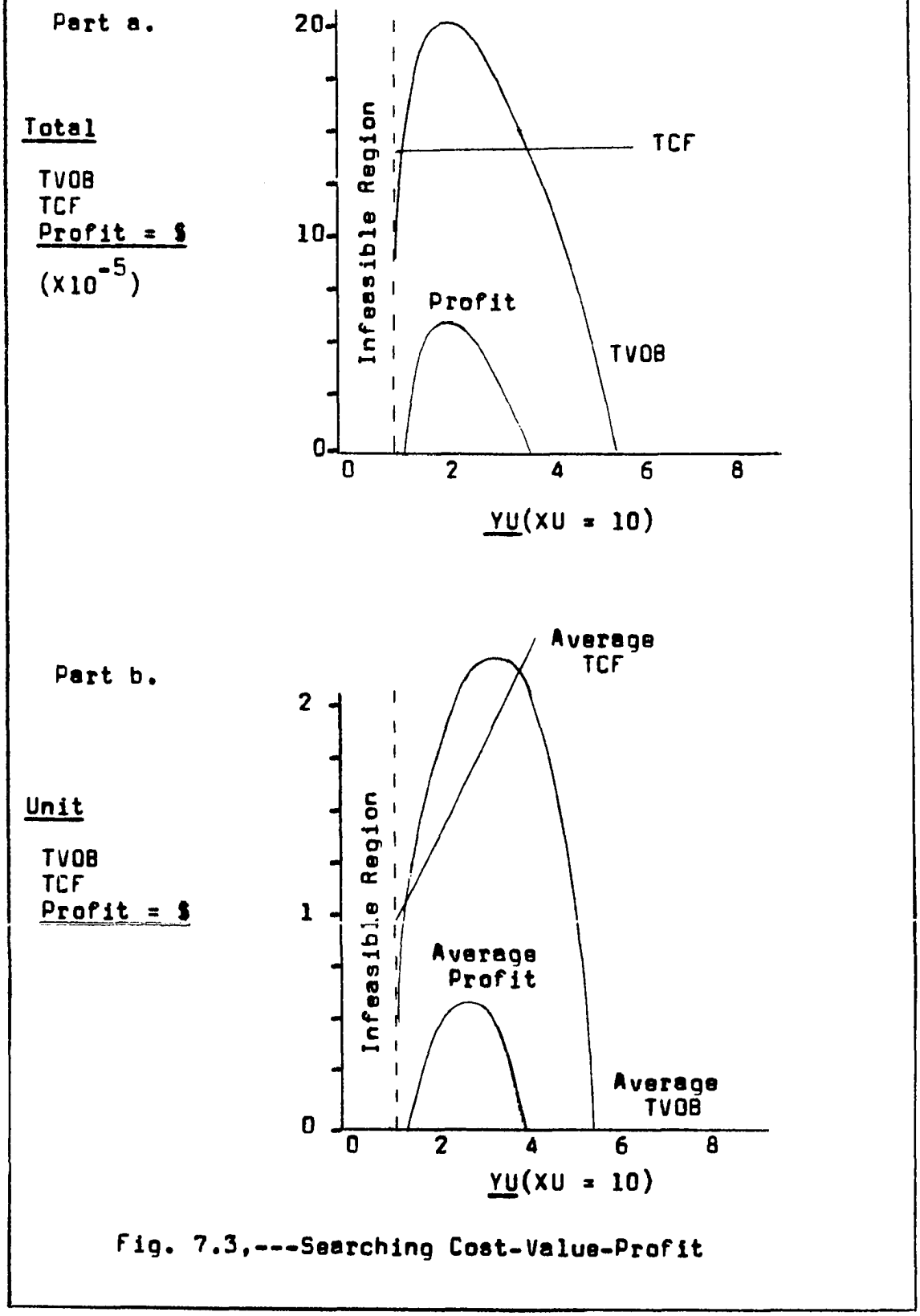


Fig. 7.3,---Searching Cost-Value-Profit

TVOB changes significantly with the number of search terms and is a maximum at YU = 2. TCF shows very little increase with the number of search terms. Total profit, therefore, is a maximum at the same value as TVOB, which is YU = 2.

