DOCUMENT RESUME

ED 057 843 LI 003 353

AUTHOR	Landry, Bertrand Clovis
TITLE	A Theory of Indexing: Indexing Theory as a Model for Information Storage and Retrieval.
INSTITUTION	Ohio State Univ., Columbus. Computer and Information Science Research Center.
SPONS AGENCY	National Science Foundation, Washington, D.C. Office of Science Information Services.
REPORT NO	OSU-CISRC-TR-71-13
PUB DATE	Dec 71
NOTE	282p.; (1,034 References)
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	*Indexing; *Information Retrieval; *Information Storage; *Information Theory; Models; *Relevance (Information Retrieval)
IDENTIFIERS	*Indexing Theory; Information Transfer

ABSTRACT

Present day shortcomings in information retrieval are the results of a failure to properly contend with the problem of data representation. The index provides the necessary linkage between a multiplicity of sources and a single receiver. Whether considering the source/document-space interface or the query/index interface, the elements of the underlying communication phenomena are the same: sets of documents, sets of attributes, and sets of relations expressing a connection between documents and attributes. The essential operation of the indexing system is the creation of a representation of the document space. The analysis -- document transformations and the final index-query transformations are shown to be, respectively, a prerequisite to, and function of, the document space representation. The operating characteristics of the indexing system are modeled by means of the index space. From a different point of view, the concept of error, organization, information and search are introduced through a consideration of the indexing process as a thermodynamic system. Thus, indexing is viewed as an order-increasing operation that identifies common data elements and relations between data elements present in the input document stream. (Author/MM)

(OSU-CISRC-TR-71-13)

A THEORY OF INDEXING: INDEXING THEORY AS A MODEL FOR INFORMATION STORAGE AND RETRIEVAL

by

Bertrand Clovis Landry

Work performed under

Grant No. 534.1, National Science Foundation

Computer and Information Science Research Center
The Ohio State University
Columbus, Ohio 43221

December 1971

PREFACE

This work was done in partial fulfillment of the requirements for a doctor of philosophy degree in Computer and Information Science from The Ohio State University. It was supported by Grant No. GN 534.1 from the Office of Science Information Service, National Science Foundation, to the Computer and Information Science Research Center of The Ohio State University.

The Computer and Information Science Research Center of The Ohio State University is an interdisciplinary research organization which consists of the staff, graduate students, and faculty of many University departments and laboratories. This report is based on research accomplished in cooperation with the Department of Computer and Information Science.

The research was administered and monitored by The Ohio State University Research Foundation.

ACKNOWLEDGMENTS

I am indebted to my good friend and advisor Professor James E. Rush, not only for his help with the preparation of this work, but for his encouragement and direction throughout my years of graduate study. I should also like to extend my thanks to Professor Harold B. Pepinsky for a very thorough reading of Chapter V -- as always, his comments and suggestions have been very helpful. I am appreciative of Dr. Naomi M. Meara's help with the design and subsequent testing of the Extended Bruner Experiment. I am also grateful to Professors Marshall C. Yovits and Ronald L. Ernst for serving as members of the committee who read this dissertation. I am also indebted to Mrs. Mary Kimball for her help in transforming the manuscript of this dissertation to typed copy.

Partial support of this work has been provided by a grant (GN-534.1) from the National Science Foundation to the Computer and Information Science Research Center, through a Title II-b Fellowship in Library and Information Science awarded by the Office of Education and through a Graduate Fellowship awarded by the I.B.M. Corporation.

Finally, I extend my gratitude to my wife, Diane, for not only providing moral support, but for learning to contend with my world of *data* and *information*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

QUOTATION REFERENCES

PAGE	
xii	G. Bachelard, *La Psychanalyse du Feu*, Collection Psychologie, NRF, Gallimard, Paris, 1938, 50.
1	J. R. R. Tolkien, *The Two Towers*, Houghton Mifflin Co., New York, N.Y., 1965, 260.
13	A. Kaplan, *The Conduct of Inquiry*, Chandler Publishing Co., New York, N.Y., 1964, 268.
26	M. G. Mellon, *Chemical Publications: Their Nature and Use*, McGraw-Hill Book Co., New York, N.Y., 1965, 1.
39	R. L. Collison, *Indexes and Indexing*, John de Graff Inc., New York, N.Y., 1956, 126.
39	A. Waley (translator), *The Analects of Confucius*, G. Allen & Unwin, Ltd., London, 1956.
116	M. G. Mellon, *op. cit.*, 17.
142	C. Dickens, *The Pickwick Papers*, The New Oxford Illustrated Dickens, Oxford Press, Oxford, 1956, 719.
219	L. Carroll, *Through the Looking Glass*, Random House, New York, N.Y., 1946.
232	J. Howell, *Discourse Concerning the Precedency of Kings*, Printed for Rowland Reynolds, London, 1664, 219.

"On ne peut étudier que ce qu'on a d'abord rêvé. La science se form plutôt sur une reverie que sur une éxperience et il faut bien des éxperiences pour effacer les brumes du songe."

Gaston Bachelard

CHAPTER I. INTRODUCTION

> 'But I should like to know...' Pippin began. 'Mercy,' cried Gandalf. 'If the giving of information is to be the cure of your inquisitiveness, I shall spend all the rest of my days answering you.'
>
> J. R. R. Tolkien, *The Two Towers*

Pippin is a Hobbit and it is well known, at least Tolkien tells us so, that Hobbits are, by nature, a very inquisitive people. At the slightest provocation they will produce a barrage of questions that will eventually dull even the sharpest of minds. In the brief exchange quoted above, Tolkien has identified (perhaps unknowingly) several important issues that are worthy of consideration. Like Hobbits, we certainly would want to learn more about the meaning of the following words and phrases: "I should like to know," "information," "inquisitiveness," and "spend all the rest of my days in answering you." Somehow the ideas these terms convey are all vaguely familiar since we have all experienced the need to know something and, sometimes, we have actually received answers to our questions. Perhaps the most troublesome point rests with the concept of *spending all the rest of one's days in answering*. Is it possible that "information" is indeed an unlimited qunatity--a resource beyond measure? Already we have started to ask questions.

Most authors begin a treatise dealing with topics in Information Storage and Retrieval with an authoritative and somewhat threatening statement concerning the "information explosion." Often, this term is also used as a convenient catch-all phrase designed to suggest knowledge of the IS&R field. We shall attempt to be somewhat more careful and confine our remarks to things that we "know for sure." For instance, from research in *cognitive dissonance*

(see Weick [1]) it is known that the value placed on a message is directly related to the magnitude of the effort required to understand it. Well, we do know that people have worked hard at *trying* to understand the message called *information explosion*, but as Hobbits, we would certainly want still to ask several questions because we feel we know very little about the information phenomenon.

From all appearances, our society has evolved to a state of dependence on the recorded message. Thus, instead of dealing with actual experiences, we manipulate facsimiles of them. Manipulation of such facsimiles is an activity called information storage and retrieval. Bar-Hillel [2] provides a working definition of the problem central to information storage and retrieval (IS&R):

> Assuming that there exists somewhere a body of recorded knowledge--in technical terms, a collection of documents--and assuming that someone has a certain problem for the solution of which this collection might contain pertinent material, how shall he decide whether there are in fact documents in this collection that contain such pertinent material, and, if so, how shall this material be brought to his attention?

In this chapter we shall consider the nature of the process of "bringing material to one's attention". We will find it helpful to ask questions about what we *seem* to know about this process, the origins of the need for IS&R and, finally, the fundamental nature of the main problem of IS&R. Partial answers to these questions will be provided through a consideration of both a schematic for the IS&R process and directions for further research.

1. The "Information Explosion"

Bonnard [3], commenting on the rapid growth of source materials opines: "The library grew not only because classical works were bought, but because of the extraordinary prolificacy of contemporary authors." This statement

accurately expresses present-day trends, but Bonnard refers, not to a growing modern metropolitan library, but to the great literary expansion that took place during the Hellenistic period of ancient Greece. People have been, and will no doubt continue to be, concerned about the rapid proliferation of documents. Obviously, the concept of an "information explosion" is not a new one.

Another device that is frequently used to highlight the "information explosion" is the figure depicting the exponential growth rate of the literature of various disciplines. One source [4] has estimated that there are recorded, in one form or another, 10 trillion alphanumeric characters. Furthermore, this collection appears to be growing at a rate of about 10 billion characters every twelve years. Based on the estimates, it is understandable that the number of scientific journals has grown to over 100,000 during the last 300 years [5]. These figures make it easy to envision bleary-eyed researchers attempting diligently to read through a rapidly growing mountain of reports and data. It is senseless to dispute any of the data concerning the *growth* of documentation. Rather, let us consider briefly some of the conditions present in both society and science that have caused this "explosion."

There are, no doubt, numerous phenomena that, in some way, have contributed to the growth of the store of recorded materials; however, we shall restrict our attention to a brief outline of just six basic factors (adapted in part from Mikhailov [6]):

- The shift from folklore to the written tradition.

In Medieval Europe, most history, literature and tradition was transmitted from generation to generation by oral means (songs, tales, *etc.*). The invention of the Gutenberg press brought on an almost complete reliance on the printed word as the vehicle for communication. A collection of "records" was no longer limited to the confines of human memory.

- The increased popularity and application of the scientific method.

Following the Renaissance, the scientific method became the dominant philosophy of the Western World. The rapid growth of the various "sciences," the emphasis placed upon theory and the need for experimentation, have all interacted to increase the amount of data that must be recorded, stored and communicated.

- An increase in resources expended in discovery.

Written scientific communication has increased simply because of the increase in the number of people involved in research. The expenditure of other resources (as reflected in costs) has given rise to the need for numerous "progress" and "justification" reports.

- The decrease in time lag between discovery and application.

This is a reflection of the rate of "progress" and, in terms of documentation, means more papers, reports, patents, abstracts and other types of documents.

- The increased need for reliable decision-making information.

Complexity in science, government, industry and society in general creates a need for "processed" data for decision-making.

- The increase in the amount of data resulting from scientific experiments.

Because of the technology which research makes possible, individual experiments can be made to yield extremely large amounts of recorded data.

From a brief consideration of these six factors one may conclude that any problems associated with the storage and retrieval of data arise not as a consequence of a sudden "information explosion" but as a consequence of a normal growth process. Quantity is certainly one of the leading problems in IS&R; every discipline is confronted with so many recorded documents (on paper, magnetic tape, film and other media) that they defy organization, manipulation and retrieval. However, it should be clear that problems that we experience today are really a consequence of a prolonged history of disorganization, a lack of planning and (of greatest importance) a fundamental misunderstanding of the techniques of data organization. These problems are only beginning to be dealt with in the field of IS&R.

Perhaps the toughest problem in IS&R centers on data representation. If vast collections of data are to be used and used effectively (*i.e.*, incorporated into the processes of the scientific method and decision making), then they must be amenable to accurate and complete searching. But the value to be derived from accurately-represented and well-organized data hinges on the assumption that, if the searcher is able to make effective use of these collections, a costly duplication of effort can be avoided. Fugmann's [7] estimate that almost 30% of the work done in chemistry is a duplication of previous work suggests that there is much work still to be done in IS&R.

Furthermore one might ask what it would profit him to turn to data retrieval systems for answers to his queries when 30% of the world's documents dealing with *phenothiazines* (for instance) are misindexed [8] by such systems? As a result of all of this, most researchers are perceptive and pragmatic about their information problems. Their information gathering procedures follow these steps (Mellon [9]):

> ... first, by inquiring of the individual who knows; second, by performing the experimental investigations necessary to ascertain the desired facts; and, third, by consulting the scientific literature, where a record may be found of the published reports of others' work on the subject in question.

In the face of all of these problems, scientists attempt to reduce the burdens of communication by becomming more specialized. While specialization in itself is useful and probably is a logical outgrowth of increased scientific activity [10], it creates the possibility of intellectual isolation. This means that the researcher may increasingly fail to become aware of significant work, carried out in other disciplines, that may impinge directly upon his own efforts. Thus Information Science, and more specifically IS&R, must not only find ways of dealing with a growing collection of documents but, more importantly, must find ways of overcomming the growing isolation of scientific disciplines.

2. The Information Retrieval Process

Numerous solutions have been offered of the growing information transfer problem. A consequence of these "solutions" is the current trend of moving responsibility for effective scientific communication away from the librarian and into the hands of system designers. This trend has resulted in the development of a profusion of information storage and retrieval systems,

each designed to solve the information transfer problem. These systems, whether manual or automatic, all attempt to provide searchers with answers to the question: "I should like to know...".

Information storage and retrieval systems can be described in the terminology of Marschak [11] as a combination of two *purposive processing chains*. These two chains are depicted in Figure 2.1; I have chosen to call them *document processing* and *retrieval processing*. Both processes, which are greatly simplified in this figure, actually involve multi-level and multi-step operations designed to expedite the transfer of data. The document processing chain shows the flow from document creation to document acquisition (by the IS&R system), representation and document storage. The first three stages are paralled by the retrieval processing chain in the conception (realization) of the information need, the clarification of the request, and the representation (coding) of the request. Both chains share, through the representation stage, the operations of selection, content analysis, indexing and coding. Document storage involves, in addition, the process of accumulation. The two processing chains merge at the searching operation where retrieved data (potentially information) are disseminated for evaluation. The dotted lines in the figure indicate the possibility of repeated cycling through the retrieval process.

The successful merging (in terms of answered questions) of these two chains depends on the achievement of common understanding [12] between the mechanics of the storage system and the actions of the searcher. Operationally, common understanding is only made evident through the success of the data searcher. But theoretically, common understanding can be evaluated in terms

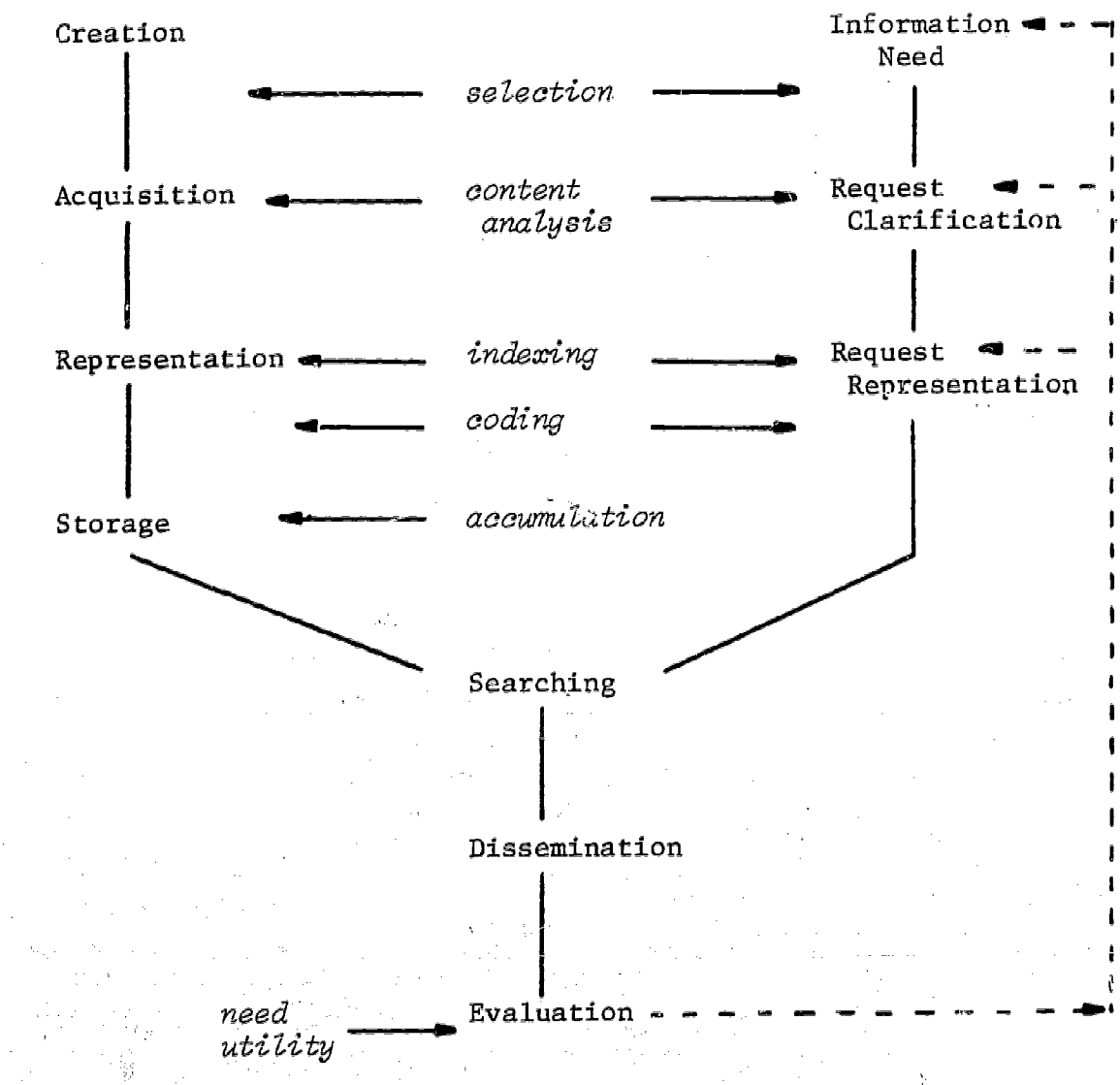


Figure 2.1: The Generalized IS&R Process.

of *shared agreement* about the worth of the product (*i.e.*, the retrieved data) and in terms of *adherence to the rules* of interaction. I shall discuss the nature of this common understanding in subsequent chapters of this dissertation.

As a final comment, one cannot help but be impressed by the diversity of research efforts in and the copiousness of the literature of Information Storage and Retrieval.* Almost every branch of mathematics has been employed in an attempt to satisfactorily model the processes depicted in Figure 2.1. Generally, such efforts have been inconclusive. While quite a bit is known about algorithms for data storage and file handling operations, little is known about the proper techniques for document selection and representation. Even less is known about the manner in which searchers go about and, eventually, satisfy their information needs. In a general way, one may say that workers in the field of IS&R are at present only promulgating a type of professional folklore. Individual studies are difficult to comprehend, let alone evaluate, in the absence of an underlying theory. Theory is a prerequisite to the successful recording of a "tellable history."

3. Directions

The major conclusion to be derived from this brief overview of the origins and nature of the information problem is that we should stop being overly concerned about the *quantity* of data produced by the sciences. Rather, our attention should be directed toward improving the *quality* of the representation

---

* We shall not here attempt to review this large body of literature. The reader is directed to Volumes 1-6 of the Annual Review of Information Science and Technology [14] as a suitable starting point for such a review.

of these data. It should be obvious that the sciences have not provided the needed "ordering framework" for the representation and dissemination of their myriad results. In fact, it could even be argued that many of the problems that we must work to overcome are caused, as Fugmann [13] puts it, by "...the lack of order that science has tolerated among its own results from the very beginning." To overcome this lack of order, the field of Information Storage and Retrieval must have as its goal the establishment of the requisite ordering between the sciences. Graziano [15] makes this a little more precise:

> The proper concern of the science of documentation [IS&R] then may be thought of as consisting of the operational *methods* of identifying elements, distinguishing elements from each other, and for transmitting sets of patterns from one time and/or place to another in such a way so as not to destory the power of the symbols to convey exact concepts.

Throughout this chapter it has been implied that the operation of document representation is crucial to the success of information retrieval. In fact, the central theme of this dissertation is an analysis of the nature and role of this representational activity. Indexing is identified as the prime exemplar of this activity. It is believed that a comprehensive theory of the indexing process will adequately serve to represent the nature of the common understanding called information retrieval.

In Chapter 2 I will further consider the topic of the role of indexing in IS&R processes. Attention will be directed to the form of a theory of indexing, especially with respect to the generalized role of theory. Finally, a statement will be made concerning the problem associated with present-day indexing practices. In Chapter 3, by way of historical review, the sparce

literature of indexing theory will be briefly explored. Following this, Chapter 4 presents the proposed indexing theory. The conceptualizations derived from this theory will then be used in Chapter 5 for a reconsideration of "information need," "inquisitiveness" and "relevance." Finally, in Chapter 6 I summarize the previous chapters and comment on the possibilities for applications and for future research.

References

1. K. Weick, "Meaning and Misunderstanding," *Contemporary Psychology* 14 (7), 1969, 357.

2. Y. Bar-Hillel, *Language and Information*, Addison-Wesley, New York, N.Y., 1964, 330.

3. A. Bonnard, *Greek Civilization: From the Iliad to the Parthenon*, The MacMillan Company, New York, N.Y., 1958.

4. J. C. R. Licklider, "A Crux in Scientific and Technical Communication," *American Psychologist* 21(11), 1966, 1044.

5. M. G. Mellon, *Chemical Publications: Their Nature and Use*, McGraw-Hill Book Company, New York, N.Y., 1965, 8.

6. A. I. Mikhailov, A. I. Chernyi and R. S. Gilyarevskii, "Informatics: its Scope and Methods," in *On Theoretical Problems of Informatics*, International Federation for Documentation publication 435, All-Union Institute for Scientific and Technical Information, Moscow, 1969, 9-12.

7. R. Fugmann, "Theoretical Aspects of Communication in Chemistry," *Angewandte Chemie International Edition* 9(8), 1970, 556.

8. *Ibid.*, 564.

9. M. G. Mellon, *op. cit.*, 1.

10. S. Gorn, "The Computer and Information Sciences and the Community of Disciplines," *Behavioral Science* 12(6), 1967, 433-52.

11. J. Marschak, "Economics of Information Systems," *Journal of the American Statistical Association* 66(333), 1971, 195.

12. H. Garfinkel, *Studies in Ethnomethodology*, Prentice-Hall, Englewood Cliffs, N.J., 1967, 24.

13. R. Fugmann, *op. cit.*, 574.

14. C. A. Cuadra (Ed.), *Annual Review of Information Science and Technology*, Encyclopaedia Britannia, Inc. (Vol. 3-6); John Wiley, Inc. (Vol. 1-2), 1966-71.