Part b shows the relationship of average value, cost, and profit per needed retrieved citation, where these results are plotted versus the number of search terms. As indicated TVOB/(ZNU·S) is a maximum at YU = 3. Total cost per citation, TCF/(ZNU·S), generally increases with YU since the number of citations, ZNU, decreases with YU. Average profit is, therefore, a maximum at YU = 3, showing the interactive effect of value and cost. As noted there is a differenct relationship of values versus level of indexing as compared to that of searching.

Profit Maximizing Values

The values of profit maximization in integer values and in more precise real numbers are shown in Table 7.2. A series of points in the vicinity of the optimum value of XU and YU were evaluated. The parameters determined were cost, value, and profit both on a total and on a needed retrieved citation basis. Using a system of decreasing intervals of investigation, the calculated optimum value was found within the accuracy of the four-digit data used. The values of the various parameters are shown in Table 7.2. The maximum total profit is found to be $575,108 at XU = 9.79, YU = 2.35.

Table 7.2. Solution Cost-Value-Profit Summary

No. of Terms		No. of References Per Search[1]	Parameters of Value - Cost			Cost Per Search
			TVOB	TCF	Profit	
Index XU	Search YU	ZNU	Total Citation	Total Citation	Total Citation	
9	2	2,298	$1,822,472 1.584	$1,301,955 1.132	$520,517 0.452	$2,602
9	3	1,619	$1,740,141 2.150	$1,303,321 1.610	$436,820 0.540	$2,606
10	2	2,431	$1,924,492 1.580	$1,400,095 1.152	$524,397 0.432	$2,800
10	3	1,752	$1,842,161 2.102	$1,401,461 1.600	$440,700 0.502	$2,803
9.79	2.35	2,136	$1,956,053 1.832	$1,380,945 1.294	$575,108 0.538	$2,762

[1] 500 searches

The maximum total profit with the decision variables expressed in integers was $524,397 at XU = 10, YU = 2. In addition to the values reviewed here, this level of operation achieved the highest number of needed references retrieved per search, ZNU = 2431. The point of maximum profit per citation is located at XU = 9, YU = 3, however, this level of inputs provides the lowest level of total profit of any of the four levels depicted.

As shown in Table 7.2, the cost per search near the optimum level of indexing ranges from $2,602 to $2,803 per search. This cost is considerably higher than that charged to most users even though it is their implied cost. This aspect supports Penner's comment, discussed on page 43, that society is not paying the costs of library information services. The numerical results obtained have been rounded off to the nearest integer value since index and search terms are also integers.

The results of this approach is that the point at which profit is maximized has been quantitatively defined based on the number of index and search terms.

Experience has shown that this procedure, discussed in this section, is valid. However, this is not to be construed as a general procedure. Other types of problems can show that rounding to the nearest integer solution might not be feasible; or if it is feasible, it may be suboptimal. If the entire population space is small enough, exhaustive

enumeration of all integer feasible solutions should be used. In this particular problem, rounding of the solution XU = 9.79 and YU = 2.35 to 10 and 2 respectively did define the maximum integer solution.

CHAPTER VIII

CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER WORK

The intent of the model developed has been to devise a series of submodels that can be used to simulate the range of activities of a reference retrieval system and provide a means of optimizing this system, including the direct monetary costs and the value of the output along with the user's costs of preparing requests and determining the applicability of output.

This model is designed to simulate a system which would have its main facilities at a central location and would have a series of satellite retrieval centers. Each of these satellite centers would handle requests for retrospective searches.