15. E. E. Graziano, "On a Theory of Documentation," *American Documentation* 19(1), 1968, 86.

CHAPTER II. INDEXING: ART, THEORY AND MODEL

> Without a theory, however provisional or loosely formulated, there is only a miscellany of observations, having no significance either in themselves or over against the plenum of fact from which they have been arbitrarily or accidentially selected.
>
> A. Kaplan *The Conduct of Inquiry*

This chapter is designed to provide the supporting framework for a *statement of the problem* central to the research the results of which are reported in this dissertation. Accordingly, further attention will be directed toward a consideration of the role of indexing in information storage and retrieval processes. Data relevant to this topic will be obtained through an analysis of alternate definitions of indexing and through a consideration of the intrinsic importance of the indexing operation. Some unanswered research questions will then be contrasted with present-day indexing practices and guidelines. Finally, since this dissertation presents a theory of indexing, special attention will be directed to an analysis of the functions of *theory*, *model* and *definition* in the organization and understanding of a miscellany of observations.

1. Research Trends

In Section 2 of the previous chapter I commented briefly on the proliferation of research in the field of information storage and retrieval. I find it difficult, if not impossible, to completely categorize all IS&R-related studies. At best, only a general grouping can be effected. Taulbee [1], in 1967, identified six broad areas of investigation (or, shall we say, activities) in IS&R. It is believed that this classification remains valid

with respect to present-day efforts.

- fundamental investigations (*e.g.*, sentence parsing and associative storage)
- reports of experiences with operating systems
- guidelines for system design and modification
- document relevance assessment
- the "how-to" for implementation
- bibliographies

At present, the activities that fall under the "fundamental" label include the development of storage and retrieval algorithms, natural language semantic and syntactic analysis, and the development of question-answering systems [2]. Conspicuous by its absence from the above list is the development of a cohesive theory of information storage and retrieval. Although researchers often refer to *information retrieval theory* it appears that this theory is an unstated (perhaps unformed) amalgamation of theories of specific retrieval functions--*e.g.*, logic, searching and storage techniques. The same accusations can be leveled at the discipline of indexing. Markus [3], in the early sixties, outlined areas of much needed research in indexing. Some of the following were included: index format; index use patterns; the teaching of indexing; increased indexing speed; equipment modification and computer program development. It is again interesting to note that *indexing theory* was not (and is still not) one of the areas mentioned.

It is correct to assume that Computer Science can be of utility in the solution of the many information retrieval problems. However, it is not correct to assume that such solutions can be effected essentially overnight.

Unfortunately, most of the computer-based storage and retrieval systems that are in existence today are the result of the "urge" to apply the remedial powers of the computer to the problem without full appreciation of the problem being "remedied." Consequently, the following research and development cycle is firmly established [4]: the need for computer-based retrieval systems is felt; computer-based systems are created; the lack of suitable evaluation criteria is felt; research is conducted on evaluation techniques; new systems are built; and so on... I conclude that this cycle must be broken if measurably effective progress is to be made. A theoretical basis for IS&R must be developed that, for a given application (specific information retrieval problem), will yield appropriate systems evaluation criteria. Such a development is prerequisite to systems implementation. It is believed that the elements of such a theory will emerge from a consideration of indexing as viewed from the interdisciplinary philosophical framework of Information Science.

2. Indexing as Art

The absence of a unifying theory for information storage and retrieval (as for most of its component processes) is emphasized by the many divergent definitions of indexing. Indeed, there appear to be as many definitions of indexing as there are individual indexing applications and studies. A limited sampling of these definitions includes the following conceptualizations: a representation of content; a systematic guide to the content; an identification tag; a product serving to point out, direct and guide; a search access point; a dictionary of nomenclature; an association between concepts and terms; a means of making information available. At this point

the reader may ask: 'But I should like to know...'. My only rejoinder, when faced with so many definitions, is a terse conclusion; the confusion between the definitions of the concepts of information, index, indexing, retrieval and system (to name a few) *must* be resolved in the development of a useful theory.

In addition to the number of definitions of indexing, the situation is further complicated by the variety of types of documents indexed and of the number of resulting indexes.* One wonders if there is not some formal connection, or relationship, between these different types of indexes. This question remains unanswered in the literature of IS&R. In present-day indexing practices the analysis of document content and the resulting indexing decisions are mainly treated as an artful practice. Mellon [5] emphasizes this point: "The making of indexes is an art in itself, involving more than a comprehensive knowledge of the general subject being covered, and the use of indexes is no less an art." It is not surprising therefore that there exists no comprehensive treatment of the process of indexing--one only finds suggestions or examples of how indexing ought (in the opinion of the writer) to be done. Even publications purporting to provide "indexing standards" are really just promulgating "suggestions." Consider the statement of purpose of the *USA Standard Basic Criteria for Indexers* [6]:

> It [USA Standard] does not attempt to set standards for every detail or for all the diverse techniques of indexing; these should be determined for each index on the basis of the type of material indexed and the type of user for whom it is designed, among other factors.

---

* Author, subject, title, citation, patent number, formula, *etc.*

Furthermore, when indexing rules are provided [7-11] the emphasis is on the "cook-book" approach to indexing. Favorite topics include the standardization of headings, the treatment of synomyms, cross-references, how to index names, how to check index entries, *etc.* [12]. Such an "artistic" approach only reflects the lack of an underlying theory of indexing. A theory of indexing must be provided that emphasizes the importance and centrality of the index operation in IS&R processes.

One of the primary goals of research in IS&R and in indexing is outlined by Baxendale [13]:

> ...starting with a collection of 50,000 documents which covers four subject fields, which is to grow at the rate of 2000 documents per year, which is to be purged on the basis of activity, and which will be subject to approximately 75% specific data requests and 25% general queries, what type of indexing device will best accomodate these conditions?

Before the development of such an indexing/retrieval-system nomogram can be realized, attention will have to be directed, for lack of better terminology, toward the identification and explication of first principles. It is quite a conceptual distance between "cook-book" indexing and indexing-system nomograms! The following are samples of things we will need to know more about:

What is the nature of the selective transmission of data?

Who decides what data is to be passed on to the user?

What is the nature of the decision making operation performed by indexers and analysts?

What is the function of index languages and devices?

How are index terms to be selected?

How is the quality of the index to be evaluated?

While several additional questions could be added to this list, those given are sufficiently indicative of the type of problems that will have to be considered in the development of indexing theory.

3. Theory and Model

In the previous discussion I have presented evidence in support of the conclusion that an underlying theory is needed for the processes of information storage and retrieval. I also conclude that a theory is needed for the crucial representational process called indexing. Considering the emphasis that is being placed upon the concept of theory, it is appropriate, at this point, to briefly discuss the role of *theory* in the sciences. This discussion is intended as a brief prologue to a consideration of the role of indexing theory with respect to information storage and retrieval. The interested reader is referred to Kaplan [14] for a detailed discussion of the nature of theory and to Caws [15] for insightful investigations into the nature of definition.

An area of investigation, or subject area, is composed of an assortment of observables* variously labeled as data, knowledge, experience and fact (we will defer a discussion of the validity of these labels to Chapters 4 and 5). The study of observables without a basic organizational framework is judged to be of low utility. Consequently, theory is an attempt to provide an organization for a set of observables or expected observables. Emphasis is placed upon the unification, systematization and representation of observables (see Figure 3.1). Properly formulated, theory is a state-

---

* Things that come into our perview through our sensory mechanisms.

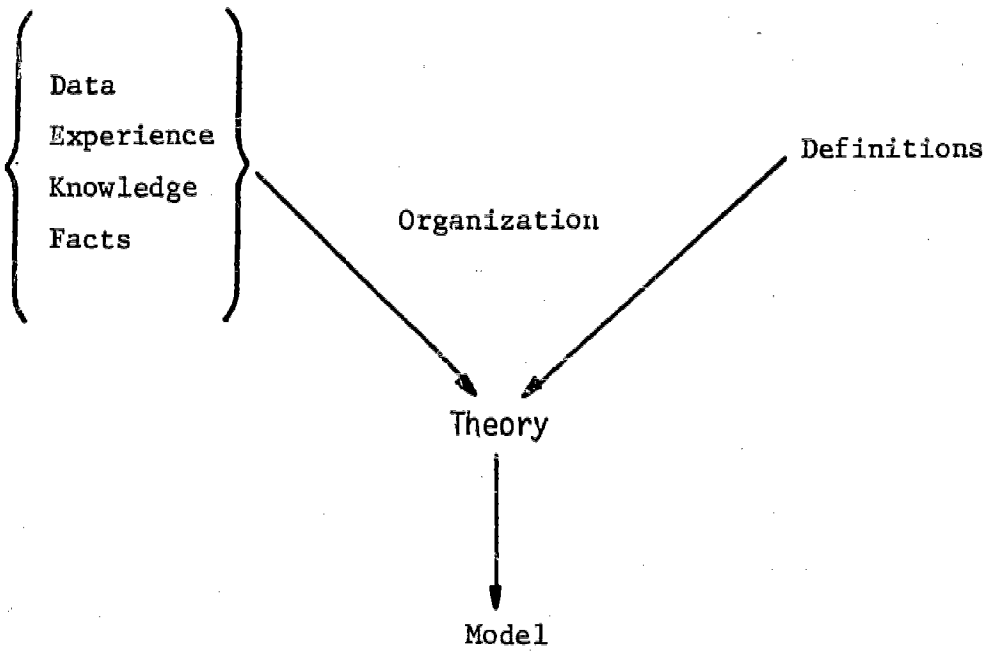


Figure 3.1: Theory and Model in Science

ment about the inherent structure of the observables. Thus, theory simultaneously describes and analyzes the collection of observables. However, as Kaplan points out [16], theory must provide more than a simple description of observables:

> A theory is more than a synopsis of rules... it sets forth some idea of the rules of the game by which moves become intelligible.

Theory is also expected to be evaluated in terms of its predictive ability. Consequently, theory, as a linkage between observables and hypotheses, is a guide to the collection and subsequent interpretation of data. The structure of the theory is provided by a cohesive set of definitions about the observables and about the relations that exist between observables. Caws makes this clear [17]:

> Ostensive definition [*i.e.*, definition by example] is clearly not enough. Moreover, a set of isolated statements about isolated phenomena is not yet science; only when the terms in the statements are related to one another does scientific theory emerge.

Generally, the order that is imposed on the observables by the theory is a consequence of the order that exists between the component definitions. The central role of definition in theory cannot be over emphasized.

Some theories act as models. Frequently some theory about observables may either be too difficult to construct, too difficult to understand or else unsuitable for the symbolic manipulation of observables. Fortunately, one theory may act as a model for another theory. The minimal condition for a theory/model relationship between two theories is resemblance in form. Thus, there is said to be a structural analogy or even isomorphism between the theories. If theory A is easier to understand and to manipulate than theory B, and if the theories are isomorphic, then the development of theory

A will serve as a model for theory B. Hopefully, an increased understanding of B will result from this modeling activity. Finally, apart from being a conceptual analogy, a good model will be the source of relevant hypotheses to be tested on the set of observables (that is, the theory will give rise to experimentation).

3.1 Indexing as a Model of IS&R

Figure 3.2 presents a schematic which illustrates the role of the indexing theory that is being proposed in this dissertation. I believe that, at our present state of knowledge, a comprehensive and workable theory of information storage and retrieval is unobtainable. Thus, we have labeled the theory an "unknown theory." Nevertheless, the indexing theory that is presented in Chapter 4 is, I believe, a suitable working model for many of the processes of Information Storage and Retrieval in addition to its standing independently as a theory of the processes of indexing and index creation. While it is not likely that indexing theory is *the* theory of IS&R, it is believed that it provides a novel and useful interpretation of the associated observables.

4. Statement of the Problem

The previous chapters have emphasized the necessity for research in information storage and retrieval. However, it is concluded that the most significant problem associated with current efforts is the lack of a useful theory of information storage and retrieval. Because of the absence of such a theory, it is difficult to properly evaluate present-day research and systems-design efforts. But it would be foolish not to acknowledge the difficulty of formulating an adequate, all-inclusive theory of IS&R. So I

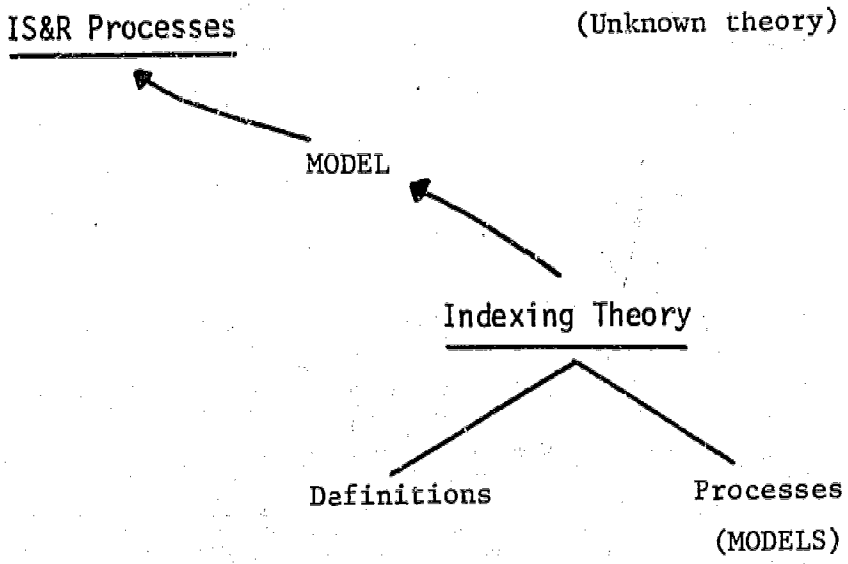


Figure 3.1.1: Indexing Theory as a Model for IS&R.

have devoted my attention primarily to the operation of indexing, working under the assumption that indexing is the central and crucial operation for the successful retrieval of information. Thus, it is believed that a theory of indexing can serve to model the essence of the information storage and retrieval processes.

As I have said, there is, at present, no comprehensive, unifying theory of indexing available for these applications. Repeatedly, indexing viewed as an "art" has failed to provide the necessary theory. Consequently, the problem is to develop a theory of indexing that satisfies two criteria: first, it must provide the basis for the systematic analysis of both indexing procedures and resulting indexes, and, second, it must provide a conceptual basis for the evaluation of IS&R systems. It is toward these goals that the research which lead to this dissertation has been directed.

References

1. O. E. Taulbee, "Content Analysis, Specification, and Control," in *Annual Review of Information Science and Technology* Vol. 3, C. Cuadra, ed., Encyclopaedia Britannica, Chicago, Illinois, 1968, 106.

2. J. Minker and S. Rosenfeld, "Introduction and Perspectives for the 1971 ACM Information Storage and Retrieval Symposium," *Proceedings of the Symposium on Information Storage and Retrieval*, The University of Maryland, 1971, 1.

3. J. Markus, "State of the Art of Published Indexes," *American Documentation* 13(1), 1962, 16.

4. J. E. Rush, "Theory and Practice in Information Retrieval," in *The Social Impact of Information Retrieval*, Seventh Annual National Colloquium on Information Retrieval, A.D. Berton, ed., Medical Documentation Service, The College of Physicians of Philadelphia, Philadelphia, Pa., 1970, 60-63.

5. M. G. Mellon, *Chemical Publications: Their Nature and Use*, McGraw-Hill Book Company, New York, N.Y., 1956, 203.

6. *USA Standard Basic Criteria for Indexers*, United States of America Standards Institute, Z39.4-1968, 1968, 7.

7. E. T. Harris, *A Guide for the Preparation of Indexes*, The RAND Corporation, 1965, AD-615-605.

8. R. L. Collison, *Indexes and Indexing*, John de Graff, Inc., New York, N.Y., 1959.

9. M. T. Wheeler, *Indexing: Principles, Rules and Examples*, The New York State Library, Albany, N.Y., 1957.

10. M. Taube, *Studies in Coordinate Indexing*, Documentation Incorporated, Bethesda, Md., 1956.

11. G. N. Knight, ed., *Training in Indexing - A Course of the Society of Indexers*, The M.I.T. Press, Cambridge, Mass., 1969.

12. C. L. Bernier, "Indexing Process Evaluation," *American Documentation* 16(4), 1965, 326.

13. P. Baxendale, "Content Analysis, Specification, and Control," in *Annual Review of Information Science and Technology* vol. 1, C. Cuadra, ed., John Wiley & Sons, Inc., New York, N.Y., 1965, 96.

14. A. Kaplan, *The Conduct of Inquiry*, Chandler Publishing Co., 1964, 302-326.

15. P. Caws, "The Functions of Definition in Science," *Philosophy of Science* 26(3), 1959, 201-228.

16. A. Kaplan, *op. cit.*, 302.

17. P. Caws, *The Philosophy of Science*, D. Van Nostrand Co., Inc., Princeton, N.J., 1966, 44.

CHAPTER III. PREVIOUS INDEXING THEORIES

> Very little advance in culture could be made, even by the greatest man of genius, if he were dependent for what knowledge he might acquire upon his own personal observations. Indeed, it might be said that exceptional mental ability involves a power to absorb the ideas of others, and even that the most original people are those who are able to borrow most freely.
>
> Libby

Almost any study that is undertaken has a corpus of related and relevant literature that must be considered, and, the research reported here being no exception, this chapter contains a summary and critical evaluation of two previous attempts toward the formulation of a theory of indexing. It should be noted that these early efforts of theory development were not continued beyond their initial exposition in the late 1950's and the early 1960's. Nevertheless, as we shall see, some valid comments were made about the indexing process.

The material in this chapter is presented in three short parts: 1) an examination of Jonker's [1-3] indexing-continuum theory, 2) an examination of Heilprin's [4] model of indexing and, 3) a discussion of the questions which these two studies left unanswered.

1. The Indexing Continuum

By way of introduction, Jonker identified three central problem areas in IS&R: 1) the indexing problem (the problem of document representation), 2) the coding problem (the problem of the conversion from a document description to a machine recognizable code) and, 3) the machine systems problem (the problem of the selection of the most advantageous code-processing system). He concluded that these three factors ultimately reduce to the constraint of

cost. Consequently, Jonker believed that the goal of IS&R research (with respect to indexing) is to construct the most economical indexing system, all the while providing the maximum depth of indexing.*

From this reviewer's viewpoint, the effectiveness and the utility of an indexing system are far more important than the cost. Cost cannot simply be equated with dollar outlay. One might, rather, consider a comparison between the cost of processing a document and the cost (*i.e.*, loss of utility) of the failure of the system to provide that document to a user. I realize that a consideration of cost, in terms of dollars, is important to the system designer, but cost should be relegated to a position of lesser importance in the development of a theory of indexing.

Jonker does, however, focus on one of the central problems directly associated with indexing:

> The inescapable conclusion seems to be that no true understanding of existing indexing systems and problems seems possible, unless all systems can be seen in the light of more general common precepts, linking all those systems together into a closed single system. [5]

Thus, Jonker's theory of indexing is best described as an attempted taxonomy for the classification of the various indexing systems. This taxonomy is based on the belief that IS&R systems do not deal with items of information

---

* Jonker uses the term "depth of indexing" as a synonym for the number of index entries per document. This term usually refers to the hierarchical specificity of the index entry, *i.e.*, depth of detail.

(sic)* themselves, but, rather, with what he calls *index-information*. An item of index-information is defined by a specification of both the "number of terms" it contains and the connections between such terms.** Jonker proposes that the characterization of index-information is provided by the *indexing continuum* which is a composite of the *terminological continuum* and the *connective continuum*.

1.1 The Terminological Continuum

An indexing system (by inference, a system that deals with index-information) provides a meta-language for document description. Such an *intensified language* provides a one-word name for every important concept (Jonker did not discuss the nature of "important concept"). Consequently, the terms of the language share a close relationship with both the symbol (document word) and the meaning represented (Jonker did not define "meaning"). Figure 1.1.1 is a representation of the continuum of intensified language terms--the terminological continuum. Term size and the divisibility or the *permutivity* of the term are assumed to increase as one moves to the right in the continuum. Permutivity refers to the variability of representation. A particular retrieval system is designed, as he saw it, by attempting to place the needs of the average user on the terminological continuum.

1.2 The Connective Continuum

The data contained in a single document may be utilized in a variety of ways. Jonker defines *diffuseness**** as the number of potential indexing

---

* The word "information" is occasionally so marked in this discussion to remind the reader that the word "data" is meant, as I see it.

** To generalize, I interpret Jonker to mean that an item of index-information is an n-term whose elements share a common connective relationship.

***This can be interpreted as indexing "breadth."

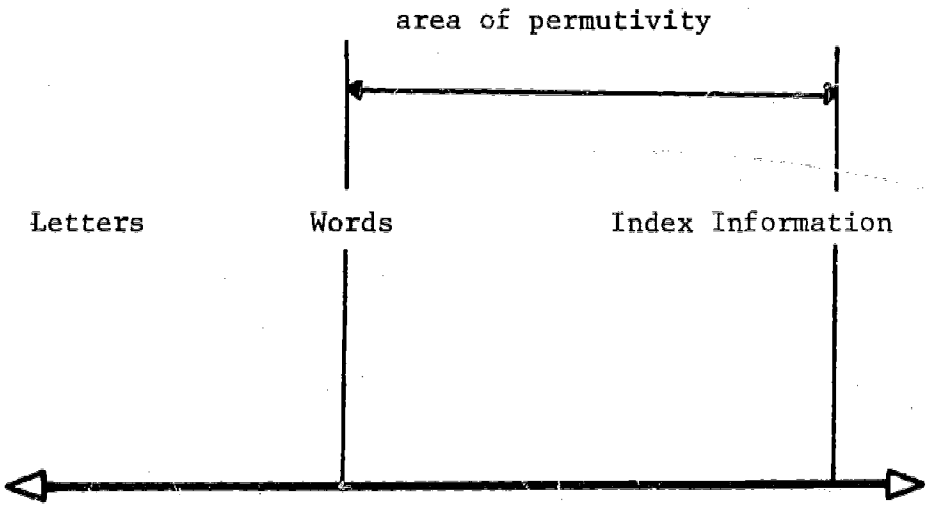


Figure 1.1.1: The Terminological Continuum (from [1]).

points for a given document. He claimed that it is the diffuseness of the information (sic) which characterizes (and defines) the information problems that exist today. Consequently, diffuseness of information is treated as the controling parameter in his generalized theory of indexing.

Jonker believed that the number of index entries per document is directly related to the number of *indexing viewpoints* that can be identified. However, he gives no word on how one might go about choosing a set of exhaustive viewpoints. It seems almost obvious that a document should be represented in a form amenable to processing by any scheme of organization and/or retrieval. Jonker nearly reached the same conclusion:

> If properly indexed, an item of information is indexed by any keyword that is or could possibly become of importance to any potential user of the item of information. [6]

Apparently, imporved indexing is achieved by selecting index terms from the right of the terminological continuum. The resulting entries are characterized by an increased number of words per term which is believed to be directly proportional to the degree of *hierarchical connectedness* of the metalanguage. Figure 1.2.1 shows the resulting connective continuum ranging between short terms and long (multiple word) terms. Jonker believed that short terms are representative of coordinate systems and long terms are representative of hierarchically-based indexing systems. Thus, the connective continuum characterizes indexing systems by their degree of generic character (see Perry and Kent in [7]). Jonker associated hierarchical classification with a low degree of diffuseness. Consequently, retrieval based on the short end of the continuum is fluid and arbitrary, whereas the long end is characterized by rigid retrieval. It is inferred that this is the essential

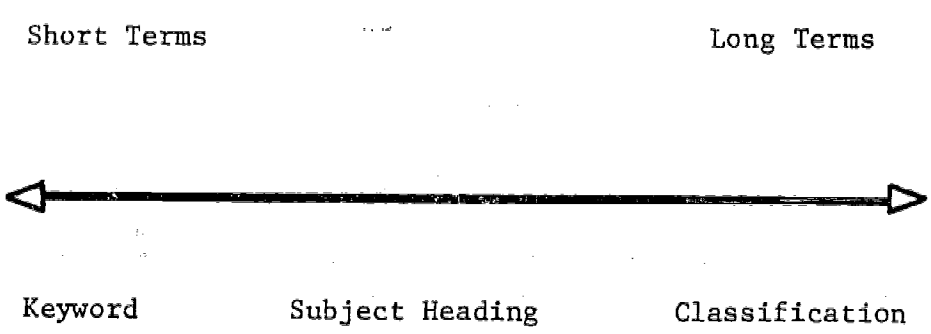


Figure 1.2.1: The Connective Continuum (from [1]).

difference between complete independence and dependence of terms.

To summarize, the short end of the continuum is characterized by high diffuseness, high permutability and low hierarchical connectedness, H. The long end is characterized by low diffuseness and permutability, and high H. Jonker concludes that the diffuseness and permutability taken together determine the retrieval power, R, of the system. This is a measure of both the accuracy with which information (sic) can be indexed and the detail by which it can be retrieved. Jonker states that R • H = constant*, or, an increase in H can only be obtained at an expense of R, and conversely.

2. An "Intuitive" Mathematical Model of Indexing

Heilprin [4] attempted to provide a mathematical treatment of the general theory of indexing developed by Jonker. His first step was to provide a formalization of the concepts of diffuseness, permutivity and hierarchical connectedness. For convenience of analysis, Heilprin chose to replace (the average term length) by n (the mean number of independent terms per stored item at point ℓ in the descriptive continuum). Figure 2.1 shows the assumed inverse relationship between the two variables.

Heilprin introduced the concept of a *search path* corresponding to the number of "paths" from questions to documents. He contended, as I believe rightfully, that the search method is independent of the number of available paths--rather, search paths (or permutations) depend only on the index. Furthermore, since most indexing systems do not permit full permutability,

---

* Jonker overlooked the fact that the precision of the retrieval usually increases with increased hierarchical definition. In other words, the success of retrieval is not directly related to the number of permutations available from a term.

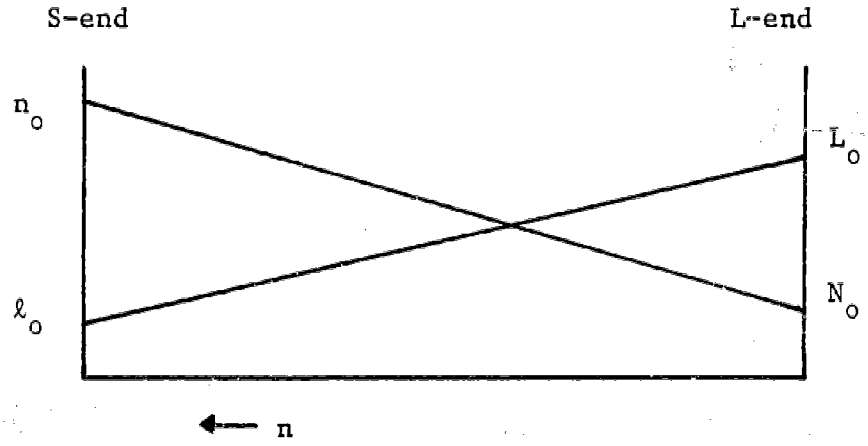


Figure 2.1: The inverse relationship between ℓ and n in the Descriptive Continuum (from [2]).

Heilprin introduced a *noise of permutability*, N, which is an expression of the deviation from the maximum. Maximum permutability is reasoned to be the set of all permutations of n terms taken q (query terms) at a time--*i.e.*, (n)q. Similarly, Heilprin introduced a hierarchical noise, M, which represents the discrepancy between the ideal number of hierarchical levels and the number of levels by which a document is indexed. The following equations were derived:

$$D = n$$

$$P = (1-n)(n)q$$

$$H = n_0 (1-M)/n$$

Accordingly, Heilprin represented all possible indexing systems by a 3-space formed by the n, H and P values (see Figure 2.2). This is a restatement of the assumed fundamental equation:

$$R \bullet H = D \bullet P \bullet H = n \bullet P \bullet H$$

Clearly, for a single positioning on the descriptive continuum, there can be various values of P and H depending on the independent, but confounding, action of the N and M variables. This creates a family of curves all generated by a single value of n. Heilprin contends that a complete family of such curves would fill much of the index region. There will be as many index curves and regions as there are values of n_0 (recall that n_0 is the value of n at the short end of the continuum). Although this concept has not been further developed, it seems to me that the size and relative positioning of the index region could serve as an analytical representation of a given indexing system.

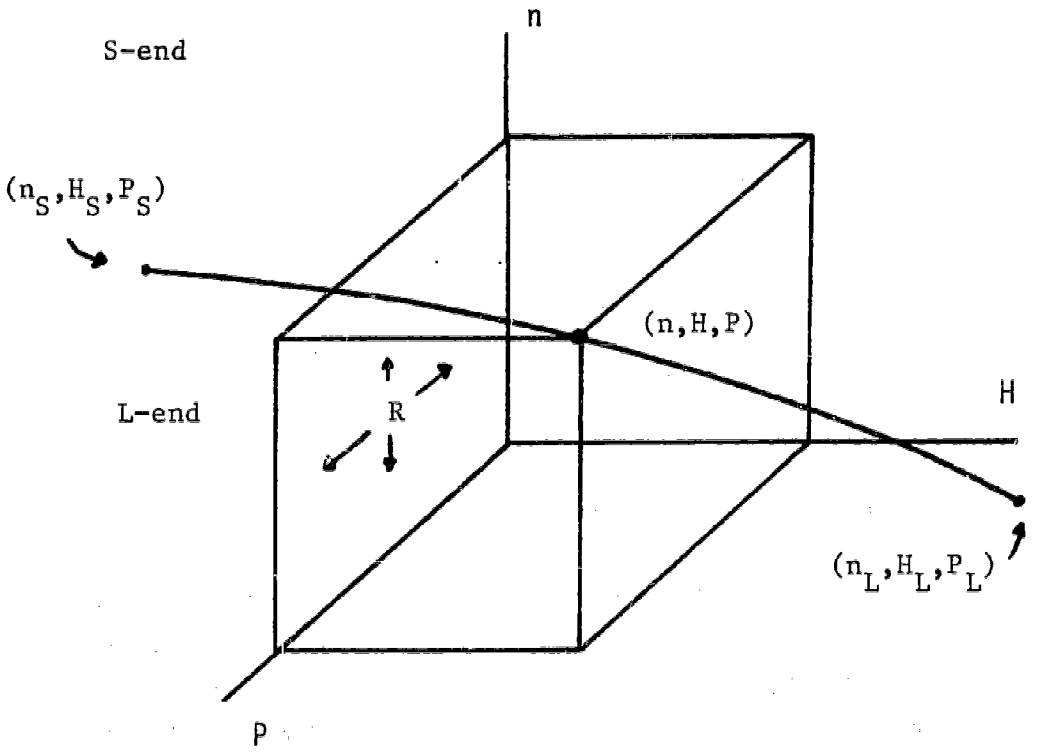


Figure 2.2: The Index Curve and the Index Region (from [4]).

3. A Brief Discussion of the Two Indexing Models

Although the presentation of the Jonker and Heilprin models has been very brief, essential concepts, definitions and relationships which they proposed have been presented. I hope the reader can, at least, gain an appreciation of the general "tone" of their theory.

An important point relative to this brief review is that *a* theory has not been presented. The mere enumeration of some of the components of the indexing process* (and associated variables) does not answer such questions as why index, what should an index provide, or what is the role of the index in the process of information storage and retrieval or in human behavior in general? However, considerable impetus for the creation of such a general theory is provided by their presentations.

Jonker must be faulted for his over concern for economy and cost factors--this is not a constraint for a theory, but rather, just another variable to be considered. In addition, it is believed that he has a fundamental misconception of the concept of information. Does an index store information or does it store data? At least some discussion on this point should preceed any general use of, or reliance on, the word "information." Just as the definition of information is loose and non-precise, the *descriptive continuum*

---

* There exists an extensive and growing literature concerning the analysis of indexing parameters. The usual assortment includes: the type of classification scheme; the depth and breadth of indexing; the number of terms per entry; the number of entries per document; the number of documents per term; the indexing language used; the type of indexing aid used (*e.g.*, links, roles, weights). It is emphasized that the behavior of these parameters is frequently analyzed for systems patterned after existing ones. Frequently, the analysis of parameter behavior is modeled by simulation studies (see [8] for a review of these studies).

must be criticized for lack of a quantitative measure. It is questionable whether term length is a meaningful variable for a *meaningful* characterization of an indexing system. Heilprin's use of the number of terms/item is more realistic, but remains essentially unsupported. Indeed, it is not clear that Heilprin's functional analysis (as presented) really adds anything new to Jonker's theory. A formal presentation of informal concepts must retain an element of uncertainity and informality.

In a more positive vein, both Jonker and Heilprin present concepts that merit further consideration and development. Each concept surely will find its place in *a* theory of indexing. I conclude this chapter with a listing of these concepts:

- Indexing systems represented by a single closed system.
- An index entry represented as a term/relationship structure.
- *Diffuseness* of "information."
- Indexing structure dependence and independence.
- Noise must be accounted for in any valid theory.
- The term-query *search path*.
- The *index region* as representative of indexing systems.

References

1. F. Jonker, *The Descriptive Continuum, a 'Generalized' Theory of Indexing*, Jonker Business Machines, Inc., 1957, AD-132-358.

2. F. Jonker, *Outline of a General Theory of Index Terminology and Indexing Methods*, Jonker Business Machines, Inc., 1961, AD-272-820.

3. F. Jonker, *Indexing Theory, Indexing Methods and Search Devices*, The Scarecrow Press Inc., New York, N.Y., 1964.

4. L. B. Heilprin, *Mathematical Model of Indexing*, Documentation Inc., 1957, AD-136-477.

5. F. Jonker, *op. cit.*, 1961, 2.

6. *ibid.*, 36.

7. B. C. Vickery, *On Retrieval Systems Theory*, Archon Books, London, 1968.

8. B. C. Landry, "An Indexing and Re-indexing Simulation Model," *Computer and Information Science Research Center Technical Report* No. 69-14, The Ohio State University, Columbus, Ohio, 1969.

CHAPTER IV. A THEORY OF INDEXING

> The Master said, Yu, shall I tell you what knowledge is? When you know a thing, to know that you know it, and when you do not know a thing, to recognize that you do not know it. That is knowledge.
>
> *Analects of Confucius* (Waley's Translation)

> An index is an array of symbols, systematically arranged, together with a reference from each symbol to the physical location of the item symbolized.
>
> Mortimer Taube: *Studies in Coordinate Indexing*

1. Introduction

In Chapters I and II, I have emphasized that research in information storage and retrieval has as its goal the discovery of solutions to the problem of efficiently organizing man's expanding knowledge. Although a variety of approaches have been applied in attempts to solve the problem, the discipline suffers from the absence of any underlying model, models which are fundamental to any well defined science.* It appears that much of the effort to develop such models (see Chapter III), although well-intentioned, is misdirected because there is little appreciation of the theoretical foundations of information storage and retrieval. As a start toward resolving some of these difficulties, the elements of a basis for a theory of information storage and retrieval are set forth in this chapter. It is hypothesized that the theory can best be formulated and expressed in terms of a general theory of indexing.

---

* Recall the definitions of *model* and *theory* given in Chapter II.

In the first section of this chapter is stated the basic premise of the theory, and a number of fundamental definitions are given. Following this there is a discussion of the similarities between the indexing process and the general communication process. Attention is then directed to the view that indexing is an order increasing operation, and some thermodynamic notions are invoked to aid in this description. The concept of a "theoretical index" is then elaborated and compared with real-world indexing systems. Finally, the contribution of the human performance variable to the efficacy of an indexing system is considered.

Just a note on organization. This chapter is divided into two parallel parts, each of which contains nine sections. The first part provides a concise exposition of the *theory* of indexing. The second part gives supporting data and discussion related to the materials presented in the first part of the chapter.

2. Some First Definitions and Postulates

It is assumed that any theory about processes in the real world must involve the operation of measurement and the specification of units. Accordingly, the concept of *data element* is postulated to be the fundamental unit of documentation. The following four definitions are treated as antecedent to the definition of data element.

Def. 2.1 Measurement: Measurement is the process of selecting among a set of possible alternatives exactly those which characterize the attribute under observation.

Def. 2.2 Attribute: An attribute is any discriminable feature of an event that is susceptible to some discriminable variation from event to event (Bruner [1]).

or, An attribute is a subset of the set of all possible observables associated with an event.

Def. 2.3 Unit of Measure:

A unit of measure is a *metric* which is defined by the function $A \times A \rightarrow N$ (natural numbers), which assigns to each pair $a,b \in A$ a non-negative real number $\rho(a,b)$ and such that the following properties hold:

1) $\rho(a,b) = \rho(b,a) \; \forall \; a,b$
2) $\rho(a,b) = 0$ iff $a = b$
3) $\rho(a,b) + \rho(b,c) \geq \rho(a,c)$

Def. 2.4 Precision: Precision is the number of alternative values for the result of the operation of mesaurement.

or, Given $a,b \in A$, a metric ρ is more precise than a metric ρ' if $\rho(a,b) < \rho'(a,b)$.

Thus, we now have:

Def. 2.5 Data Element: A data element, d, is the smallest thing which can be recognized as a discrete element of that class of things named by a specific attribute, for a given unit of measure with a given precision of measurement.

The following definitions build on the concept of data element:

Def. 2.6 Relation: Given sets of data elements $d_1, d_2, d_3, \ldots, d_n$ (where $d_k = \{d_{k_1}, d_{k_2}, \ldots\}$, $1 \leq k \leq n$), form the cross product $d_1 \times d_2 \times d_3 \times \ldots \times d_n = \prod_{k=1}^{n} d_k$. A relation, R, is a subset of this conjunctive set: $R \subset \prod_{k=1}^{n} d_k$.

Def. 2.7 Ordered set: A set of data elements is said to be ordered by a relation R (over the data elements) if the relation is transitive and satisfies the trichotomy law ($d_i R d_j$ or $d_j R d_i$ or $d_i = d_j$ where $d_i, d_j \in \prod_{k=1}^{n} d_k$).

Def. 2.8 Well-Ordered Set: An ordered set of data elements is said to be well ordered if its every non-void subset has a first element.

Def. 2.9 Document: A document, D, is a well-ordered set of data elements.

Def. 2.10 Document Space: A document space is an ordered set of documents. This set is denoted by: $\mathcal{D} = \{D_1, D_2, \ldots\}$.

Def. 2.11 Index Space: An index space, $\mathcal{I}$, is a representation of the data elements, d, and relations, R, found in the *indexing system* (defined in Section 4).

Theorem 2.1: $\mathcal{I}$ is a document and $\mathcal{I} \in \mathcal{D}$.

Def. 2.12 Index: An index, I, is the image* of composite order-preserving mappings performed on the document space $\mathcal{D}$.

Theorem 2.2: I is a document.

Def. 2.13 Query: A query, Q, is a well-ordered set of data elements such that $Q \subset I$ (*cf.* Def. 9.1).

Theorem 2.3: Q is a document

Postulate 2.1: $I = f(\mathcal{D}, \mathcal{I})$, where f is the *indexing* process (*cf.* Def. 4.1).

Postulate 2.2: Accurate retrieval depends upon the exactness of the indexing.

3. Communication and Indexing

Information storage and retrieval is inherently a part of communication. In fact, it can be argued that information storage and retrieval is central to all of our activities. It is thus necessary to formalize the nature of the ties between information storage and retrieval and communication.

---

* Given a function $f: S \rightarrow T \ni \forall s \in S \exists f(s) \in T$, we say that f(s) is the image of the mapping defined by f.

Def. 3.1 Communication: Communication is a closed system consisting of an effector, a receptor, a transmission channel and a feedback unit (*cf.* Fig. 3.1).

Def. 3.2 Flow Rate: The communication or flow rate is measured in data elements per unit time.

Postulate 3.1: The items transfered from element to element in communication are data elements and associated relations.

Postulate 3.2: Any theory or practice of communication which causes a loss of data elements, either through their misrepresentation or by restricting their flow, must be considered inadequate.

Accordingly, it is assumed that accurate and effective communication is the goal of an IS&R system. The following definitions consider the nature of *effective* communication.

Def. 3.3 Experience Set: The source's or receiver's memory is modeled as an ordered set of data elements and relations. Denote the experience set by (ES).

Theorem 3.1: An experience set is a document.

Def. 3.4 Interface Experience Set: The interface experience set (IES) represents the data elements and relations that are used in the actual communication between the source and the receiver.

Def. 3.5 Effective Communication: Effective communication is obtained when the intersection of the source experience space, $(ES)_s$, and the receiver experience space, $(ES)_r$, is non-empty. i.e., $(ES)_s \cap (ES)_r \neq \phi$.

Theorem 3.2: Effective communication is maximal when $(ES)_s = (ES)_r$.

Def. 3.6 Experience Set Transformations: Experience set transformations are defined by sets S and R whose domains are $(ES)_s$ and $(ES)_r$, respectively. These transformations

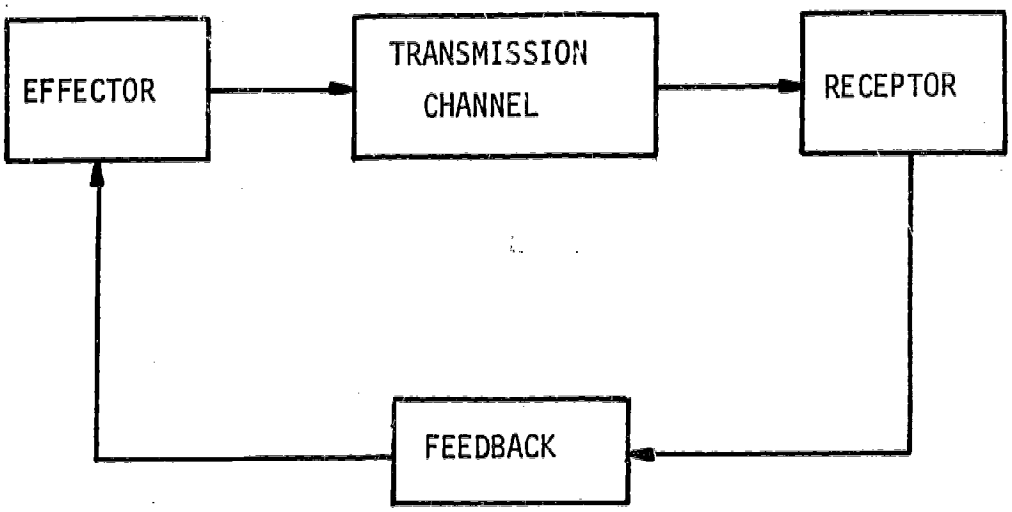


Figure 3.1: A General Model of Communication.

have the following property:
$S \cdot (ES)_s = (IES) = R \cdot (ES)_r$
(see Figure 3.2).

Theorem 3.3: An IS&R-system user must have knowledge of the organization and representation of the data elements in the system to achieve effective communication with it.

Postulate 3.3: The *indexing system* (*cf.* Def. 4.3) provides the interface experience set and the transformations required for effective communication.

One of the transformation functions in the indexing process deals with the order of the data elements that occur in the communication link. This order-defining transformation is based on the definition of five exhaustive, overlapping* classes of data-element relations:

Def. 3.7 Data-Element Relations:
A data element relation is an element of the set of relations, REL = {E,G,P,F,T} defined over sets of data elements $d = \{d_1, d_2, \ldots\}$ and sets of attributes A = {a,b,c,...}.

The relations comprising the set REL are defined as follows:

Def. 3.8 Equivalence Relation:
An equivalence relation, E, satisfies the following properties:
$d_i E d_j$ (reflexivity)
$d_i E d_j = d_j E d_i$ (Symmetry)
$d_i E d_j \ \& \ d_j E d_k \Rightarrow d_i E d_k$ (transitivity)

---

* That is, a pair of data elements may be related by combinations of these relations. For instance, we write aEFb , to mean both relations E and F operate on data elements a and b.