The fundamental problem posed is determination of the optimum level of usage of two variables, each of which independently has the probability of omission and commission of error. A three-phase model for ascertaining the optimum level of usage of two decision variables of a reference model has been developed. The procedure for obtaining the optimum level of usage of the decision variables has been achieved. A numerical example is included to demonstrate the model's feasibility by use of data from a heuristically derived situation designed to reflect a real reference retrieval

system. The profit maximization concept is used to determine the best level at which to operate the system and to reflect all factors measured by costs: direct expenditures, operating costs, and user costs. Use of the pure competition model for profit determination precludes any variation in the value per unit of the final product, which is the number of retrieved needed references to documents. Given the physical parameters of the system in question, running the necessary experiments, ascertaining the market price of the inputs and outputs (on a unit basis), and assuming the existence of a market provide the basic conditions for application of the model.

Conclusions

The conclusions drawn from this study relate to the three phases of the model: the error detection technique, the performance model, and the optimization procedure. These conclusions are as follows:

1. A means of expressing physical quantities and costs of inputs to a reference retrieval system which correlates with the quantity and value of output has been derived by use of simulation and mathematical models.

2. Numerical values of the parameters of such a system have been derived, the model has been tested, and numerical output was achieved.

3. A means of optimizing the system, having pure competition on the purchase of factors (inputs) and sale of product (outputs), have been devised. Application of a reference retrieval system has been simulated, and numerical values that describe the level of operation of the model were calculated.

4. Numerical values of the parameters of the value of the system, cost, and profit, along with physical parameters expressed by the optimum level, have been calculated.

5. The calculations show that the maximum profit, (at an integer value, are at XU = 10, YU = 2, total profit is $524,397, with a total value of $1,924,492 and total cost of $1,400,095. However, the direct users are probably not being charged the appropriate level of costs.

6. Total output is proportional to the number of index terms and inversely proportional to the number of search terms. Final solution is XU = 10, YU = 2.

Recommendations for further work must stem from a review of the limitations of the current model, then putting these limitations into a constructive framework for analysis and possible implementation.

A review of the limitations of the present total cost-value system and discussion of a means of coping with them can be presented by technical specialty or by inspection of the sequential phases of the model. This sequential review of the phases starts with the error detection technique, followed by the performance model and the optimization procedure.

Error Detection Technique Limitations

The errors of indexing and retrieval stem from errors in language itself and the inconsistencies in the application of particular terms to describe specific facet(s) of a document. One significant limitation is the difficulty to distinguish between the errors of language and the errors

generated by indexers. Efforts to improve the efficiency of the systems rest on the ability to improve the quality of indexing and searching, especially the first few terms. In turn this action necessitates improving the index-search term vocabulary. A major benefit derived from the more efficient usage of terms would be a reduction in the number of terms used. This model has assumed that the individuals reviewing the indexing and searching do not commit errors. Also the error for each sequential term used has been expressed deterministically. These data were generated by a series of repeated experiments. These sample data were then converted to the best estimator of the parameter, $\bar{P}$. Another limitation is that the number of index and search terms is fixed.

These limitations could be removed by modification of the performance model so that much of the data would be stochastic, such as the value of the probability distribution of each of the index terms and also the number of terms used. The search terms can be handled similarly. Another possible variation includes removing the assumption that the indexing evaluator be considered as a final authority. Instead this determination of the applicability of terms would be treated stochastically as to their certainty. The number of terms needed to express fully all aspects of the concept, assuming that a level of no error of commission has been committed, could be treated as a variable rather than as a constant.

A formulated schedule of variations that can be considered in the error detection procedure follows.

Indexing and Searching. The various concepts that are applicable to indexing are developed.

1. The number of concepts to be recognized are

   a. the number of concepts to recognize is specified by management,

   b. the number of terms needed to present all aspects of these concepts is a function of the indexing vocabulary. [Also the length of document, coverage, type of report (analytic descriptive, derivative application)],

   c. the procedure for implementing this aspect could be derived in the error determination technique.

2. Under the conditions of a and b above, the number of terms actually used can then be limited, prescribed by management. This application could be expressed by using a probabilistic approach with the following procedure.

   a. There would be no overlap of information, and each term would contribute the same amount of information, which would be unique,

   b. Assume mutually exclusive but varying amounts of information for each term but the information carried would be unique.

   c. Assume independent, but not mutually exclusive, content for each term. Each term would contain a finite amount of information. After application of the first term, the marginal contribution of each term would decrease.

   d. Assume that each term is independent, not mutually exclusive. Each term would be carrying varying amounts of information. However, in application of a term to indexing, the probability of marginal contribution of each additional term would decrease.

e. Assume that each term is independent, not mutually exclusive. Application of some one term would imply the application of other associated terms, each one contributing some unique portion of information. These terms could have

   1) equal contribution or
   2) unequal contribution.

f. Assume that each term is independent, not mutually exclusive. Application of some one term would imply the application of these associated terms. The probability of marginal contribution of each term would decrease with the number of terms added.

Error Detection. Error determination can be developed further as discussed below.

1. Presently, assuming perfect evaluation, relaxing of this assumption is a realistic consideration. Significant work concerning this relaxation has been done, but it is not presented in a manner that allows direct application to the error detection model.

2. Presently, the system is presumed to operate error free in this area. All of these errors have some finite probability of occurrence. Other errors can be grouped as follows:

   a. no clerical error,
   b. no false drops, output is adequate,
   c. no mechanical errors, output available, output readable.

Performance Model

Limitations of the performance model relate to the lack of means of formulation of searches in an algebraic manner and with dendritic terminology controls.

Formulation of searches in an algebraic manner has been restricted to intersection of terms as shown in equation

(3.5). The means of expressing a union search or negations have not been developed in this model although it exists in real reference retrieval systems. Therefore, searches based on any combinations of intersection, union, and negations can not be handled presently. Procedures for handling mutually exclusive terms, in addition to the prescribed independently distributed terms, are desired. Also a procedure for formulating searches using a combination of intersection and union of search terms would be useful. In addition formulation of searches using the negation of terms, (A•B - A•B•C) is recognition of the real state of the art.

A more systematic procedure for handling the variations in ranking, usage, and misuage of index versus search terms, by relaxing the concepts discussed in Chapter III under vocabulary usage errors, has application.

The use of a dendritic terminology control with the application of levels of authority of terms is currently not feasible. Lack of this feature restricts the applicability of the model to the "real world."

Optimization Procedure

Several factors limit the optimization stage of the system, including some features with overall limitations; time dependent factors and value-cost considerations.

Time Dependent Factors. Time impinges on information systems in the form of growth of the document collection

size, which is caused by an increase in the rate of publishing. In addition, if growth in the basic file of references is considered, the level of the number of concepts that are recognized could increase. This increase in the number of file items would necessitate an increase in the number of terms used in indexing, with or without an increase in recognition of the number of concepts in the document. An increase in the number of terms in the vocabulary would be a prerequisite for increased file volume of references to documents.

Another aspect of time is the time lag between receipt of a search request from a user and the return of the file of citations of references obtained. Reducing this time lag will reduce the loss in value of the information to the user, but it will normally increase the cost of operating the system.

Value-Cost Considerations. The effect of time on value of information has been discussed, however; the means of determining value itself expecially as it relates to the number of references retrieved in a single search have not. The basis for and the extent of decreasing marginal returns of value would be desirable. Also much work has been discussed by various authors concerning the need for cost data, but Helmkamp (see page 42) and Penner (see page 43) indicate limited awareness of the need for appropriate costs data, and there are few procedures for obtaining such data.

The expression of the concept of value has been based on a pure competition model which produces a constant price per unit of output or per reference to a document. Other approaches would be to consider other economic models having a deterministic solution. Also models having a stochastic price determination would be evolved. Application of utility concepts to depict measure value is another approach for measuring the value of output.

The negative value assigned to nonrecalled desired references could be similarly handled.

Specific Suggestions For Further Work

This sequential review of the total system has described several features that could be implemented so that the model can be used to represent more readily "real life" reference retrieval systems. The features that present the most favorable areas of research are listed below in preferential order.