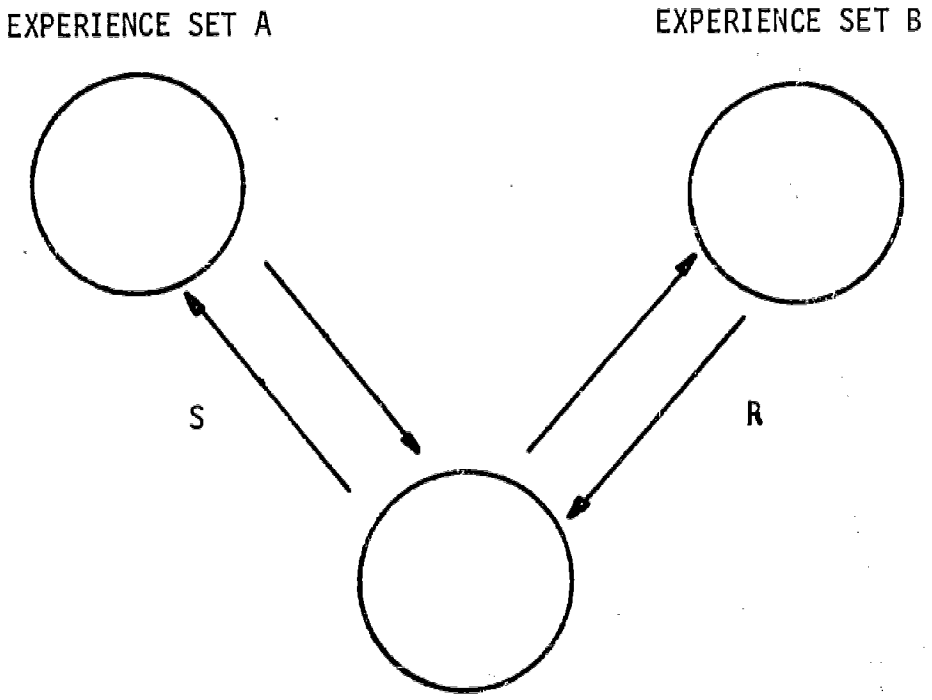


Figure 3.2: The Experience-set Interface. S and R are Experience-set Transformations. The Common Experience Set Represents A ∩ B.

Def. 3.9 Generic-Specific Relation:
The generic-specific relation, G, is defined by d_i "is generic to" d_j or, equivalently, by $d_i \geq d_j$. G is reflexive, transitive but not symmetric.

Def. 3.10 Part-Whole Relation:
A part-whole relation, P, is defined by: d_i "is a part of" item X, or, equivalently, by $d_i \in X$. P is only reflexive.

Def. 3.11 Difference Relation:
A difference relation, F, is defined by: d_i "is not equal to" d_j or, equivalently, by $d_i \neq d_j$. F is symmetric and transitive.

Def. 3.12 Intensional Relation:
An intensional relation, T, is defined by: d_i "is defined as" d_j where d_i is an item and d_j is a name. T is only transitive.

Thus, the order-defining transformation, $\mathcal{O}$, is defined as follows:

Def. 3.13 Order-Defining Transformation:
An order-defining transformation $\mathcal{O} \in S$ (*cf.* Def. 3.6) is a mapping from strings of data elements into REL:
$\mathcal{O}(d_1,d_2,\ldots,d_n) \rightarrow REL(d_1,d_2,\ldots,d_n)$.

Theorem 3.4: Transformation $\mathcal{O}$ partitions D.

Theorem 3.5: Transformation $\mathcal{O}$ partitions $\mathcal{D}$.

Postualte 3.4: The $\mathcal{O}$ transformation identifies patterns of data elements

4. The Role and Position of the Indexing System in the Communication Process

In Section 3 a general definition of the term *communication* has been given. In addition, some preliminary remarks have been made concerning the nature of the indexing operation. At this point, the position of the index in communication is viewed in terms of an adaptation of the Shannon-Weaver generalized communication scheme [2]. Definitions are now presented to

characterize the nature of the *transmission channel*.

First, let us consider the definitions of *indexing process, system* and *indexing system*.

Def. 4.1 Indexing Process:
The indexing process is characterized by the operations of identification (recognition) and representation of data elements and relations.

Def. 4.2 System: A system is that portion of the universe chosen for observation and measurement.

Def. 4.3 Indexing System:
An indexing system is a system for the application of the indexing process to the document space. The output from the indexing system is the index.

Now, we shall define the position of the indexing system in the communication process.

Def. 4.4 The Location of the Indexing System in Communication:
The indexing system is an intermediary between the transmission channel and the receiver.

The indexing system is affected by noise. The output of the indexing system, the index, is viewed as intermediary between the channel and the receiver (see Figure 4.1). The input to the indexing system is characterized as a document stream.

Def. 4.5 Document Stream:
The input to the indexing system from the communication channel is called a document stream which is defined as a heterogeneous collection of apparently un-related documents, ordered by their time of arrival at the indexing system.

For convenience of definition we shall grant the indexing system the ability to sample the document stream for fixed periods of time. Accordingly,

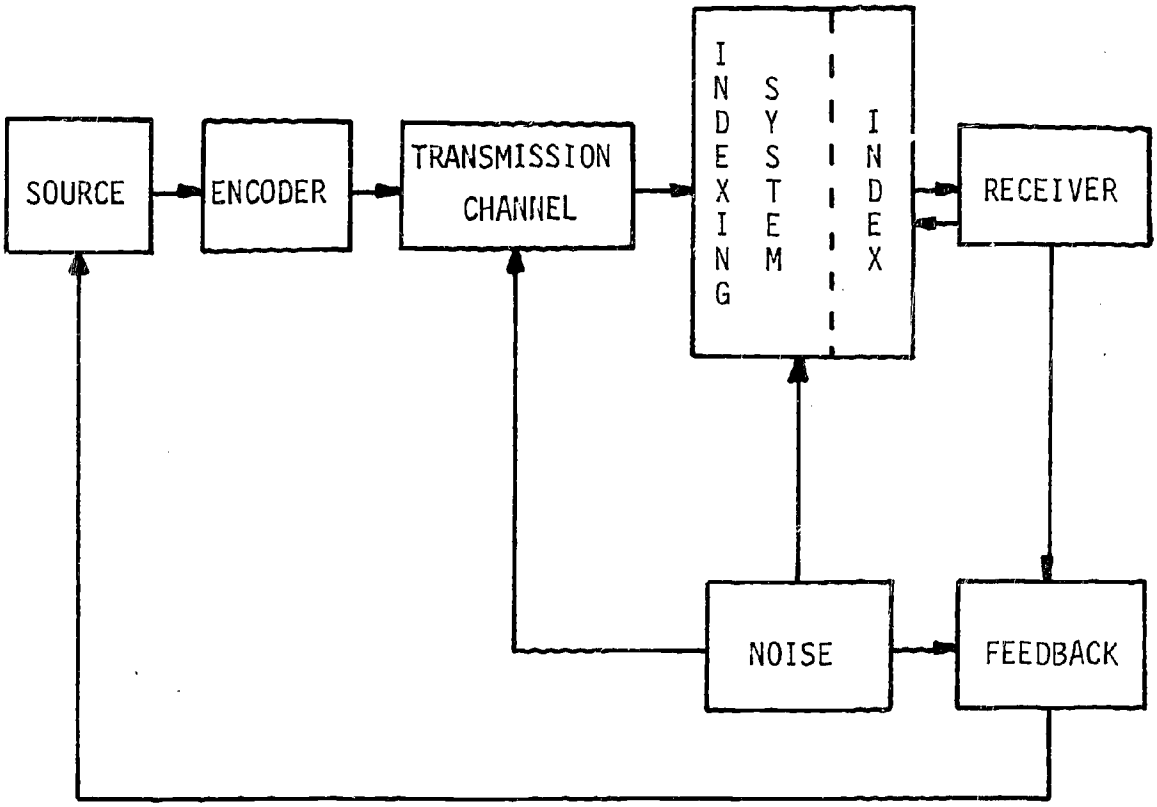


Figure 4.1: The Shannon-Weaver Model of Communication Adapted to Include the Indexing System and the Index. Note the Role of Noise and Feedback.

Def. 4.6 Input Time Slice:
An input time slice is a section of the document stream corresponding to a fixed interval of time t, $t \ll T$ (where T is the time required to receive the entire document under consideration), that is isolated for observation and processing.

Postualte 4.1: The indexing system must recognize relations (from REL) between data elements both within and between time slices.

Theorem 4.1: The indexing system recognizes inter- and intra-document data element relations.

The *role* of indexing can now be defined:

Def. 4.6 The Role of Indexing:
Indexing is a procedure for identifying relations that completely specify the flow of data in the document stream at any point in time.

Theorem 4.2: The indexing process is reversible: it must allow for the reconstruction of the original document flow.

Unfortunately, real-world indexing practices deviate considerably from the effective structure of the indexing system described above. The following postulate allows for the existence of error.

Postulate 4.2: Current indexing practices serve to obscure the unique organization between data elements in documents.

5. The Ordering Properties of the Index

In the first four sections of this overview of *indexing theory*, we have considered successively the definition of some fundamental concepts, the definition of communication, the nature of experience-set transformations, types of relations applicable to document representation, and, finally, the role and position of the indexing system in the communication process. Attention is now directed to a further characterization of the indexing

system, considering especially the *index* as a bi-directional interface between the document collection and the receiver.

A document has been described as an author-assembled, well-ordered collection of data elements. It is inferred that these data become information only when they are assimilated or put to use by the receiver(s). Accordingly,

Def. 5.1 Information: Information is defined as data elements of value in decision making.

Clearly, data elements must be available at the proper time and in the proper form to be of value in the decision-making process. To insure accurate data transfer, the indexing system must produce an index that is a facsimile of the system's parent documents. Thus,

Theorem 5.1: Accurate and complete document representation is the function of the indexing system.

The indexing system draws on a bipartite document space to effect this representation. The two components of the document space are defined as follows:

Def. 5.2 Input Documents: Input documents, $\mathcal{D}_i$, are documents which arrive at the indexing system *via* the transmission channel. These are the documents that the indexing system will represent.

Def. 5.3 Analysis Documents: Analysis documents, $\mathcal{D}_a$, are documents which describe the transformations, S. These documents reside permanently in the system and are used as aids in the representation operation.

The document space can now be described.

Theorem 5.2: $\mathcal{D} = \mathcal{D}_i \cup \mathcal{D}_a$.

The representation of an input document by the indexing system can be expressed as a set-product operation:

Theorem 5.3: Input-document representation = $\mathcal{D}_i \otimes \mathcal{D}_a = (\mathcal{D}_i) \cdot \text{REL}.$

It follows as a consequence of definition 2.12 that:

Theorem 5.4: $I = g(\mathcal{D}_i \otimes \mathcal{D}_a).$

The function, g, generates the index entries, where:

Def. 5.4 Index Entry: An index entry $i \in I$ is an expression such that the following data-element relation holds: $d_j \text{REL} d_k$. Where for each j, $\exists \{k\} \ni d_j \text{REL} d_k$ $\forall j,k$.

Figure 5.1 is a pictorial representation of these operations. It is interesting to note that this framework allows for a recursive definition of an index; an updated index, I_n, is formed through a combination of the old index, I_o, and the new elements of the document space:

Theorem 5.5: $I_n = (\mathcal{D}_i \otimes \mathcal{D}_a) \cup I_o.$

The operation of the indexing system is characterized by the *index space*, $\mathcal{I}$. The following definitions are required for this characterization.

Def. 5.5 Vocabulary: The vocabulary, V, is a set of possible data elements in a document space, ordered by precision of measurement. Subsets $V_i \subseteq V$ of this continuum describe those data elements recognized by a particular indexing system.

Def. 5.6 Transmission Decoding: Transmission decoding, TD, is a set of possible productions defining strings of data elements over V. Subsets $TD_i \subseteq TD$ of this continuum describe those productions employed by a particular indexing system.

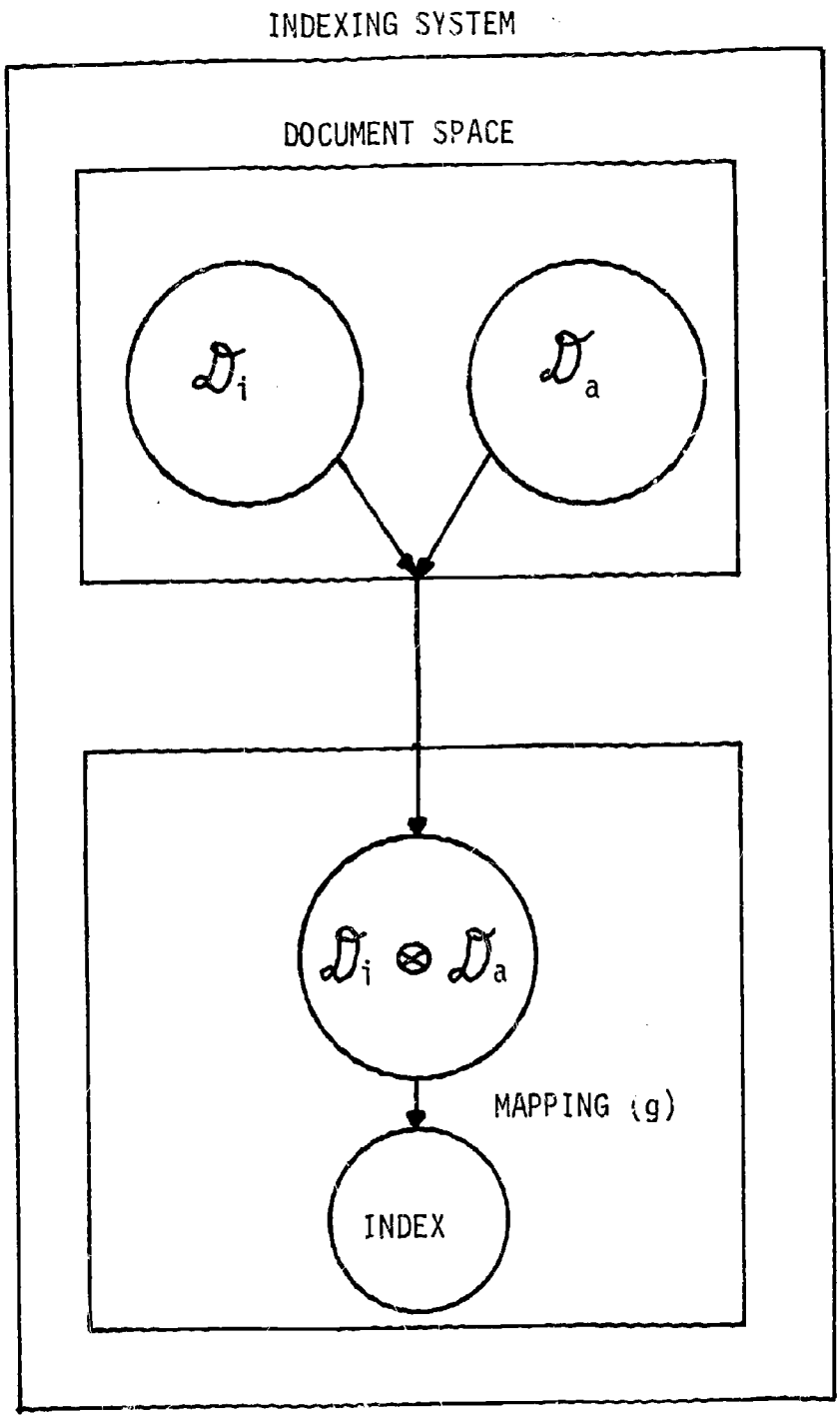


Figure 5.1: The Indexing System.

Def. 5.7 Language: Language, L, is a set of possible expressions (strings defined by TD plus relations from REL). Subsets $L_i \subseteq L$ of this continuum describe the expressions employed by a particular indexing system.

The index space is now alternatively defined as follows: (*cf.* Def. 2.11).

Def. 5.8 Index Space: The index space, $\mathcal{J} = V \times TD \times L$.

The concept of an index space provides a useful framework for analyzing the retrieval process. A specific request initiated by the receiver must be formulated as an element of the index space:

Theorem 5.6: $Q \subset \mathcal{J}$, *i.e.*, $V_Q, TD_Q, L_Q \in \mathcal{J}$.

Thus,

Def. 5.9 The Process of Retrieval: A homomorphic mapping of the request data-elements and relations into the index space.

Consequently, we have the following homomorphic mappings:

Theorem 5.7: $Q \rightarrow \mathcal{J} \rightarrow I \rightarrow \mathcal{D}$, $\mathcal{D} \rightarrow I$

Corollary 5.1: There exist as many homomorphic mappings ($Q \rightarrow \mathcal{J}$) as their exist individual receivers in communication during a specified time interval.

Corollary 5.2: I is a bi-directional interface between Q and $\mathcal{D}$.

6. Indexing as an Entropy-Reducing Operation

We now consider an alternative way of characterizing the operation of the indexing system, namely, that the indexing system increases the order of the data elements in the document space. More explicitly, the specification of a structure upon which measurement is effected yields a reduction in thermodynamic entropy by increasing the intrinsic order of the

system under study. The imposition of an explicit order (*i.e.*, order relations selected from REL) upon the elements of the structure also amounts to a decrease in communication entropy. For the moment we shall content ourselves with a fuzzy definition of entropy (the reader is referred to the parallel section 15 for an overview of the alternate definitions of "entropy"):

Def. 6.1 Entropy: "... a measure of the lack of information about the actual structure of the system." (Brillouin [3])
or, A measure of the incompleteness of the data from which we infer the state of the system.

Documents, $\mathcal{D}_i$, that arrive at the indexing system are (ignoring chronology) in a highly disordered state because there exist no overt data-element connections across document boundaries. Accordingly,

Postulate 6.1: The indexing system recognizes and makes explicit inter-document data-element relationships.

The indexing system, in its organization and recognition operations, defines a *phase space* of data elements intermediary between the document space and the receiver. Thus,

Def. 6.2 Phase Space: A phase space is a definition of the accuracy of measurements based on the division of the document space into well defined units.
or, The specification of two document coordinates:
a) configurational coordinates that depict which data are stored, and
b) momentum coordinates that determine the particular sequence of configuration coordinates involved in the document representation.

The two coordinates of the *phase space* describe the storage and search operations associated with the use of the index. Consequently,

Theorem 6.1: The phase space is isomorphic with the index space, J .

The ordering of data elements, by means of a phase space, amounts to a reduction in entropy:

Theorem 6.2: The order-preserving and increasing properties of an indexing system amount to a reduction in the entropy of the document-space/document-space-searcher system.

Since such a reduction in entropy must be accompanied by an increase in entropy (*i.e.*, by an expenditure of energy) elsewhere in the system, we have:

Theorem 6.3: The entropy decrease which results from the creation of the index, is balanced by the entropy increase associated with the effort needed to obtain the coordinates of data elements in the phase space.

Finally, we postulate a relationship between the work expended in indexing (the specification of phase space coordinates) and the information desired:

Postulate 6.2: The probability of a given set of data elements becoming information is a function of the work expended by the indexing system.

7. The Concept of Benefit

We have so far been concerned with the recognition and representation of data by the indexing system. It has been emphasized that data must be in the proper form and must be available at the proper time to be of use to the receiver (decision maker). When the conditions of form and availability are fulfilled, we say that the data becomes information. However, there remain the questions: What is *data of value?* and How is the searcher to *benefit* from the existence of such information? The answers to these questions are found in a consideration of the concepts of *goal, hypothesis testing* and *decision making.*

It is assumed that a *goal* represents a desired end product or end state of the receiver. A goal may be as simple as "the retrieval of any document on subject X" or as complex as "the winning of a game of chess." Thus,

Def. 7.1 Data of Value: Data are of value when they are used in the accomplishment of a goal.

In the retrieval process, the goal of the searcher is achieved through a hypothesis-testing and decision-making chain. Hypotheses are posed by the receiver concerning the data store (*e.g.*, concerning the contents of the document space) and the retrieved data *may* provide information leading to the decision which results in goal achievement. Figure 7.1 shows a structure of possible goal-directed paths of which we define two extreme cases:

Def. 7.2 Path of Maximum Benefit: The H - D - G path is the path of maximum benefit where H = hypothesis, D = decision and G = goal.

Def. 7.3 Path of Minimum Benefit: The H - H - H ... hypothesis path, denoted by H - H - H ... is the path of minimum benefit.

Clearly a decision must be made in the minimum benefit case to formulate a new hypothesis based on the data retrieved in support of the previous hypothesis. But this decision will be treated as less significant than the goal-achievement decision associated with definition 7.2. Thus,

Def. 7.4 Meta-Decision: A meta-decision is a decision which does not lead directly to goal achievement (frequently associated with the progression between hypotheses).

and, in addition:

Def. 7.5 Meta-Information: Meta-information is data elements of value in meta-decision making.

Clearly,

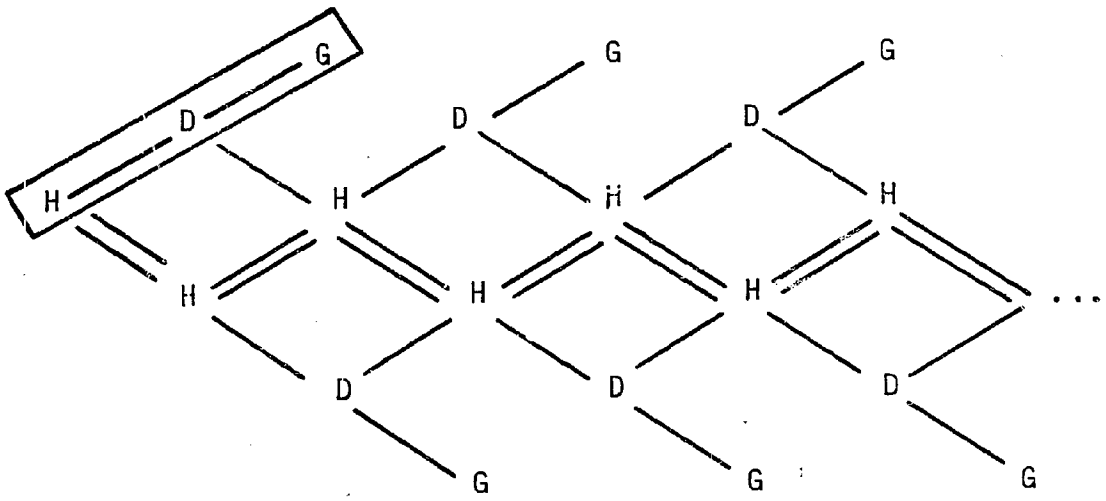


Figure 7.1: The Hypothesis-testing and Decision-making Chain. The double-bonded path represents the minimum benefit case; the enclosed path represents the path of maximum benefit. (H = hypothesis, D = decision, G = goal)

Theorem 7.1: Information is not equivalent to meta-information.

At this point we are better equipped to define the concept of benefit:

Def. 7.6 Benefit: Benefit is a relationship between the information obtained and the number of decisions required to reach a goal.

In addition, benefit, B, obtained from indexing is accumulated over a considerable interval of time (measured at times t_i), thus:

Theorem 7.2:

$$B = \sum_{i=1}^{N} B_i ,$$

where B_i is benefit measured at time t_i and,

$\exists N \ni B_i = 0 \quad \forall i > N.$

8. Theoretical vs. Real-World Indexes

Indexing systems are, by definition, imperfect because the associated ordering measurements (classification) are inherently uncertain. The Heisenberg uncertainity principle [4] applies to the specification of elements of the indexing phase space. Thus, there is always the possibility of misinterpretation (misrepresentation) of data elements. Clearly, then, there is a certain amount of "noise" or "error" built into an indexing system because of the inherent limitations of the associated classification method.

Based on the premise that indexing systems are imperfect, one must be able to distinguish between "perfect" and "imperfect" indexing systems. This distinction is sharpened through the definition of *theoretical* and *real-world* indexes. Thus:

Def. 8.1 Theoretical Index: The theoretical index represents all inter- and intra-document relations between data elements in the document space. Order-

preserving operations are employed at all steps of the indexing process.

If we assume that there are d data elements in a given document of the collection, then the theoretical index must permit at *most* the existence of 2^d connections between these data elements. But, since there are many documents (say m of them) in a given document space, one must allow for the existence of many more data-element relationships.

Theorem 8.1: The theoretical index must be able to represent any subset of $\{2^{d_1} \ldots 2^{d_m}\}$ data-element relations.

The real-world index is now defined:

Def. 8.2 Real-World Index: Real-world indexes contain, for a given document space, a number of valid index entries (*cf.* Def. 5.4) N_R such that $N_R \ll N_T$, where N_T is the number of valid index entries contained in the theoretical index for the same document space.

Thus,

Postulate 8.1: For a given $\mathcal{D}$, real-world indexes fall short of the theoretical index because the indexing of the document space is incomplete. (*cf.* Postulate 4.2).

9. The Human Limitation

The presentation, up to now, has been concerned with a systematization of information storage and retrieval by means of a theory of indexing. This theory rests essentially on a formalization of the notions of data element and relations. However, the implementation and modeling of an information storage and retrieval system are not simply abstract constructs, but are engineering processes involving the human factor. This section will

consider the nature of the interface between the indexing system (the index) and the receiver, and show that the abstract construct of an ideal index must be tempered by a fuzzy theory of human query-formulation and decision-making processes.

Very rarely is the receiver, who utilizes a real-world index, completely satisfied with the result of an initial query. Either he has an incomplete understanding of the organization of the system or he is unable to adequately formulate a hypothesis about its contents. The following definition of query is an extension of Definition 2.13.

Def. 9.1 Query: A query is a hypothesis about the contents of the document space, $\mathcal{D}$. (*cf.* Def. 2.13).

Postulate 9.1: The maximal and minimal paths (Def. 7.2 and 7.3) of inquiry have a small probability of occurrence.

Consequently,

Theorem 9.1: The first data element retrieved, in response to an initial query, is likely to be only partially beneficial. (*cf.* Def. 2.13).

Corollary 9.1: Benefit can only be maximized through repeated interaction between the receiver and the index.

In the maximal benefit case (Def. 7.2), the data element that provides the information is said to have maximal *utility* or *value*. However, in any intermediary case, the utility of an information-providing data element is decreased because the data element is one of a sequence of retrieved data elements. Thus,

Def. 9.2 Decay: A data element received at time t_{i+1} has lower value than the same data element received at the time t_i. (Here time is measured relative to the start of the interaction between receiver and index.) This

decrease in value is called *decay*.

Theorem 9.2: The decrease in utility of a data element is directly related to its position in a string of retrieved data elements.

Corollary 9.2: The utility of any data element decreases with the number of hypothesis-testing and decision-making steps which preceed its retrieval.

Corollary 9.3: The utility of the last data element used to reach a goal is a function of the benefits derived from the use of the previous data elements.

It is postulated that data elements exhibit a Poisson-like behavior in their role in decision making. Consequently, the value of any data element in decision making diminishes with time (see Figure 9.1). However, as the hypothesis becomes more specific, the rate of loss of utility of a data element also decreases (see Figure 9.2).

Postulate 9.2: The value (utility) of a data element, with respect to goal achievement, is Poisson distributed.

Postulate 9.3: Data elements are indistinguishable with respect to their value distribution:

$$\text{value}(d_1)_{t_n} = \text{value}(d_2)_{t_n}$$

and, finally

Postulate 9.4: The rate of decay of the utility of newly retrieved data elements decreases with increasing path length in the H - D - G structure (Figure 7.1).

10. Interregnum

To R. L. Collison [5], "The trouble with indexing is that even today we are still at the elementary stage of learning how to do it. We do no know enough about its technique" and we certainly do not know enough about

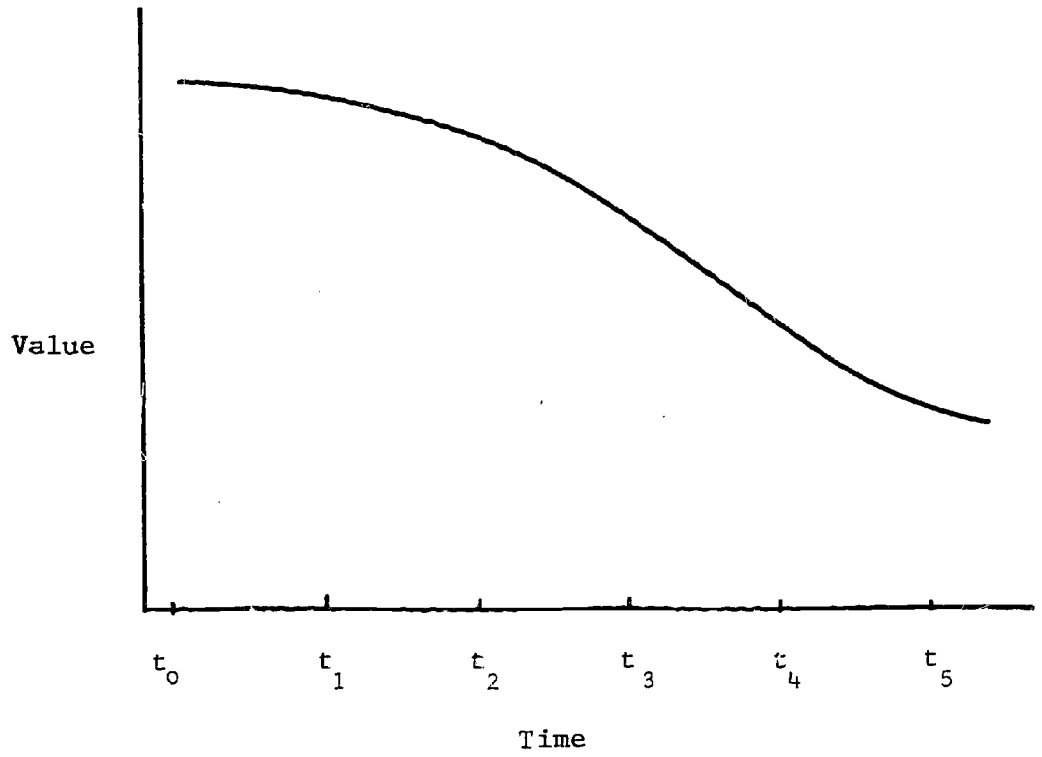


Figure 9.1: Data-element-value Distribution over Time.

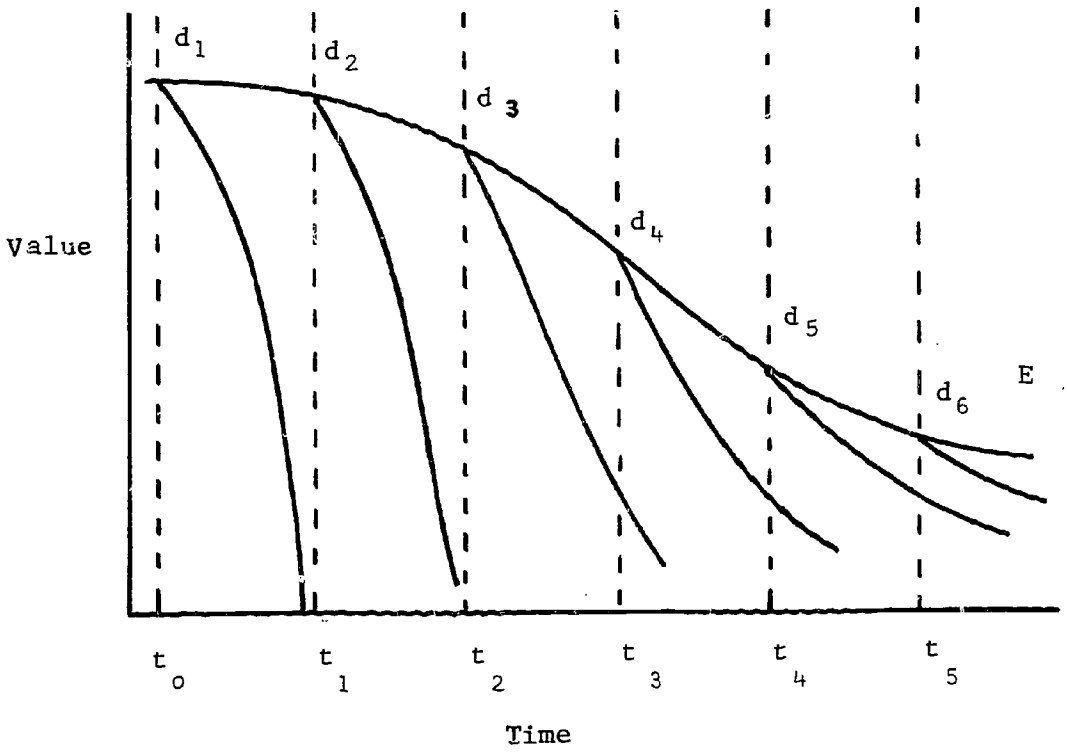


Figure 9.2: Data-element-value Decay as a Function of Successive Interactions Between User and Index.

its theory. Indexing, and its associated paraphernalia, constitute a strange process. Consequently, the researcher is confronted by an interesting situation: on the one hand examples of the product of indexing, the index, are plentiful and ubiquitous; on the other hand, attempts to formalize either the process of indexing or the relationship between its exemplars are virtually nonexistent. The previous nine sections of this chapter constitute an attempt to remedy this situation by presenting a formal description (and interpretation) of the "indexing process".

The exposition of the theory was designed to be brief and terse, consequently, a summary is not easily presented. As a form of summary, the postulates presented in the previous sections are listed, as a group, as being indicative of the scope of the theory presented.

2.1: $I = f(\mathcal{D}, \mathcal{J})$, where f is the *indexing* process.

2.2: Accurate retrieval depends upon the exactness of the indexing.

3.1: The items transfered from element to element in communication are data elements and associated relations.

3.2: Any theory or practice of communication which causes a loss of data elements, either through their misrepresentation or by restricting their flow, must be considered inadequate.

3.3: The *indexing system* provides the interface experience set and the transformations required for effective communication.

3.4: The $\mathcal{O}$ transformation identifies patterns of data elements.

4.1: The indexing system recognizes inter- and intra-document data element relations.

4.2: Current indexing practices serve to obscure the unique organization between data elements in documents.

6.1: The indexing system recognizes and makes explicit inter-document data-element relationships.

6.2: The probability of a given set of data elements becoming information is a function of the work expended by the indexing system.

8.1: For a given $\mathcal{D}$, real-world indexes fall short of the theoretical index because the indexing of the document space is incomplete.

9.1: The maximal and minimal paths of inquiry have a small probability of occurrence.

9.2: The rate of loss of utility of a data element is Poisson distributed.

9.3: Data elements are indistinguishable with repsect to their value distribution.

9.4: The rate of decay of the utility of newly retrieved data elements decreases with increasing path length in the H - D - G structure.

For further clarification the reader is referred to Figure 10.1. This figure presents an overview of the various schema associated with the indexing theory. The conceptual steps that lead from generalized communication to the characterization of the indexing system are depicted. The final level of analysis, the index space, $\mathcal{I}$, is interpreted as a representation of the operating limits of the indexing system.

It is interesting to note that the discussion of the previous sections has been predicated upon the existence of three conceptual classes: sets of documents, sets of attributes, and sets of relationships expressing a connection between documents and attributes. These are the fundamental entities of any IS&R system and must be incorporated in the characterization of *effective communication*. The ideal index has been chosen as the standard for effective indexing. Albeit unobtainable, the ideal index serves as a useful comparative device. By analogy, the ideal index operates in a manner similar to the ideal game player (adapted from Garfinkel [6]): He never overlooks a message; he extracts from the message all the data it bears; he names things properly and in the proper form; he never forgets;

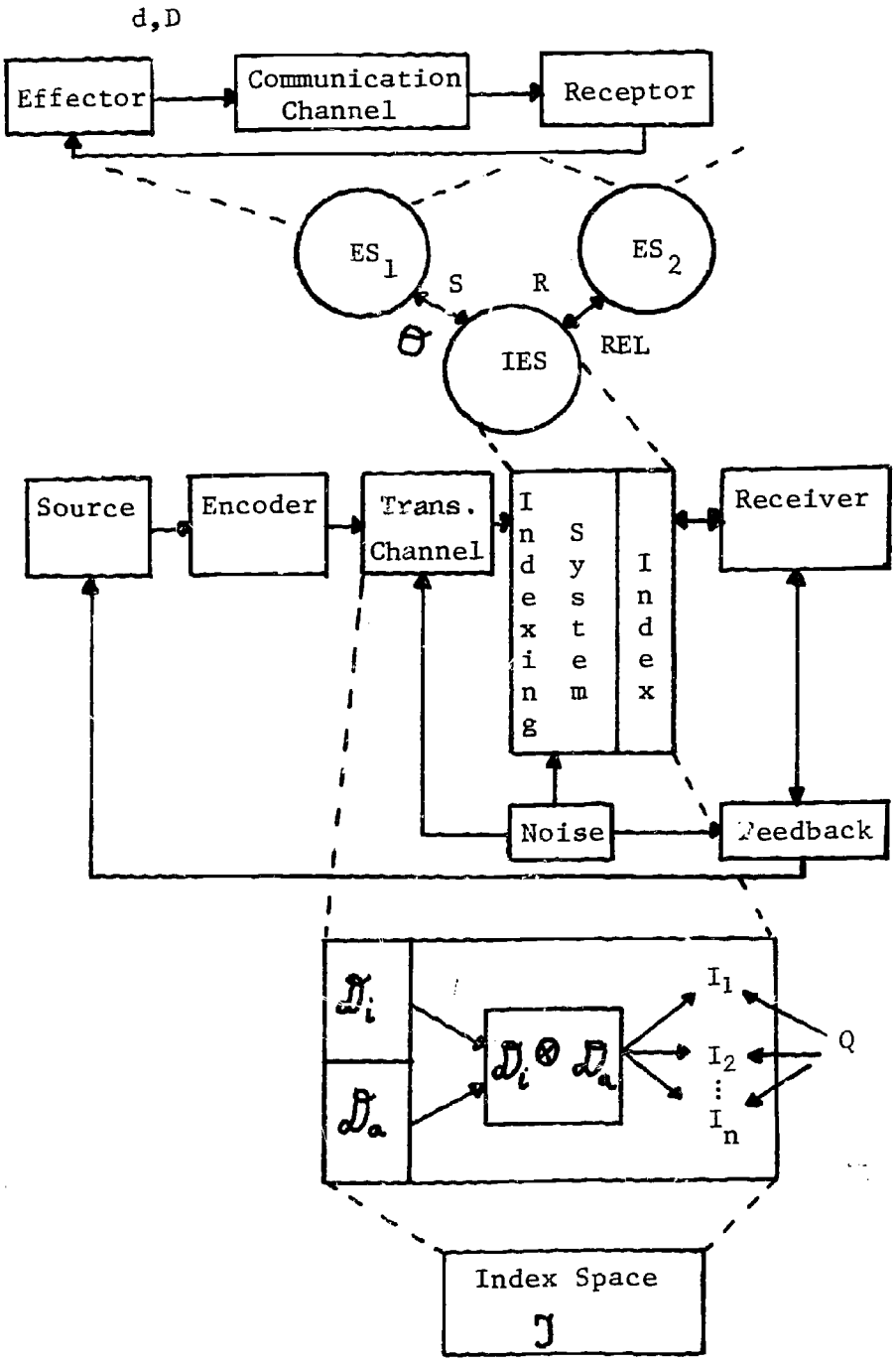


Figure 10.1: An Indexing-theory Schema.

he stores and recalls without distortion; he never acts on principle but only on the basis of an assessment of the consequences of a line of conduct for the problem of maximizing the chances of the effect he seeks.

A theory of indexing must obviously account for error but, more importantly, it must provide guidelines for the maximization of document representation fidelity. The next eight sections (each one parallel to a section of the overview part of this Chapter) will present arguments for and a further exposition of the indexing theory. The goal is to at least partially establish isomorphism between real-world-indexing practices and the interpretations of these practices embodied in the theory.

11. Data Element as the Basis

The theory of indexing that has been presented in the previous sections has relied heavily on the concept of *data element*. It has been assumed that data element is the fundamental unit of documentation and, accordingly, provides the basis for many of the concepts and relationships developed in the theory. Following Sorgel [7] (who was concerned with the concept of *keyword*), three important features of the concept of data element can be identified:

1) The concept of data element allows for independent manipulation.

2) A data element does not decompose into two or more units.

3) A data element has a definite meaning or interpretation.

These features were incorporated (albeit implicitly) into the definition of data element (*cf.* Def. 2.5) and were viewed as consequent to the definitions of *measurement*, *attribute*, *unit of measure* and *precision*. The presentation in Section 2 began, rather abruptly, with the definitions of measurement and attribute. As an alternative, and to counter a possible objection that

these first definitions were "pulled from the air", we shall consider a more formal development of the concept of measurement, a concept that is antecedent to data element.

Before undertaking the further development of *data element*, let us introduce a document by means of which the various concepts discussed may be exemplified. This document is shown in Figure 11.1 together with examples of index entries involving differing definitions of data element. Many alternative derivatives of this document appear throughout the remainder of this chapter.

Let us adopt an essentially mechanistic view of the world and consider that all events (the word "event" is left to the reader to define) are the outputs of machines. Accordingly,

Machine: A machine is a black box which accepts inputs and emits outputs, (see Figure 11.2).

Thus an output, or event, is somehow paired with an input by means of a "black box." Although such a definition is all inclusive, it offers little in a descriptive sense. For increased specificity the following definition incorporates a theory of the operation of the black box:

Turing Machine: A turing machine, Tm, is denoted by
Tm = $\{K,\Gamma,\delta,\sigma,F,q\}$ where:
K is a finite set of states;
Γ is the finite set of symbols from which the inputs and outputs are obtained;
δ: $K \times \Gamma \rightarrow K$ is the next state function;
σ: $K \times \Gamma \rightarrow \Gamma$ is the output function;
$F \subseteq K$ is the set of final states; and
$q \in K$ is the start state.

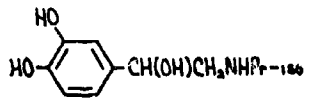
The next three definitions arise immediately from that of turing machine:

Observables: Observables are elements of Γ.

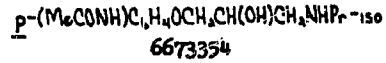
Lancet

820 EFFECT OF A SELECTIVE BETA-ADRENERGIC BLOCKER IN PREVENTING FALLS IN ARTERIAL OXYGEN TENSION FOLLOWING ISOPRENALINE IN ASTHMATIC SUBJECTS.
LANCAo,2,7630,69,1092-3
Palmer KNV, Legge JS, Hamilton WFD, Diament ML;Dep. Med., Univ. Aberdeen, Aberdeen, Scot.

6673354 4-(2-Hydroxy-3-isopropylaminopropoxy)acetanilide (practolol)
(20 mg/subject, i.v.), a β-ADRENERGIC BLOCKING agent
selective to the HEART, prevented the decrease in ARTERIAL
OXYGEN TENSION in 11 ASTHMATIC patients following
7683592 isoprenaline (0.1 mg/subject, aerosol inhalation) treatment
without significantly decreasing the BRONCHODILATOR action
of
isoprenaline.



7683592



6673354

Figure 11.1: An Example Document
[CBACA$_3$, vol. 11(2), 1970, p. 119]

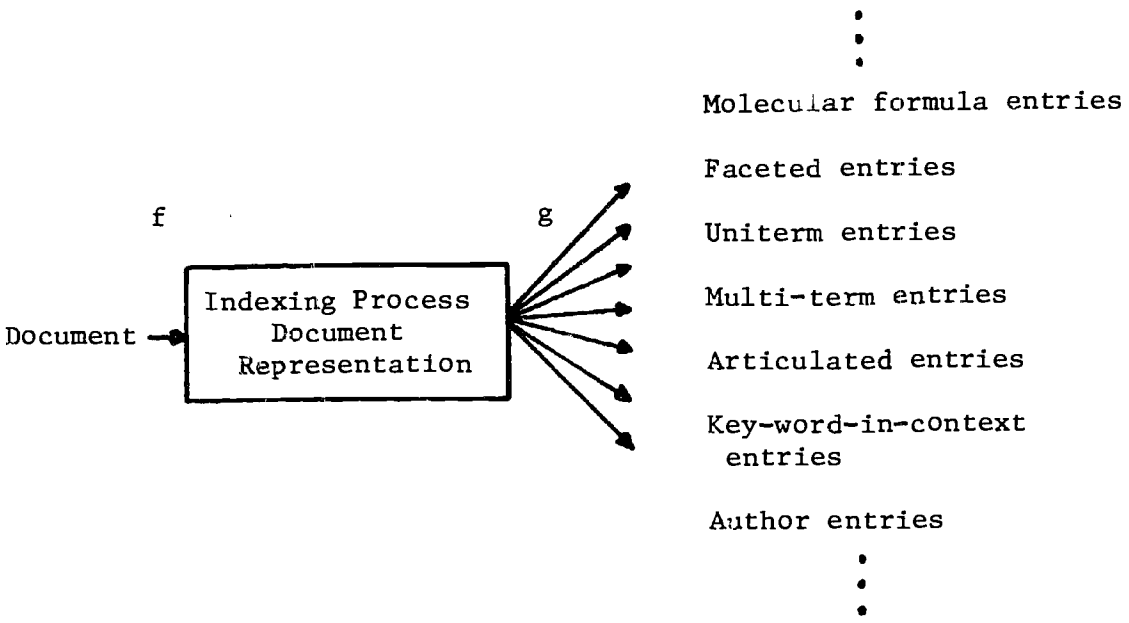
(Reproduced with the permission of the
Chemical Abstracts Service, Columbus, Ohio)

LANCET, 2, 7630, 69, 1092-3.