1. A procedure for formulation of searches in manners other than intersection is needed. This formulation could include a combination of intersection and union, negations, and the usage of dendritic terminology control with the application of levels of authority of terms.

2. The inclusion of growth factors in the file of indexed documents and vocabulary size is needed, and would include the number of index and search terms.

3. Significantly more work is needed in the cost-value area to identify the relevant items based on the physical basis and cost factors. The first consideration is with physical units which

can be determined as follows:

a. Ascertain the underlying factors that control the physical costs of the system,

b. Obtain measures of quantification of these factors,

c. Present these quantified values so that they can be used as predictors for the system and thus provide a more comprehensive series of production functions.

4. A means of vocabulary control and upgrading is needed.

APPENDIX

The details of the computer program used in Chapter IV, including the program and the description and procedures for its use, have been documented.

These documented proceedings, which are not included in this report, have been deposited in report form with the library of the School of Industrial Engineering of the University of Oklahoma, title "Reference Retrieval Simulation Model," TR 73-1

REFERENCES

1. Luhn, H. P. "A Statistical Approach to Mechanized Literature Searching." IBM Research Center, Research Paper RC-3, January 30, 1957.

2. Lancaster, F. Wilfrid. Information Retrieval Systems. New York: John Wiley & Sons, Inc., 1968.

3. Sharp, John R. Some Fundamentals of Information Retrieval. New York: London House & Maxwell, 1965.

4. Murdock, John W., and Liston, David M., Jr. "A General Model of Information Transfer: Theme Paper 1968 Annual Convention." American Documentation. Vol. 18, No. 4 (October, 1967), pp. 197-208.

5. Jahoda, G. "Correlative Indexing Systems for the Control of Research Records." Columbia University D. L.S. thesis, 1960, University Microfilms, Mic. 60-3082.

6. Landry, Bertrand C., and Rush, James E. "Toward a Theory of Indexing ...II." Journal of the American Society for Information Science. Vol. 21, No. 5 September - October, 1970), pp. 358-367.

7. Zipf, George Kingsley. Human Behavior and the Principle of Least Effort. Cambridge, Mass.: Addison-Wesley Press, Inc., 1949.

8. Houston, Nona, and Wall, Eugene. "The Distribution of Term Usage in Manipulative Indexes." American Documentation. Vol. 15 (April, 1964), pp. 105-114.

9. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

10. Morse, Philip M. Library Effectiveness: A Systems Approach. Cambridge, Mass.: The M.I.T. Press, 1968.

11. Raver, Norman. "Performance of IR Systems." Information Retrieval. Edited by George Schecter. Washington, D.C.: Thompson Book Company, 1967, pp. 131-142.

12. Long, John M.; Barnhard, Howard J.; and Levy, Gertrude C. "Dictionary Buildup and Stability of Word Frequency in a Specialized Medical Area." American Documentation. Vol. 18, No. 1 (January, 1967), pp. 21-25.

13. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

14. Houston, Nona, and Wall, Eugene. "The Distribution of Term Usage in Manipulative Indexes." American Documentation. Vol. 15 (April, 1964), pp. 105-114.

15. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

16. Leimkuhler, Ferdinand F. "Systems Analysis in University Libraries." College and Research Libraries. Vol. 27, No. 1 (January, 1966), pp. 13-18.

17. Fussler, Herman; Simon, H.; and Julian, L. Patterns in the Use of Books in Large Research Libraries. Chicago: The University of Chicago Library, 1961.

18. Jain, A. K. "Sampling and Short-Period Usage in the Purdue Library." College and Research Libraries. Vol. 27, No. 3 (May, 1966), pp. 211-218.

19. Morse, Philip M. Library Effectiveness: A Systems Approach. Cambridge, Mass.: The M.I.T. Press, 1968.

20. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

21. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

22. Uhlmann, Wolfram. "Documents, Specification and Search Strategy Using Basic Intersections and the Probability Measure of Sets." American Documentation. Vol. 19 (July, 1968), pp. 240-246.