1069 tokens
408 types

Index Entry (KWIC)	Frequency in text
Acetanilide	1
Adrenergic	2
Aerosol	1
Arterial	7
Asthmatic	4
Blocker	1
Bronchodilator	2
Heart	1
Hydroxy	1
Inhalation	1
Isoprenaline	14
Isopropylaminopropoxy	1
Practolol	9
Tension	3

Figure 11.1 (cont.): Frequencies in Original Text.



Where the following are examples of the index-entry transformation, g:

g_a: Molecular formula entries

820 $C_{14}H_{22}N_2O_3$ 6673354

820 $C_{11}H_{17}NO_3$ 7683592

g_b: Faceted entries

820 Receptor, beta(1)
820 Receptor, beta(2)
820 Drug, beta-blocking
820 Muscle, Bronchial
820 Asthma, Bronchial

Figure 11.1 (cont.): Example Indexes.

g_c: Uniterm entries

820 Isoprenaline
820 Receptors
820 Drug
820 Myocardium
820 Myocardial
820 Contractility
820 Oxygen
820 Tension

g_d: Multi-term entries

820 Beta-blocking Drug
820 Myocardial Contractility
820 Bronchial muscle
820 Oxygen tension
820 Bronchodilator activity
820 Blood-gas tension
820 Bronchial Asthma

g_e: Articulated entries

820 beta-adrenergic blocker
- effect of, in preventing falls following isoprenaline in asthmatic subjects
- in preventing falls following isoprenaline in asthmatic subjects, effect of

820 asthmatic subjects,
- effect of beta-adrenergic blocker in preventing falls in, following isoprenaline
- following isoprenaline, effect of beta-adrenergic blocker in preventing falls in

Figure 11.1 (cont.): Example Indexes.

g_f: Key-word-in-context entries

sopropylaminopropoxy) ACETANILIDE(practolol)(20 mg/subject, i.v.
OF A SELECTIVE BETA- ADRENERGIC BLOCKER IN PREVENTING FALL
/subject, i.v.)a β- ADRENERGIC BLOCKING agent selective to the
line (0.1 mg/subject, AEROSOL inhalation) treatment without signifi
N PREVENTING FALLS IN ARTERIAL OXYGEN TENSION FOLLOWING ISO
ented the decrease in ARTERIAL OXYGEN TENSION IN 11 ASTHMAT
OXYGEN TENSION in 11 ASTHMATIC patients following Isoprenaline(o
OWING ISOPRENALINE IN ASTHMATIC SUBJECTS.
CTIVE BETA-ADRENERGIC BLOCKER IN PREVENTING FALLS IN ARTERI
cantly decreasing the BRONCHODILATOR action of Isoprenaline.
gent selective to the HEART, prevented the decrease in ARTERIAL O
4-(2- HYDROXY-3-isopropylaminopropoxy)acetanilide(
1 mg/subject, aerosol INHALATION) treatment without significantly
IC patients following ISOPRENALINE (0.1 mg/subject, aerosol inhal
GEN TENSION FOLLOWING ISOPRENALINE IN ASTHMATIC SUBJECTS.
NCHODILATOR action of ISOPRENALINE.
4-(2-Hydroxy-3- ISOPROPYLAMINOPROPOXY)acetanilide(practo
propoxy)acetanilide (PRACTOLOL)(20 mg/subject, i.v.), a β-ADRE
LS IN ARTERIAL OXYGEN TENSION FOLLOWING ISOPRENALINE IN AST
se in ARTERIAL OXYGEN TENSION IN 11 ASTHMATIC patients following

g_g: Author entries

Diament ML, 820
Hamilton WFD, 820
Legge JS, 820
Palmer KNV, 820

Figure 11.1 (cont.): Example Indexes.

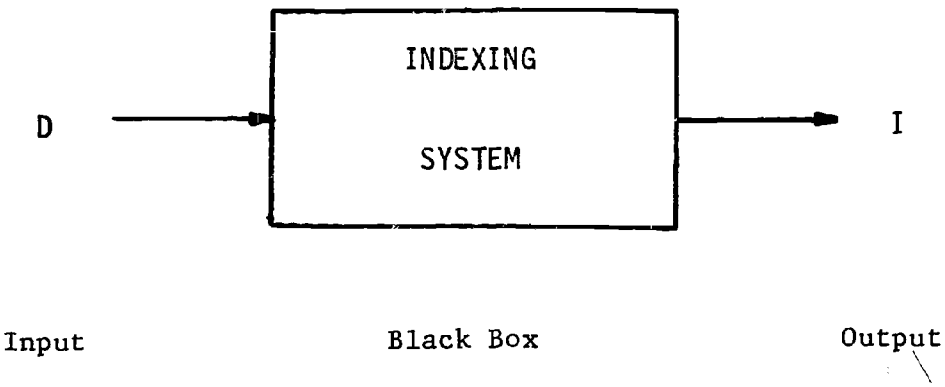


Figure 11.2: The Indexing System.

Procedure: A procedure is a turing machine.*

Attribute: An attribute, A, is a subset of the set of all possible observables associated with a procedure. $A \subseteq \Gamma$ (*cf.* Def. 2.2).

With the addition of Definitions 2.3 and 2.4 (unit of measure and precision) the following definition of measurement can be presented.

Measurement: Measurement is a procedure which:
a) Isolates the attribute;
b) Applies a unit of measure to the attribute; and
c) specifies a precision.

We shall define the result of measurement as *data*. Clearly, the data obtained must be a subset of the set of symbols, Γ, associated with the turing machine that embodies the attribute under observation. Following Definition 2.5, then, a data element is the smallest datum in the class of data arising from the repeated measurement of an attribute. Accordingly, a data element is the smallest datum in a well-ordered (*cf.* Def. 2.8) set of data and serves as the *differentia* for class membership.

From the derivation of the definition of data element, it should be clear that a data element can be any desired entity. The essential point is that the specification of a data element must be accompanied by the description of the associated measurement. This means that when referring to a data element, one also refers to the name of the attribute measured, the unit of measure and the precision of measurement. Any omission yields a meaningless entity. For example, the statement "the building is 21" is a meaningless statement since "21" is not defined. The building could just as easily be 21 years old as it might be 21 stories tall. On the

---

* Notice that the term *theory* can not be defined as a set of definitions associated with a procedure.

other hand, to say that a *word* is a data element demands a meaningful specification such as: "a string of characters delimited by blanks." The attribute is a string of characters, the unit of measure is a non-zero distance between blanks and the precision of measurement is the recognition of a character. Similarly, the data element *character* might be specified as " a unique and unambigous pattern of bits of length six". We could continue to cite examples, but it should now be clear that an infinite variety of data elements could be identified. Fortunately, the set of data elements which must be dealt with is finite since the measuring (recognition) devices associated with a given indexing or retrieval system have a finite (manageable) number of outputs.

The possibility of a data element being any desired entity is a convenience both from a descriptive and a theoretical view point, for on the one hand it becomes possible to identify a continuum of data elements which includes characters, words, strings of words, titles, sentences, abstracts, full documents, numbers, frames of film, varying lengths of video and audio tape, to name but a few. On the other hand, the specification of an indexing method or system serves to define those data elements that can be recognized and subsequently processed by the system. The obverse of this statement is also valid: the specification of a data element defines those systems capable of representing it. Consider the following as an example of the defining role of data element. For most automatic classification and indexing systems, the data element is defined as a word (keyword); however, if the data element is re-defined to be a character then a new *theory* of the classification process is obtained--*e.g.*, error detection in spelling.

An understanding of the concept of data element (its definition is an extension from present-day, computer-orientated usage) leads one quickly into the concepts of document, document space, index, index space and query. These concepts will be treated in detail in subsequent sections, but an introduction to their importance is provided as follows.

A document is viewed, by this author, as a well-ordered set of data elements. Although a data element can, as we have seen, be any desired entity, it is frequently associated with the definition of word, clause and/or sentence. These are the units of written-document communication. Data elements of this type are well ordered by their physical position or occurrence within the document. In addition, some subsets of the data elements of a document are well ordered with respect to membership in a classification hierarchy, where the ordering relationship is denoted by genus-species. Section 12 contains a more detailed discussion of the significance (and utility) of data-element relations.

The input documents to the indexing system (and to the indexing process) are represented by the document space. These documents are ordered by their time of arrival at the indexing system, or, possibly, by subject content (a shared data element and/or relation). It should be clear that some subsets of documents in the document space can be well ordered. Furthermore, an important feature of the document space, as a collection of documents, is that some documents will contain identical data elements and data-element relationships. The recognition of document similarity is one of the essential functions of the indexing process; this process will be described in Section 13.

The index is the output from the indexing process and it is postulated to arise as a function of both the document and of the indexing process. Also, both the index and the description of the indexing process characterize the operation of the indexing system. We will just mention, at this point, that the indexing system is further characterized by a description of its index space which is itself a representation of those data elements and relations that can be recognized by the system.

Hence, Theorem 2.1 (Proof): A representation of permitted data elements and relations is itself a well-ordered set of data elements, hence a document. The document space is viewed as an R-set (following Russell's notation [8]--that is, it is a set that contains its own description), hence its description, $\mathcal{J}$, is a member of $\mathcal{D}$.

It should be obvious, following the remarks of the previous paragraph, that if the index is the result of successive order-preserving mappings performed on the document space, then the data elements contained in the index must preserve the original data-element/relation structure of the document space.

Hence, Theorem 2.2 (Proof): An index, by definition, must preserve the well ordering of its parent documents and the ordering of the document space. Thus, an index is a well-ordered set of data elements (trivially well-ordered by alphabetization).

It follows therefore that the purpose both of the indexing process and of the creation of the index center on the representation and subsequent retrieval of documents. In fact, the fundamental assumption of this section is that accurate retrieval depends on the exactness of document representation in the indexing process. Furthermore, it is intuitively reasonable to assume that a document should have the same representation both in

storage and in retrieval. This observation is thus the basis of the definition of a query as a well-ordered set of data elements which is a proper subset of the index.

Hence, Theorem 2.3 (Proof): If a query is a well-ordered set of data elements then it is, by definition, a document.

12. Communication and Indexing -- II

12.1 Communication

Information Science is endowed with a multitude of models and definitions of communication. The various views of communication can be conveniently classed, following Weaver [9], as describing either technical, semantic or pragmatic information transfer. While such models are useful in the description of specialized modes of communication, we have chosen to introduce the communication function of indexing in terms of a more generalized view of communication. Cherry's [10] definition embodies such a general view:

> Communication. Broadly, the establishment of a social unit from individuals, by the use of language or signs. The sharing of common sets of rules, for various goal-seeking activities.

The really important point brought to light by this definition is that communication involves the *sharing* of behavioral elements. From a sociological point of view, the sharing of behavioral elements leads to shared agreement, common understanding* and, finally, concerted action between the communicants. Furthermore, the concept of sharing is implied in the view that communication is the relationship between the transmission of stimuli and the evocation of responses [11].

---

* The term common understanding is due to Garfinkel [12] and is further explicated in Landry, Meara, Pepinsky, Rush and Young [13].

We have chosen to model the mechanism (or everyday setting) which permits the sharing of behavioral elements by a closed system consisting of an effector (the source of the stimulus), a receptor (the recipient of the stimulus and the source of the response), a transmission channel and a feedback unit (see Figure 3.1). It is postulated that the items which are transfered or "communicated" are data elements and associated relations. An interesting confirmation of the above view of communication comes from Pierce's philosophy of Pragmatism. In the late nineteenth century, Pierce [14] posited the triadic nature of every sign situation. The triada were designated as "sign-designatum-user" or, "sign--that-which-is-refered-to--user" and embodied the view that communication involves the expression of the *intent* of the sign. Consequently, a sign (*i.e.*, a collection of symbols) never stands in isolation, but must possess a relationship to other signs. The acknowledgement of the "understanding" of the intent of sign relationship comes from the feedback elicited from the original receiver. In Pierce's terms, this is the *development* of the sign. As a final observation, the definition of data element (Def. 2.5) is a triad involving the specification of the relationship between an attribute, a unit of measure and a precision. Thus, viewed as triads, data elements are the correct elements of communication, at least in the sense of Pierce.

12.2 Experience Set

The elements of communication are data elements. To go one step further it is assumed that the "memory" of the source and the receiver can be adequately modeled as an ordered set of data elements and relations. We shall refer to the ordered set as an *experience set*. Hence,

Theorem 3.1 (Proof): The data elements and relations of the experience set are well-ordered with respect to their order of insertion into the "memory structure". Hence, by definition, the experience set is a document.

Data elements and relations are selected from the source's experience set for transmission and, upon reception, these same data elements and relations are evaluated in terms of the receiver's experience set. The data elements and relations selected for transmission constitute what I have called the interface experience set (IES). In a sociological sense, the elements of the IES are the participant's "informative displays"*. As we have defined communication, it can involve transmission between any combination of men and machines. For example, in communication between a programmer and a computer the IES is some programming language (see Figure 10.1).

An interesting alternate model of the experience set is provided by Mackay [15]. He argues that the pertinent states and relations (we call these data elements and relations) are represented by a conditional probability matrix (CPM). The transition probabilities of the CPM indicate those relations recognized by the particular experience set. To Mackay, the meaning of a communicated data element can only be evaluated in terms of a change of state (or probability) in the CPM. Consequently, the source decides whether the receiver has properly interpreted the "meaning" of the communication by carefully observing its effect (*e.g.*, the response). In this way the source (and, in return, the receiver) draws inferences about the CPM modification. The concept of a conditional probability matrix will be further developed in Chapter 5 by means of the construct called a *hypothesis structure*.

---

* See Landry, *et al.* [16] for a discussion of "informative display."

The minimum condition for effective communication is that there be some overlap between the participant's experience sets. Overlap will permit an informative display signaled by one of the participants to be properly interpreted (and acted upon) by the other participant. Of course, the reliability of the interpretation depends on the degree of commonality between the respective experience sets and the IES. Hence,

Theorem 3.2 (Proof): Effective communication is maximal when there is commonality between the ES's and the IES for all possible messages -- *e.g.*, when $(ES)_s = (ES)_r$.

and,

Theorem 3.3 (Proof): The IS&R-system user cannot know precisely which data elements are stored in the system. however he must understand how the system stores, organizes and represents documents. Effective communication with the system is achieved when the system IES relations and transformations are known and understood by the user.

12.3 Transmission Analysis and Indexing

To this point we have been concerned with a generalized presentation of the concept of communication. Although somewhat esoteric in nature, such a discussion provides a theoretical basis for a consideration of the essential role of transmission representation and indexing. Namely, we postulate that any theory or practice of communication which causes a loss of data elements, either through their misrepresentation or by restricting their flow, must be considered inadequate.

Accordingly, the problem becomes one of representing messages that come from a number of unrelated sources. The initial collection of these messages forms what we have labeled the document space. It should be

obvious that message (document) representation must be effected so as to guarantee the maximum degree of overlap between the experience set that *is* the representation and the experience sets of the class of potential receivers (searchers). In this way, the "meaning" or intent of the document will be preserved.

The particular document representation that is employed must serve two distinct functions: 1) it must allow for the creation of "stores" of document content, and 2) it must provide a basis for search operations. This type of representational activity is implicit in Graziano's [17] view of the process of documentation (information storage and retrieval):

> ...the operational methods of identifying elements, distinguishing elements from each other and for transmitting sets of patterns from one time and/or place to another in such a way so as not to destroy the power of the symbols to convey exact concepts.

The IS&R representational activity, as described above, must be concerned not only with what the document *says* (*i.e.*, the message proper) but must be concerned with what the document is *about* (*i.e.*, content analysis). Since IS&R systems store data and retrieve information, it is the purpose of the system to effect (permit) the transformations between data and information. Obviously, the ability to effect these transformations depends on the fidelity of the representation.

Following Fairthorne [18] it is believed that the maximal representation of a document depends on the number of distinct configurations* that can be observed in it. Transformations are employed to reduce the

---

* This author believes that there exist a small set of structures (*e.g.*, patterns of data elements) in a language. Thus there exist a finite number of relations, so that the variability one observes in a language is only achieved through data-element substitution.

redundancy of the data-elements in the document. (This does not imply that the *code* used to represent the data content will be shorter--indeed, the shorter the code the less structural information is preserved.) These transformations involve an order-preserving representation of the document. Included are representation by *compression* (*i.e.*, the preparation of abstracts and extracts) and representation by *symbolic substitution* (*i.e.*, the creation of index entries through the use of thesauri, word control lists, etc.). The main observation, at this point, is that indexing performs a communication function. We postulate that the index provides the IES and the transformations required for effective communication. We must now consider the nature of indexing, the inherent drawbacks of present day real-world indexing, and finally, the types of relations and transformations required to create the IES.

Compare these two statements as descriptive of the nature of the indexing problem:

> No person who is engaged in the work of extracting information from printed sources... can fail to be aware of the frustration constantly presented by knowing that the information exists without knowing where it exists. [19]

> What constitutes a good index? The test is to determine whether or not an index will serve as a reliable means for the location, with a minimum of effort, of every bit of information (*sic*) in the source covered which, according to the indexing basis, that source contains. To meet this test an index must be accurate, complete, sufficiently precise in the information supplied, and so planned and arranged as to be convenient to use. [20]

Of course the ideal system would store the complete document, and each stored document would be searched in response to every request. Since this is a practical impossibility, a representation of the document is effected through indexing. Grems and Fisher [21] have provided an interest-

ing characterization and comparison of the nature of indexing and retrieval. The essence of their description is presented here in tabular form:

Indexing	Retrieval
objective	subjective
analysis	synthesis
impersonal	personal
algorithmic	heuristic

We will dwell at some length on the nature of these retrieval characteristics in Chapter 5. In any event, indexing is viewed as an algorithmic process for producing a document surrogate.

12.4 Indexing Failures

Indexes range in size from a few entries to entire sets of volumes. However, one should not make the false assumption that the quality of indexing is comensurate with the number of index entries chosen per document. Usually error is introduced in the creation of the index through limitations of the representational vocabulary. Under the constraint of a controlled vocabulary either the system totally lacks the power to describe the contents of a document, or, if means are available, they lack the required precision of description. In either case, most systems (either manual or automatic) force the user to supply alternate index entries. This occasions a lengthy index search for the satisfaction of an information need.*

Mellon [22] cautions:

> The searcher must guard against relying too heavily on the indexes. Too often they merely index titles or words, and at best they probably never contain entries for all of the important points covered by the articles.

---

* The concept of *information need* is defined and discussed in Chapter 5.

If the user adopts this negative view of the index, then he is left to his own devices to supplement or to supplant its contents. As Skolnik [23] reports, chemists frequently supplement available indexes by personal in-depth card files. In desperation, some researchers adopt the "random scan technique" of covering indexes and documents in an effort to find important items that have not been properly indexed (or else have been totally ignored). A case history which typifies the problem is given in Appendix A, page 136.

Apparently, in the current process of indexing, a document is viewed as a collection of a few "important" concepts. (The word important is placed in quotes because importance as determined by a system is likely to be considerably different from that determined by a user based on his experience set.) Once these "important" concepts have been identified, they are given labels and placed into an ordered list together with similar concepts from other documents. Ordering is based upon commonality of data elements with virtually no regard for relations shared by them. By example, a back-of-the-book index can be viewed as an alphabetically arranged collection of N ordered pairs of index terms and addresses. These entries correspond to large sections of text, causing potentially important information to be lost because of a lack of index terms which refer to specific data elements within the section. In addition, index entries rarely refer to all of the occurrences of a data element; rather they represent the (often implicit) imposition of a gross classification scheme on them.

Consider, for example, that there are 620 pages of text in Pauling's The Nature of the Chemical Bond [24] and 19 pages of index (both subject

and author), or stated differently, approximately 220,000 words of text and 2,100 index entries. The assumption that one can retrieve data from the text using the back-of-the-book index is a tenable as the assumption that only 2,100 of Pauling's words are of consequence, the remaining 218,000 words serving simply as filler. Yet, when such indexes are discussed and created, this is the assumption which is made in every case.

In order to show that the phenomenon exemplified by Pauling's book was general in nature the following experiment was performed. Eleven texts were selected at random from various fields. A chapter from each was then selected randomly, and the text types* contained in the chapter were identified. Those that appeared in the back-of-the-book index (index types; index entries which refered to the chapter in question) were then counted and the index-type/text-type ratio was calculated (see Table 12.1). These ratios cluster around 3 percent. Interestingly, the number of single entries (non-faceted) accounted for approximately 50 percent of the index entries associated with the chapters in question. The total index-size/book-size ratio was, on the average, 0.6 percent.

Geballe [25] in a recent review of *The McGraw-Hill Encyclopedia of Science and Technology* (containing 120,000 index entries and 15.8 entries/document) faulted the index for its treatment (or non-treatment) of synonyms and lack of uniformity in cross-indexing. He concludes [26]

> ...no editor used a wide-angle lens. The indexing appears to have been accomplished in a mechanical fashion; it suffers from a kind of aimlessness and inattention to overall considerations.