23. Raver, Norman. "Performance of IR Systems." Information Retrieval. Edited by George Schecter. Washington, D. C.: Thompson Book Company, 1967, pp. 131-142.

24. Perry, James W.; Kent, Allen; and Berry, Madeline M. Machine Literature Searching. New York: Interscience Publishers, 1956.

25. Cleverdon, Cyril W. Aslib Cranfield Research Project. Report on the first stage of an Investigation into the Comparative Efficiency of Indexing

Systems. Cranfield, England: The College of Aeronautics, 1960.

26. Cleverdon, Cyril W. Aslib Cranfield Research Project. Interim Report on the Test Programme of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1962.

27. Cleverdon, Cyril W. Aslib Cranfield Research Project. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1962.

28. Lancaster, F. Wilfrid. Evaluation of the MEDLARS Demand Search Service. Bethesda, Maryland, National Library of Medicine, January, 1968. Report No. PB 178-660.

29. Montague, Barbara A. "Testing, Comparison and Evaluation of Recall, Relevance, and Cost of Coordinate Indexing with Links and Roles." American Documentation. Vol. 16, No. 3 (July, 1965), pp. 201-8.

30. Verhoeff, J. W.; Goffman, W.; and Belzer, J. "Mathematical Models in Systems Design for Information Retrieval." Theoretic Note No. 1, May 15, 1961, AFOSR Contract No. AF 49(683)-357.

31. Swets, John A. "Information Retrieval Systems." Science. Vol. 141 (July, 1963), pp. 245-250.

32. Swets, John A. "Information Retrieval Systems." Science. Vol. 141 (July, 1963), pp. 245-250.

33. Salton, G. "The Generality Effect and the Retrieval Evaluation for Large Collections." Journal of the American Society for Information Science. Vol. 23, No. 1 (January - February, 1972), pp. 11-22.

34. Bourne, C. P.; Peterson, G. D.; Lefkowitz, B.; and Ford, D. "Requirements, Criteria and Measures of Performance of Information Storage and Retrieval Systems." Stanford Research Institute, Menlo Park, California. Office of Science Information Service, Report No. AD 270 942, SRI Project No. 3741.

35. Pollock, Stephen M. "Measures for the Comparison of Information Retrieval Systems." American Documentation. Vol. 19 (October, 1968), pp. 387-397.

36. Luhn, H. P. "A Statistical Approach to Mechanized Literature Searching." IBM Research Center, Research Paper RC-3, January 30, 1957.

37. Wall, Eugene. "A Rationale for Attacking Information Problems." American Documentation. Vol. 18, No. 2 (April, 1967), pp. 97-103.

38. Cleverdon, Cyril W. Aslib Cranfield Research Project. Report on the first stage of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1960.

39. Cleverdon, Cyril W. Aslib Cranfield Research Project. Interim Report on the Test Programme of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1960.

40. Cleverdon, Cyril W. Aslib Cranfield Research Project. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1962.

41. Lancaster, F. Wilfrid. Evaluation of the MEDLARS Demand Search Service. Bethesda, Maryland, National Library of Medicine, January, 1968. Report No. PB 178-660.

42. Saracevic, et al. An Inquiry into Testing of Information Retrieval Systems. Part 1: Objectives, Methodology, Design, and Controls. Comparative Systems Laboratory Final Technical Report. Center for Documentation and Communication Research Case, Western Reserve University, 1968.

43. Aitchison, Jean and Cleverdon, Cyril. Aslib Cranfield Research Project. Report on a Test of the Index of Metallurgical Literature of Western Reserve University. Cranfield, England: The College of Aeronautics, 1963.

44. Lancaster, F. Wilfrid. Evaluation of the MEDLARS Demand Search Service. Bethesda, Maryland, National Library of Medicine, January, 1968. Report No. PB 178-660.

45. Saracevic, et al. An Inquiry into Testing of Information Retrieval Systems. Part 1: Objectives, Methodology, Design, and Controls. Comparative Systems Laboratory Final Technical Report. Center for Documentation and Communication Research Case, Western Reserve University, 1968.

46. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

47. Landau, Herbert B. "The Cost Analysis of Document Surrogation: A Literature Review." American Documentation. Vol. 20, No. 4 (October, 1969), pp. 302-309.

48. Lancaster, F. Wilfrid. "The Cost-Effectiveness Analysis of Information Retrieval and Dissemination Systems." Journal of the American Society for Information Science. (January-February, 1971), pp. 12-27.

49. Keith, Nathan R., Jr. "A General Evaluation Model for An Information Storage and Retrieval System." Journal of the American Society for Information Science. (July-August, 1970), pp. 237-239.

50. Block, U. and Ofer, K. D. "Experiments With an SDI System." Mechanized Information Storage, Retrieval and Dissemination. North-Holland, 1968, pp. 326-334.

51. Bourne, C. P.; Peterson, G. D.; Lefkowitz, B.; and Ford, D. "Requirements, Criteria and Measures of Performance of Information, Storage and Retrieval Systems." Stanford Research Institute, Menlo Park, California, AD 270 942, SRI Project No. 3741.

52. Bourne, Charles P., and Ford, Donald F. "Cost Analysis and Simulation Procedures for the Evaluation of Large Information Systems." American Documentation. Vol. 15 (April, 1964), pp. 143-149.

53. Marron, Harvey, and Snyderman, Martin, Jr. "Cost Distribution and Analysis in Computer Storage and Retrieval." American Documentation. Vol. 17, No. 2 (April, 1966), pp. 89-95.

54. Marron, Harvey, and Snyderman, Martin, Jr. "Cost Distribution and Analysis in Computer Storage and Retrieval, II." American Documentation. Vol. 18, No. 3 (July, 1967), pp. 162-164.

55. Kuney, J. H. "Computer Typesetting for Scientific Publications." Mechanized Information Storage, Retrieval and Dissemination. North-Holland. 1968, pp. 510-528.

56. Stanwood, R. H. "The Cost of a Computerized Information Retrieval System." Mechanized Information Storage, Retrieval and Dissemination. North-Holland. 1968, pp. 391-406.

57. Cleverdon, Cyril W. Aslib Cranfield Research Project. Report on the first stage of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1960.

58. Cleverdon, Cyril W. Aslib Cranfield Research Project. Interim Report on the Test Programme of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1960.

59. Cleverdon, Cyril W. Aslib Cranfield Research Project. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1962.

60. Barish, Norman N. Economic Analysis for Engineering and Managerial Decision-Making. New York: McGraw-Hill Book Company, Inc., 1962.

61. Montague, Barbara A. "Testing, Comparison, and Evaluation of Recall, Relevance, and Cost of Coordinate Indexing with Links and Roles." American Documentation. Vol. 16, No. 3 (July, 1965), pp. 201-208.

62. Overmyer, La Van. "An Analysis of Output Costs and Procedures for an Operational Searching Service." American Documentation. Vol. 14, (April, 1963), pp. 123-142.

63. Costello, John C., Jr. Coordinate Indexing. Vol. VII, Rutgers Series on Systems for Intellectual Transmission of Information. Edited by Susan Artandi, Graduate School of Library Science, Rutgers, The State University, New Brunswick, New Jersey: 1966.

64. Johnson, R. R. "Needed: A Measure for Measure." Datamation. Vol. (December, 1970), pp. 22-30.

65. Helmkamp, John G. "Managerial Cost Accounting For a Technical Information Center." American Documentation. Vol. 20, No. 2 (April, 1969), pp. 111-118.

66. Penner, Rudolf J. "The Practice of Charging Users For Information Services: A State of the Arts Report." Journal of the American Society For Information Science. (January-February, 1970), pp. 67-74.

67. Rogers, Frank B. "The Development of MEDLARS." Medical Library Association Bulletin. Vol. 52, No. 1 (January, 1964), pp. 150-151.

68. Rogers, Frank B. "MEDLARS Operating Experience at the University of Colorado." Medical Library Association Bulletin. Vol. 54, No. 1 (January, 1966), pp. 1-10.