---

* A *text type* is defined as a word of the language.

TEXT SAMPLE	WORDS (TYPE) IN TEXT	TEXT TYPES IN THE INDEX	INDEX-TYPE/TEXT-TYPE	SINGLE ENTRY TOKENS	SIZE OF INDEX	SIZE OF BOOK	INDEX-SIZE/BOOK-SIZE
A	1389	66	0.047	50	826	140,000	0.006
B	1300	42	0.032	20	272	59,600	0.005
C	952	24	0.025	11	1398	144,000	0.009
D	1268	132	0.104	48	801	76,800	0.01
E	694	3	0.004	1	156	45,500	0.003
F	1342	25	0.018	12	194	68,400	0.003
G	1225	96	0.078	14	381	56,700	0.007

Table 12.1: A Study of Text and Index Tokens

Text Sample References

A) H. Borko, *Automated Language Processing*, John Wiley, 1969
Chapter 4

B) P.M. Fitts and M.I. Posner, *Human Performance*, Brooks/Cole, 1967
Chapter 3

C) P.L. Garvin, *Natural Language and the Computer*, McGraw-Hill, 1963
Chapter 11

D) J.R. Sharp, *Some Fundamentals of Information Storage and Retrieval*, London House and Maxwell
Chapter 4

E) D.A. Bell, *Intelligent Machines*, Blaisdel Scientific Paperback, 1964
Chapter 8

F) D. Lefkovitz, *File Structures for On-Line Systems*, Spartan Books, 1969
Chapter 4

G) S. Artandi, *An Introduction to Computers in Information Science*, Scarecrow Press, 1968
Chapter 3

Table 12.1(cont.): Study of Text and Index Tokens.

It is concluded that current indexing practices serve not only to eliminate many of the concepts in the document, but also to destroy many of the relationships between the concepts which are selected for the index.

12.5 Representational Relations

The number and type of documentary relations employed in IS&R activities tend to reflect our general lack of understanding of the functions of language. This means that data-element relations are employed only as aids in the representational and indexing process, and serve to contribute to the complexity of the many information retrieval languages rather than to facilitate a searcher's interaction with the IS&R system. The point is that the identification of data element relations allows for the specification of a document structure which reflects the homomorphic representational operations discussed in Section 12.3.

Two broad classes of relations can be identified: semantic and statistical. Statistical relations are characterized by data element type and token counts and frequency of occurrence values. We have characterized the semantic relations by five classes of relations (see Def. 3.7-3.12): equivalence, generic-specific, part-whole, difference and intensional. We might add to this list what Levéry [27] calls the relation of "nearness" or data-element proximity. Sometimes this relation takes the form of the identification of related terms (*e.g.*, concept clustering) and sometimes it takes the form of the identification of *contextual environment*. Unfortunately this relation is at best poorly defined and serves mainly as a symptom of the linguistic short-comings mentioned above. The relations of Definitions 3.7-3.12 are, to use DeSaussure's terminology [28], defined *in absentia* while the relation of "nearness" is defined (recognized)

in praesentia. Alternatively, we can call the former relations *paradigmatic* (the identification of patterns of data elements characterizes the $\mathcal{O}$ transformation - see Def. 3.13) and the latter *syntagmatic*. Whether paradigmatic or syntagmatic relations are employed in the indexing language is really a function of the current state of knowledge. The goal is total document content analysis, with respect to the other documents in $\mathcal{D}$, so that the relations may be characterized as Fairthorne describes them [29]:

> Parts of a document are not always about what the entire document is about, nor is a document usually about the sum of things it mentions. A document is a unit of discourse, and its component statements must be considered in the light of why this unit has been acquired or requested.

It should be clear that the specification of data-element relations is an order-defining transformation of D. Thus the order-defining transformation, $\mathcal{O}$, specifies which d ε D are mapped into REL. Hence,

Theorem 3.4 (Proof): The function, $\mathcal{O}$, creates equivalence classes of data elements with respect to the relations in REL. Thus, $\mathcal{O}$ partitions D.

similarly,

Theorem 3.5 (Proof): Documents in $\mathcal{D}$ are partitioned by data-element membership in the equivalence classes defined by REL.

13. A Further Specification of the Indexing System

Let us briefly review the material that has been presented in the previous two sections. In Section 11 we considered the nature of the concept of data element and touched upon its relation to Information Storage and Retrieval. An example document together with several forms of index entries were presented in Figure 11.1. The correlation between the document and the resulting index entries was modeled by means of the

indexing system. Figure 11.2 equated the indexing system to a black box that receives documents as inputs and produced indexes as outputs. In Section 12 we consider the nature of communication and information transfer. The concepts of data element, experience set and interface experience set were discussed and the task of the analysis and the representation of the transmission was delegated to the indexing system. We equated the indexing process to the interface experience set (see Figure 10.1) and then considered the nature of current indexing-communication failures. Finally, potential intradocument/data-element relations were discussed. Attention is now directed to the role and the position of the indexing system in the communication process.

In this section we shall differentiate between indexing, the indexing process and the indexing system. The object of the *indexing process*, as was implied in Section 12, is to provide a structure to represent the various orders of the data elements in the input documents. These data elements are usually accepted by the indexing process in the form of natural language strings. Consequently, for a given data element, the indexing process must represent the following items:

- the data element itself
- the surrounding data elements (context)
- the order of the surrounding data elements (syntax)
- relations (from REL) to other data elements (semantics)

The function called the indexing process is descriptive of the internal operation of the *indexing system*. Documents are input to the indexing system; this system controls the application of the indexing process which performs order-preserving transformations to represent the data elements

and relations between data elements found in the input documents; and, finally, an index is generated as output. To perform these functions, the indexing system must reside intermediary between the transmission channel and the receiver. This is illustrated in Figure 4.1 by means of an adaptation of the Shannon and Weaver communication schema.

A feedback function is included in this adaptation of the Shannon and Weaver model in order to depict the view of communication represented by Figure 3.1. The index, by means of a citation or accession number, enables the receiver to retrieve the source's document and thus complete the communication loop. Notice that the transmission channel, the indexing system and the feedback function are all potentially affected by noise. These errors represent, respectively, document transmission error (possibly encoding error), indexing process representation error, and receiver misinterpretation of the source document. Based on the observations of Section 12.4 it is postulated that the most significant error results from an alteration of data-element order by the indexing process. Thus, the error associated with current indexing practices serves to obscure the unique organization between data elements in documents.

The input to the indexing system is characterized as a document stream consisting of previously unrelated documents. The indexing system processes this stream in fixed intervals of time, called time slices. We assume that the time required to process a time slice is significantly less than the time required to process the entire document. In both manual and automated systems, the bibliographic citation, introduction, body, tables, figures, conclusion and references all supply different kinds of data elements and must, accordingly, be isolated and processed separately. Of course, a

given type (and value) of data element may appear in several locations in a document; consequently the indexing system must recognize common data elements and relations both within and between time slices. Hence,

Theorem 4.1 (Proof): In the document stream, document boundaries are just a type of data element, hence parts of more than one document may appear in a given time slice. Since the indexing system is able to recognize data elements and relations both within and between time slices, it can recognize inter- and intra-document data element relationships.

In a manual indexing system, an indexer (a component of the indexing system) is considered excellent (other things being equal) if he cuts across document boundaries when producing index entries. This is because the information he needs to make correct decisions about data-element values and relations is usually not contained in a single document. In order to cut across document boundaries (that is, to process all data elements and relations in a time slice), the indexer must make use of, among other things, the very index he is generating. It is for this reason that adequate (perhaps we should say intelligent) automated indexing systems have seldom (if ever) been developed.

The role of indexing (the indexing process and system) is to completely specify the data elements and relations in the document stream by means of order-preserving transformations. The document representation provided by the indexing process is a homomorphic reduction, or many-to-one mapping, from document stream to index. Hence,

Theorem 4.2 (Proof): The reduction transformations preserve the data element and relation order of the document stream, hence they are reversible. Document stream reconstruction is possible up to the specification of data-element order.

14. The Index as a Bi-Directional Interface

The indexing system provides the transformations and the interface experience set required for effective communication between the source(s) and the receiver. This means that data elements and relations between data elements must be identified and represented by means of order-preserving transformations. We have assumed, from our theoretical view of the indexing process, that such transformations completely specify the *content* of the documents in the document space, $\mathcal{D}$. The index is the end product of all of this activity; namely, the index is the image of composite order-preserving mappings performed on $\mathcal{D}$. The crucial point to realize is that not only is the index the product, but that it is all that remains of the original document space. We assume that the original documents are not directly available to the receiver, hence the index is the receiver's only point of access to the document collection. Under such constraints it should be clear that accurate retrieval depends on the exactness of the indexing. In other words, the indexing system must produce an index that is a facsimile of the document space. Hence,

Theorem 5.1 (Proof): Inaccurate or incomplete document space representation will lead to retrieval error since the index is the receiver's only point of access to the document collection. The index, and the indexing system are the only intermediaries between $\mathcal{D}$ and the receiver, hence reliability of document representation is the function of the indexing system.*

Reliability is partially achieved through the completeness of the index entry. Bernier [30] makes this point clear:

---

* For further amplification of this point, see Appendix A page 136.

> There is not so much information in an index entry or vocabulary terms as in the document or parts of a document that it represents. Because of the greater context and meaning of an index entry heading and modification (modifying phrase) than of a term or word, the complete index entry serves more effectively as a guide to the information than does a single word or term.

However, the indexing system cannot assume that all (or any) statements in a document contain information--indeed, a document is just an author-assembled collection of data elements. We infer that the data elements of the document become *information* when they are assimilated or put to use by the receiver(s). Consequently, information is defined as "data elements of value in decision making" (adapted from Yovits and Ernst [31]). The index, and the subsequently retrieved documents, must provide data elements at the proper time and in the proper form to be of value in the decision-making process.

Prior to a consideration of the indexing system transformations and the index space, an overview of the concept of information is presented. This discussion is not only applicable to this section but is also preparatory to the topics to be presented in Sections 15, 16 and 18.

14.1 Information

We shall consider two approaches to the definition of information which was presented above and in Section 5 of this Chapter. First, information is defined from an organizational/operational viewpoint and, second, information is defined as an extension of the concepts of turing machine and procedure outlined in Section 11.

Briefly, we derive information from the world about us by performing a set of operations on an object under study. The result of these operations is a selection of a subset from the set of alternatives that

was available prior to the application of the operations. The operations are the experiment and the subset of alternatives is the resulting measurement. This is an informational description of the operational processes of science. Information is obtained through the reduction of the number of alternatives available to describe the object under study (the number of alternatives available before the measurement is the *precision* of the measurement). As expressed by Brillouin [32], information is the logarithm of the ratio of the *a posteriori* number of alternative values, A_a to the *a priori* number of alternative values, A_b:

$$I = \log \frac{A_a}{A_b}$$

However faithful this measure is to the statistical-mechanical conceptualization of information, it tells us nothing about the quality or usefulness of the derived information. In the real world there is, for one thing, a non-equality between alternatives; thus, a better way of evaluating experimental results is desired.

There are at least four different classes of information. They include: 1) technical or communication-theoretic information (Shannon [33]); 2) semantic information (Carnap and Bar-Hillel [34]); 3) pragmatic or effectiveness information (Yovits and Ernst [35]) and; 4) inferential or experimental information (including, Shannon [36] as an informational measure of the mean, Fisher [37] as a measure of the variance, Kullback [38] as an informational measure of the confidence in alternate hypotheses about the value of the mean). We shall direct our attention to a measure of information that incorporates the concept of the use and the effectiveness of information. To introduce the measure, the concepts of *course of*

action, decision making and *decision* are briefly discussed.

Intuitively, a *course of action* can be interpreted as a planned sequence of responses to an anticipated set of stimuli. Thus, a course of action can be defined as a well-ordered set of stimulus-response pairs that are directed toward the attainment of a goal. A course of action is specified by the enumeration of the following: the set of inputs that it can process, the set of states associated with the processor, and the next state and output functions. Of course this is the definition of a turing machine and, consequently, a course of action can be equated to a procedure. It is possible that alternative well-ordered sets of responses may exist for the achievement of the same goal. Thus, under the constraint that only one course of action may be effected during a prescribed interval of time, a choice must be made between the alternatives. This choice must take into account both the present state of the system (the system is that which executes the course of action) and the present inputs (*e.g.*, course of action). The execution of the choice may involve several sequential inputs and several intermediary outputs, consequently next state and next output descriptions must be provided. A definition of choice then, amounts to a definition of a turing machine. We shall call the process of choosing between alternative courses of action *decision making*.

Based upon the above characterization, the final output from a decision-making procedure is the selection of a course of action for subsequent execution. This final output is called a *decision*. The decision which is output is described by the relation, σ: $K \times \Gamma \rightarrow \Gamma_A$, where Γ_A denotes the set of alternative courses of action. We shall demand that the state transformations associated with this choice result in the

attainment of a final state, hence the decision-making procedure will halt.* Finally, the input symbols to the decision-making procedure which lead to a final state and to the choice of a course of action are defined as *information*. Since the input symbols are data to the operation of the turing machine *i.e.*, to the decision-making procedure, information is also defined as data of value in decision making. It should now be evident that information is context sensitive** since those data leading to a final state depend on the starting state and the sequence of inputs to the decision-making procedure.

The connection between data, information, course of action and decision making is conveniently modeled in the Yovits and Ernst [39] description of the *information transfer process* (see Figure 14.1). Notice that the observables that result from the execution of the course of action eventually become new data for the information (really data) acquisition function in the model. It is believed that this interpretation of the information transfer process embodies the desirable measures of information use and effectiveness.

Finally, it should be noted that the indexing system is really a model for the information acquisition box in Figure 14.1. Data must be carefully identified and represented so that the particular decision-making context

---

* Giving us, therefore, an *algorithm* for effectively making decisions.

** A.D. deGroot [41] has shown that after a short stimulus period, a chess-master can easily reconstruct the chess board arrangement shown to him, whereas a novice finds the task almost impossible. It is hypothesized that the Master stores the information about the board in the form of *relations* between the pieces, rather than in the form of a complete scan. The relational context creates a non-equality among the probabilities of the alternative arrangements, thus, there is no "information overload."

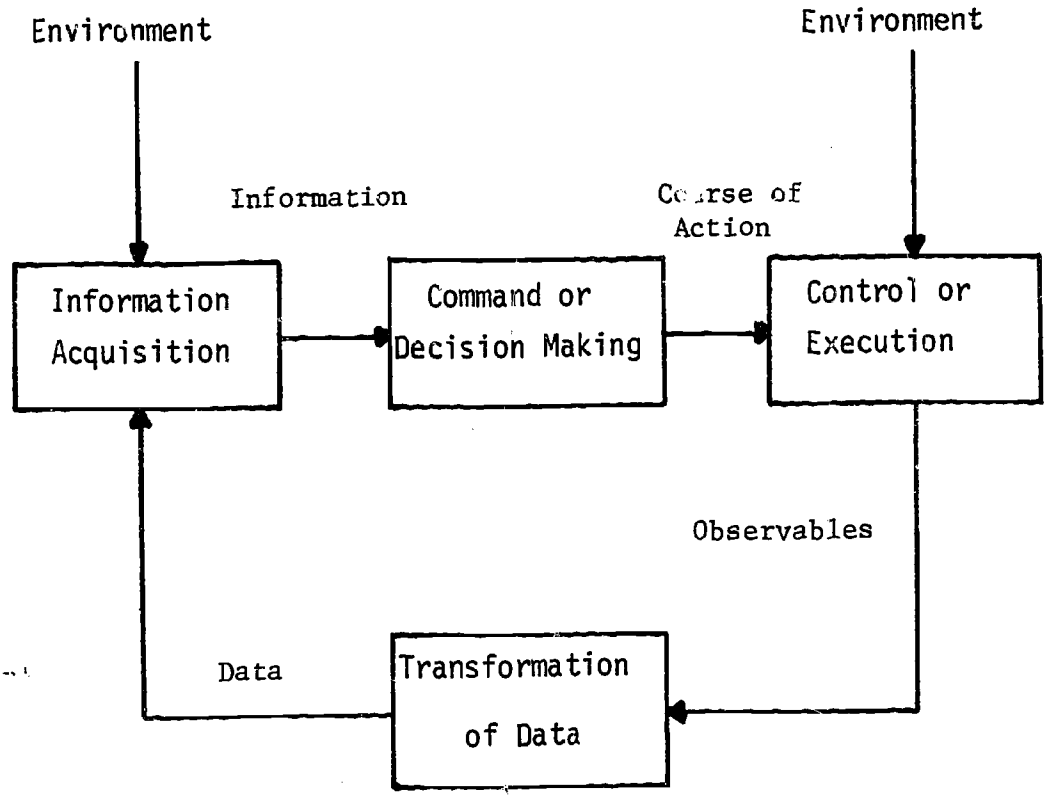


Figure 14.1: The Yovits/Ernst Model of Information Transfer. (from [31])

will define the required information. Fairthorne [40] has similarly observed that mathematical statements are just data--"the *information* that one reads into a numerical result often involves the semantic field of a particular application."

14.2 Indexing System Transformations

The indexing system effects two distinct types of transformations upon the input documents. First, data elements and relations present in the document must be *expressed* in terms of the system's data elements and relations. This transformation is effected through use of the analysis documents, $\mathcal{D}_a$. Secondly, the system effects a transformation (denoted here by the letter g) on the data element representation to create the index entry. Variations in g yield different types of index entries. Thus, the *form* of the index is specified by this second transformation. Let us consider each of these transformations in turn.

The document space, in Section 5, was described as the union of two subspaces: the input documents and the analysis documents. The input documents constitute the document stream; $\mathcal{D}_i$ is continually changing. However, $\mathcal{D}_a$ is created by the system (or, for the system) and is assumed to change at a rate which is much less than the rate of flow of documents in the document stream. Analysis documents take the form of classification hierarchies, word guides*, vocabularies, lists of formulae, syntactic classes, *etc.* In other words, the analysis documents are the embodiment of the system's representational rules, and amount to the system's realization of the set of relations, REL. For illustration, consider the

---

* A word guide is the only reasonable extension of the concept of *thesaurus*.

sample document presented in Figure 11.1. The title "Effect of a selective beta-adrenergic blocker in preventing falls in arterial oxygen tension following isoprenaline in asthmatic subjects" is input to the indexing system (let's say that it appears in the first time slice) and a representation of the title is effected by the indexing system. Figure 14.2 shows the data element transformations that have been effected by documents in $\mathcal{D}_a$. In this example $\mathcal{D}_a$ consists of role indicators (T), generic-specific relations (G), formula list (E) and a word guide or controlled vocabulary (T). The indexing system representation of the title might take the form:

> "R_{19} of a selective beta-adrenergic receptor (beta receptor) blocking drug (drug) in R_6 R_3 in arterial (cardiovascular system) oxygentation tension (airway resistance) R_{30} isoprenaline ($C_{14}H_{22}N_2O_3$) in asthma subjects."

It is obvious that this representation is a composite of $\mathcal{D}_i$ and $\mathcal{D}_a$ and since the data-element/relation/data-element triplets of $\mathcal{D}_a$ incorporate relations from REL, the representation can be expressed as $(\mathcal{D}_i) \cdot \text{REL}$ (Theorem 5.3).

Once the input document representation is effected (*i.e.*, after several input-time-slice operations), then the indexing system applies the index-entry generation function, g. Following the model shown in Figure 11.1, the particular form of the index entry depends upon which data elements and relations are selected from the representation (see the index entry examples in Figure 11.1). Hence,

Theorem 5.4 (Proof): The index entry function, by definition, effects a transformation on the representation $(\mathcal{D}_a \otimes \mathcal{D}_i)$ and this transformation is an order and relation preserving (homomorphic) mapping. This transformation is represented by $I = g(\mathcal{D}_a \otimes \mathcal{D}_i)$.

Text:

Effect of a selective beta-adrenergic blocker in preventing falls in arterial oxygen tension following isoprenaline in asthmatic subjects.

(Input)	(System Representation)	
Data Element	Data Element	$\mathcal{D}_a$
Effect	R_{19}	Role
beta-adrenergic	beta-receptor	BT
adrenergic	adrenergic-receptor	USE
blocker	blocking-drug	USE
	drug	BT
prevention	R_6	Role
falls	reduction	USE
reduction	R_3	Role
arterial	cardiovascular system	BT
arterial oxygen	arterial oxygentation	USE
arterial oxygen tension	airway resistance	RT
following	after	USE
after	R_{30}	Role
isoprenaline	$C_{14}H_{22}N_2O_3$	FORMULA
asthmatic	asthma	USE

Figure 14.2: Data Element Transformations Effected by the Application of $\mathcal{D}_a$.

Once the individual index entries (Def. 5.4) have been generated, then the physical form of the index depends only on the particular manner of index-entry ordering, source-document citation, repeated heading selection, *etc.* Also a new index is just an update of the old version, hence:

Theorem 5.5 (Proof): New index entries are created by the application of the indexing process to new documents in $\mathcal{D}_i$. Assuming that the index entry function, g, has not changed, then the new index is simply the union of the old index and the new entries.

14.3 The Index Space and Retrieval

The end product of the indexing process is the creation of the index entry and, finally, the index. Lancaster [42] provides a description of the "ideal" index entry vocabulary:

> Ideally, an entry vocabulary should contain all words and phrases used in input documents to express items of subject matter that have been recognized in the conceptual analysis stage of indexing. The entry vocabulary will refer to the code terms used to express this subject matter.

However, our discussion of the indexing system has given no clue as to whether any specific operating system can provide such an entry vocabulary. Clearly, a means of characterizing the operating level of the indexing system is required. We shall describe the operation of an indexing system by means of the *index space*, $\mathcal{I}$.*

In Definition 2.11, the index space was initially described as a representation of the data elements and relations found in the indexing

---

* Maron [43] has used the term *index space* in reference to an n-dimensional space of vocabulary terms, where connections represent sharred relationships. This is analogous to the document/term space described in Chapter 3. We shall, rather, restrict the concept of an index space to a 3-space.

system. The g-transformation and the $\mathcal{D}_a$-based transformations that were discussed in Section 14.2 can, obviously, accomodate a wide range of data element descriptions and relations. In fact, the variability of the vocabulary and of expressions that characterize the spoken and written languages applies equally to the operation of the indexing system. Consequently, we postulate that the indexing system is best characterized by the vocabulary, productions and expressions that it can recognize and subsequently represent. Accordingly, the index space is defined as a triple formed by the cross product of vocabulary, transmission decoding and language.