69. Rogers, Frank B. "MEDLARS Operating Experience: Addendum." Medical Library Association Bulletin. Vol. 54, No. 4 (October, 1966), pp. 316-320.

70. "Update of Unit: Cost of MEDLARS." Bulletin from the Dept. of Health, Education and Welfare, National Library of Medicine.

71. Cummings, Martin M. "Needs of the Health Services." Electronic Handling of Information: Testing and Evaluation. Washington, D. C.: Thompson Book Company, 1967, pp. 13-23.

72. Personal communication from William H. Caldwell, Deputy Chief, Bibliographic Services Division, National Library of Medicine, dated November 17, 1969.

73. Niland, Powell. "Developing Standards for Library Expenditures." Management Science. Vol. 13, No. 12 (August, 1967), pp. B797-B808.

74. Mueller, Max W. "Time, Cost and Value Factors in Information Retrieval." Operations Research Division, Lockheed Aircraft Corporation, California Division.

75. Good, I. J. "The Decision-Theory Approach to the Evaluation of Information-Retrieval Systems." Information Storage and Retrieval. Oxford, England: Pergamon Press, 1967, Vol. 3, pp. 31-34.

76. Emery, James C. Organizational Planning and Control System: Theory and Technology. Arkville Press, 1969.

77. Gotterer, Malcolm H. "Identification of Performance Criteria of an Electronic Information System." Electronic Handling of Information: Testing and Evaluation. Edited by A. Kent and others. Washington, D. C.: Thompson Book Company, 1967, pp. 51-62.

78. Bryant, E. C. "Modeling in Document Handling." Electronic Handling of Information: Testing and Evaluation. Edited by A. Kent and others. Washington, D. C.: Thompson Book Company, 1967, pp. 163-173.

79. Churchman, C. West; Ackoff, Russell L.; and Arnoff, E. Leonard. Introduction to Operations Research. New York: John Wiley & Sons, Inc., 1957.

80. Mueller, Max W. "Time, Cost and Value Factors in Information Retrieval." Operations Research Division, California Division, Lockheed Aircraft Corporation.

81. Box, G. E. P., and Hunter, J. S. "Condensed Calculations for Evolutionary Operation Programs." Technometrics. Vol. 1, No. 1 (February, 1959), pp. 78-95.

82. Carlisle, Donald. "The Economics of a Fund Resource With Particular Reference to Mining." American Resource Review. Vol. 44 (September, 1954), pp. 595-616.

83. Cleverdon, Cyril W. Aslib Cranfield Research Project. Report on the first stage of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1960.

84. Cleverdon, Cyril W. Aslib Cranfield Research Project. Interim Report on the Test Programme of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1960.

85. Cleverdon, Cyril, W. Aslib Cranfield Research Project. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, England: The College of Aeronautics, 1962.

86. Lancaster, F. Wilfrid. Evaluation of the MEDLARS Demand Search Service. Bethesda, Maryland, National Library of Medicine, January, 1968. Report No. PB 178-660.

87. Raver, Norman. "Performance of IR Systems." Information Retrieval. Edited by George Schecter. Washington, D. C.: Thompson Book Company, 1967, pp. 131-142.

88. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

89. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

90. Centralization and Documentation. Final Report to the National Science Foundation. Report No. C-64469, July, 1963. Little (Arthur D.) Inc.

91. Houston, Nona, and Wall, Eugene. "The Distribution of Term Usage in Manipulative Indexes." American Documentation. Vol. 15 (April, 1964), pp. 105-114.

92. Costello, John C., Jr. Coordinate Indexing. Vol. VII. Rutgers Series on Systems for Intellectual Transmission of Information. Edited by Susan Artandi, Graduate School of Library Science, Rutgers, The State University, New Brunswick, New Jersey, 1966.

93. Barish, Norman N. Economic Analysis for Engineering and Managerial Decision-Making. New York: McGraw-Hill Book Company, Inc., 1962.