The vocabulary, V, is a finite set of possible data elements each of which defines an equivalence class of symbols. These data elements can be ordered by precision of measurement. Examples of the resulting data element continuum are found in Bernier's classes of "microsemantics" and "macrosemantics": punctuation, symbols, suffixes, words, phrases, clauses, sentences, paragraphs, pages, chapters, sections, reports, books, collections... The actual vocabulary elements will be specific characters, words, suffixes, *etc.*, frequently given as a document in $\mathcal{D}_a$. Subsets $V_i \subseteq V$ of the vocabulary continuum (not necessarily continguous subsets) describe those data elements that can be recognized by the system.

Transmission decoding, TD, is a set of productions (rules) which define "recognizable" strings of data elements taken from the subsets, V_i. For example, <word><word> or <word><formula> would be considered as acceptable input or output strings, however, <formula><formula> might be labeled as unacceptable. Subsets $TD_j \subseteq TD$ of this continuum of possible data element productions describe the productions actually employed by a given indexing

system. These productions are especially useful for the characterization of the permitted data-element syntax in an index entry.

Language, L, is a set of possible index entry expressions. These expressions are built from strings over V, defined by TD_j, and from relations from REL. An example of the continuum of expressions is offered by Meadow's [44] continuum of indexing languages: hierarchical, subject heading, keyword, tagged descriptor, faceted term, phrases, natural language. Each of these languages defines a different form of index entry, especially when an index entry is viewed as an ordered set of data elements and relations (see Def. 5.4).

As we shall see, the concept of the index space provides a useful framework for analysing the retrieval process. Recall that the concept of query, Q, was defined (Def. 2.13) as a well-ordered set of data elements, such that $Q \subset I$. From the previous discussion it should be clear that either this is an idealized statement or else Q represents the query as finally accepted by the system.* Experience tells us that the latter is the case. A receiver's initial query will not immediately be acceptable to the retrieval system--indeed, the problem amounts to one of matching the user's "conceptual" terms with the system's fixed scheme of document representation, that is, of putting the query into a form acceptable to the system.

A query may deal with either specific data elements or complex combinations of data elements and relations. Typically a query takes the

---

* We assume that the indexing system provides for both the representation and the retrieval of documents--thus it can be called the "retrieval" system.

form "What are the physical and chemical properties of compound X?" or, "How does one convert compound X into compound Y?" The initial goal of any system/query interaction is to bring the two vocabularies into coincidence, which means that common data elements and relations must be discovered (the process of discovery in retrieval is the main topic of Chapter 5). The user will not necessarily know that the index lists Pentylbenzene under "Benzene, pentyl-" or that Hexylbenzenes are listed under "Hexane, phenyl-" [45]; consequently, several interactions with the index are required before user and system achieve coincidence of expression. To the designers of the indexing system the disposition of the index space (*e.g.*, which V_i, TD_j and L are implemented) is clear; however, to the user, the exact nature of the index (his conceptualization of the index space) appears to be fuzzy. This situation accounts for the several interactions required to bring the user's query expression into a form compatible with the index and the indexing system. Hence,

Theorem 5.6 (Proof): Compatibility between query and indexing system means that the query and index share the same vocabulary, productions and expressions, or, $Q \subset \mathcal{J}$.

Theorem 5.6 is interpreted to mean that the data-element ordering and relations present in the query must also be present in the index. Consequently, retrieval is viewed as a homomorphic mapping from the request into the index space. Hence,

Theorem 5.7 (Proof): From Theorem 5.6 and Definition 5.9, we know that the relations present in the query must be the same as th[illegible]e in the index and those which are de[illegible]d by the index space. The indexing process and its reversibility (Thms. 4.1 and 4.2) account for the mapping $I \leftrightarrow \mathcal{D}$.

15. The Indexing System as a Phase Space

In the previous discussion we have assumed that the indexing system is always present and functioning, but for the sake of contrast, consider the extreme case of an IS&R system without an indexing subsystem. In such a system, input documents would simply be stored by their order of arrival at the system. A user of such a system would be forced to conduct an exhaustive, sequential scan of the entire collection in response to his every "information need." Such a system would either have a small (drawer sized) collection of documents, or a fully automated time independent processor or, more likely, a vanishingly small group of users. This hypothetical situation represents the case where data elements are to be located in a collection about which there is no prior knowledge concerning its contents. We have previously postulated that effective IS&R system operation presupposes some manner of organizational scheme for document representation. Fairthorne [46] reminds us that the needed organizational scheme is not simply a communication-engineering problem:

> The communication engineer is not concerned with completed messages, but how to deal with bits of them in the course of communication. The IS&R specialist [rather] deals with spatial collections of completed messages and, after recognition and identification, questions of their ordering and disordering predominate.

The IS&R system must effect some form of organization of the input documents so as to maintain "coverage" and to provide a manageable search time. We postulate that this organization of the document space is provided by the indexing system by means of its recognition and representation of inter- and intra-document data-element relations. Consequently, the probability of a given set of data elements becoming information (recall the discussion in Section 14.1) is a function of the work expended by the indexing system.

Symbolically, the potential utility of data elements (with respect to their information content) after indexing, PU_i, is the sum of the potential utility before indexing, PU_{i-1}, and the work expended by the indexing process, W:

$$PU_i = PU_{i-1} + W$$

We shall now consider how the representational operations of the indexing system can be modeled by thermodynamic concepts and, how such considerations introduce the concept of "information benefit." First, a brief overview of thermodynamics is presented.*

Thermodynamics is concerned with the energy description of well defined systems. More specifically, thermodynamics is the study of the relationship between heat and work. In the characterization of the indexing system we shall be concerned with either open systems (systems that exchange heat and matter with their environment) or adiabatic systems (no exchange with the environment). Thermodynamic parameters include the following: entropy, mass, energy, volume, temperature, and pressure. The specification of a value for the parameters denotes the state of the system. What is important to this study is that the parametric structure is assumed and the alternative values (states) are unknown before measurement.

Thermodynamic systems are conveniently modeled by statistical mechanics. Statistical mechanics accounts for thermodynamic properties (microscopic or macroscopic) by considering a system as a collection of particles (*i.e.*, gas molecules) subject to the laws of motion. Measurements on thermodynamic systems are postulated to be performed on a *phase space* composed of 2n dimensions (n positional coordinates and n momentum coordinates). The

---

* An excellent discussion of the relationship between energy and information has very recently appeared [47] to which the reader is referred for a more detailed treatment of this subject.

specification of the state of a particle, or of the state of the system, is analogous to the identification of a point in this phase space. Caratheodory's principle [48] posits that certain adiabatic state transformations are impossible, hence there is a natural partitioning of the phase space. The resultant partitions are identified as equivalence classes of states determined by adiabatic transformations. The macroscopic property, entropy, is assumed to be constant for each equivalence class. Consequently, entropy measures the amount of missing microscopic information (*e.g.*, which state is occupied) given the energy of the system [49]. The important point is that while the structure of the phase space is known, *a priori*, entropy is a measure of the uncertainty of the state value.

The indexing system, with respect to the document space, must be treated as an open system; however, the indexing process is assumed to be effected within an adiabatic system. The phase space associated with the indexing system is a space of n dimensions corresponding to the n data elements recognized by the system. The "configurational" coordinates are those data elements which characterize documents in $\mathcal{D}$. The "momentum" coordinates, as we will see later, correspond to the concept of the *index search*. The analogue of Caratheodory's principle is the equivalence of data elements as manifested through shared relationships (from REL) between data elements. In a real sense, the indexing phase space corresponds to the range of index entry assignments permitted within the indexing system, hence,

Theorem 6.1 (Proof): A point in phase space represents a data element type, and the partitioning of the phase space amounts to a specification of the allowed data-element expressions (state transformations). Every point of phase space has its analogue in the system's index space.

In a formal sense, both indexing and measurement involve the production of a result from a classificatory act on an object of interest. In IS&R the object of interest is a document. Indexing, much like measurement, serves to reduce the uncertainity concerning which data elements are present in the input document. Clearly, careful observation (measurement) is required to narrow the *a priori* alternatives for classification. The result is the ability of the system to fit a given data element into the index. However, since most indexing systems are imperfect, the associated measurement operation must involve some uncertainty. This uncertainty corresponds to the indexing system's inability to exactly specify the correct point in phase space. This indexing "noise" or "error" is conveniently accounted for by the Heisenberg uncertainty principle [50].

Despite the existence of indexing "error", the indexing system as a phase space effects a considerable reduction in the entropy of the document-space/searcher interface (Thm. 6.2). Prior to indexing, the searcher's knowledge of the contents of the document space is minimal--hence, his uncertainty is maximal. The indexing process identifies those elements of the phase space which are present in the document space, hence uncertainty, through the use of the index as the interface, is reduced. However, since the indexing system is adiabatically closed, such a reduction of entropy must be matched by a commensurate rise in entropy elsewhere in the system (Thm. 6.3). This rise in entropy is accounted for by the effort (mental, physical) required to effect the indexing process. Thus, the change in entropy is equivalent to the work, W, expended in the indexing process.

The "momentum" coordinates of the indexing phase space are modeled by Rothstein [51] as the path of a search. If one adds the dimension of time

to the phase space, then a sequence of points represents a "search" in phase space. Since we have equated the points of phase space to the entries of the system's index, then a path in phase space also represents a search through the index. Rothstein posits the existence of an *average scan* or search that is required to retrieve the desired information from the system. The existence of an average search length (>1) is a direct result of the error or uncertainty that characterizes the structure of the phase space. We add the concept of the *ideal search* which results in the retrieval of information on the first access to the index, and the concept of the *optimal search strategy* which results in the shortest path to retrieval--short of the ideal search. It is to be expected that the retrieved data elements resulting from such forms of search have varying informational values or benefits associated with them. Such considerations will be discussed in Sections 16 and 19.

16. Course of Action as Hypothesis Testing and Decision Making

The previous sections of this chapter have contained discussions concerning how the indexing system represents the elements of the document space. Attention has been directed to both the manner and the form of this representation. We shall, for the sake of further discussion, assume that the indexing system has performed its function (the quality of the performance is another matter) so that we may now consider the nature of the process of conversion of stored data into information. We will only briefly discuss the concepts of goal, hypothesis testing and decision making, since a detailed consideration of their nature and role in information retrieval forms the substance of Chapter 5.

When a receiver (user) attempts to access the stored data, by means of the index, it is assumed that he has a goal in mind. We assume that, to the user, a goal represents a desired end product or end state. Several examples of retrieval goals can be given in the form of questions:

Are there data on compounds X and Y?

How does one convert X into Y?

What is the mechanism of the reaction of X with Y?

What is the effect of catalyst A on the reaction of X with Y?

What other (than X) compounds yield Y under the influence of A?

One can easily conceive of these goals as representing separate but conceptually related sequences of interaction with the retrieval system (*cf.* footnote on p. 107). But for each goal there is a corresponding course of action which is executed as a repeated interaction with the index and subsequent analysis of retrieved data elements. Furthermore, each interaction involves a hypothesis concerning the contents of the document space. Thus, each course of action is a sequence both of hypotheses concerning the contents of the data store *and* of decisions concerning whether or not the goal has been obtained. Recalling the discussion of Section 14.1, it is recognized that the retrieved data *may* provide information with respect to goal attainment. Thus, an initial hypothesis may be either refuted by the retrieved data or else it may be incompletely supported; in either event, a new hypothesis must be formulated and new data examined.

The progression between hypotheses, decision making and goal achievement, which was depicted in Figure 7.1, gives rise to two cases of data-element benefit, with respect to goal attainment. The sequence hypothesis-formulation/decision/goal is said to provide *maximum benefit* since the data

retrieved (information), in response to the initial hypothesis, completely "satisfy"* the goal. Conversely, the *minimum benefit* case is identified by the hypothesis-to-hypothesis path. In such a case, the data that are retrieved are not sufficient to provide information concerning goal achievement. As was previously implied, benefit intermediate between the maximum and minimum benefit cases is obtained when the information is necessary but not sufficient to reach the goal, and the formulation of a new hypothesis, based on the nature of the data already obtained, is required. Thus, when the retrieved data are not wholly suited to the testing of the initial hypothesis, a new hypothesis must be formed. Although a decision must be made to formulate this new hypothesis, we will call this a *meta-decision* since it is not directly involved in the final attainment of the goal. Furthermore, although information is obtained from the failure of an hypothesis, we shall refer to such information as *meta-information* since it is associated with a meta-decision. Data elements are of value in decision making (hence, are information) when they are directly involved with goal achievement. Hence, meta-information is not equivalent to information (Thm. 7.1).

It is interesting to note that the types of data-element search that are carried out in phase space (see Section 15) can be conveniently represented as the progression between hypotheses. The ideal search corresponds to retrieval yielding maximum benefit (the H-D-G path), whereas average search is represented by a chain of hypotheses terminating with goal attainment (H-D-H-...-D-H-D-G). The optimum search strategy represents the user's

---

* The concept of the "satisfaction" of an information need will be discussed in Chapter 5.

systematic variation of index-entry attributes in an effort to retrieve the desired information. The nature of such a strategy will be discussed in Chapter 5.

It should be clear from the previous discussion that benefit is a relationship between the information obtained by the receiver and the number of decisions (meta- and real) required to satisfy an "information need". Intuitively, the information that is retrieved through the first query and that satisfies the goal has maximal benefit. However, we infrequently experience the maximal-benefit situation--rather, benefit must be accumulated over a sequence of queries.

17. Perfect and Imperfect Indexing Systems

> By a fiction as remarkable as any to be found in law, what has once been published (no matter what the language) is usually spoken of as known, and it is often forgotten that the rediscovery in the library may be a more difficult and uncertain process than the first discovery in the laboratory.
>
> Lord Rayleigh

This dim view of a searcher's likelihood of success in library search is further supported by Reid's [52] comments: "... a point will always be reached, eventually, where all competent judges must agree that the probability of finding a reference and its possible value if or when found do not warrent the time, trouble or expense involved in continuing [searching]." It is emphasized that, perhaps, the principal postulate of the theory of indexing propounded in this dissertation is that error in information storage and retrieval stems from error in indexing. The indexing process, as usually implemented, does not accurately mirror the contents of documents in $\mathscr{D}$. As a consequence of this failure, a document indexed, for example, by the term "glass" may actually discuss a principle governing the action of metals or of undercooled melts [53]. Aside from search by "browsing", these

other "content descriptors" are forever lost. We need not prolong the examples of indexing failure (see Section 12.4 for a discussion of the failures of the back-of-the-book index); rather, let us contrast the concepts of "perfect" and "imperfect" indexing systems and attempt to draw some conclusions concerning areas for the improvement of current indexing processes.

17.1 The Theoretical and Real-World Indexes

The "perfect" indexing system operates according to the principles embodied in the indexing theory, hence, we shall call the output from this system the *theoretical index*. We define the theoretical index as serving to represent all inter- and intra-document relations between data elements in the document space. It is assumed that order-preserving operations and transformations are employed at all steps of the "perfect" indexing process. Recalling Mellon's definition of the good index presented in Section 12.3: "...an index will serve as a reliable means for the location, with a minimum of effort, of every bit of information [data] in the source covered..," it is concluded that every data element occurrence must be indexed so that the contents of the document will be available to every potential user and query. This is the essential role of the theoretical index.

If we assume that there are potentially d data elements in a given document, then each data element serves as a two-valued function--either the document has the datum or it does not. Consequently we could define 2^d subsets of data elements by the operation of set intersection. The theoretical index must provide for the existence of at most 2^d connections (shared relationships) between data elements. However, since there are multiple documents (say m of them) in the document space, $\mathcal{D}$, then one must allow for the existence of an increased number of data-element relationships. One bound

(limit) on the number of data-element relations is $2^{d_1} \cdots 2^{d_m}$ and, by definition, the theoretical index must be able to represent any subset of $2^{(d_1+\ldots+d_m)}$ (Theorem 8.1). This large number of relations is calculated under the assumption that the data elements associated with the m documents are unique. A more manageable upper bound for the number of potential relations that could appear in the theorectial index would be $2^{(d_1 \cap d_2 \ldots \cap d_m)}$. Since a query, Q, is mapped into the index, I, we can define a mapping of requests into the subsets of possible data-element relations. This mapping is between data elements of the query and entries of the index:

$$Q \rightarrow 2^{\bigcap_{i=1}^{m} d_i}$$

In contrast with the "perfect" indexing system, the "imperfect" indexing system is characterized by its output--the *real-world index*. The essential difference, as we have previously mentioned, is that real-world indexes contain significantly fewer index entries than would have been represented in the theoretical index. Consequently, there is a loss both of important data elements and of significant relationships between data elements. We postualte that for a given document space, real-world indexes fall short of the theoretical index because the indexing of the document space is incomplete.

An example of the incompleteness of "imperfect" indexing systems can be found in a comparison of the theoretical growth rate of four well-known indexing methods with their operational counterparts, the growth rates of which are severely restricted by means of word control lists and simple index-size

limitations. Figure 17.1 shows the theoretical (solid lines) versus the real-world (dotted lines) growth rates of an articulated index [54], the SLIC index [55], and the uniterm or keyword index. The theoretical growth rates are as follows: articulared--odd members of the Fibonacci series; SLIC--$2^{(n-1)}$; uniterm--n, where n is the number of data elements/document. Clearly, real-world indexes do not provide a sufficient number of index entries.* Figure 17.2 shows the performance of these indexing methods with respect to the hypothesized 2^n number of relational entries. Values above the equality-of-number-of-the-index-entries-to-2^n line represent redundant entries, whereas, values below the line indicate poor performance. Interestingly, for large numbers of terms, the simple combinations of terms show the best performance. However, it can be argued that the SLIC index performs just as well since all combinations of terms can be easily generated from this index. One can also argue that a consistent deletion of redundant entries is desirable.

In general, then, published indexes fall short of the theoretical index. Reasons for this phenomenon could be the lack of adequate technology for large index storage (we will briefly discuss this point in Chapter 6) or the prohibitive cost of the generation of a large number of index entries. It is realized that the theoretical index and the perfect indexing system are unobtainable**, however, there is a positive value in knowing what the ideal

---

* For n > 6 we have the following ordering by decreasing number of entries: permutation (n!), articulated, combinations, SLIC, double KWIC (n(n-1)) [56], uniterm.

** The third law of thermodynamics tells us that the entropy of a pure quantum state is 0; or, that complete certainty about the document space is impossible.

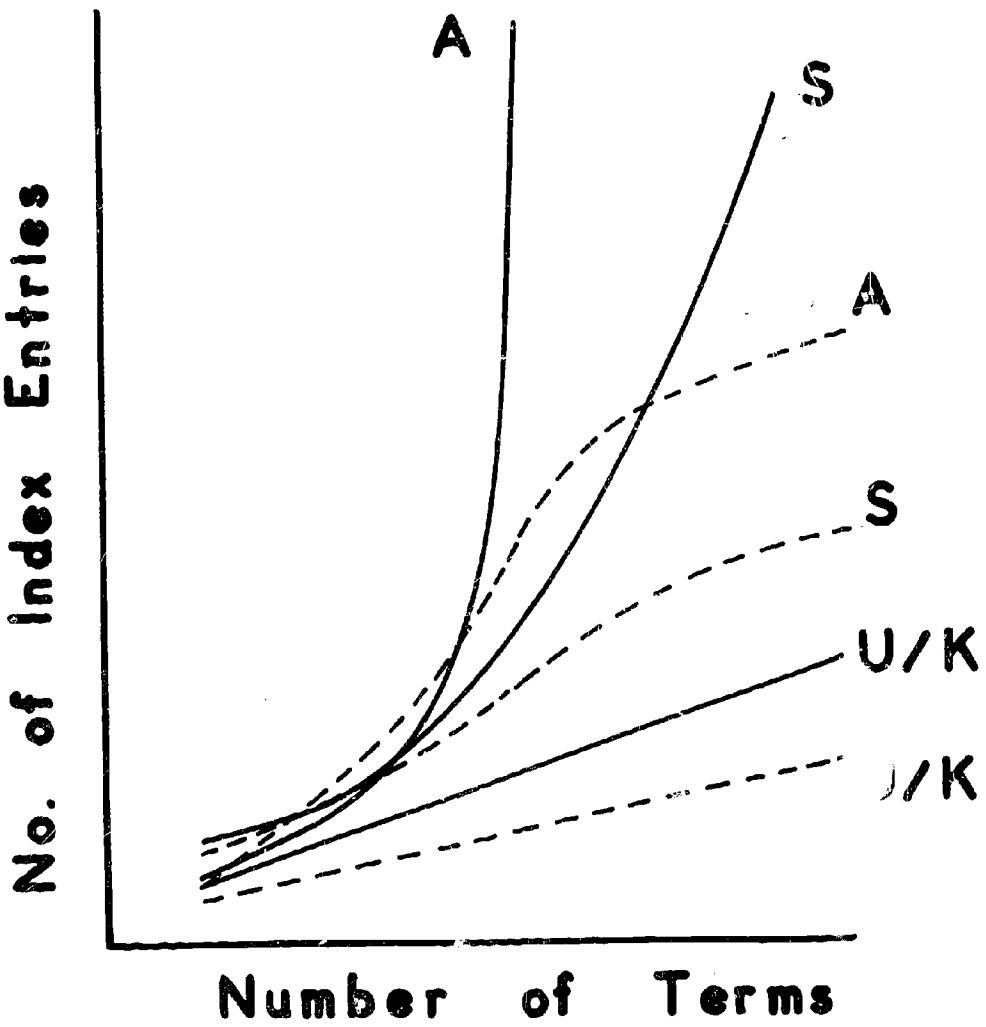


Figure 17.1: Theoretical vs. Real-World Index Growth (A = Articulated index; S = SLIC index; U/K = Uniterm or KWIC index; ... = real-world, ___ = theoretical)

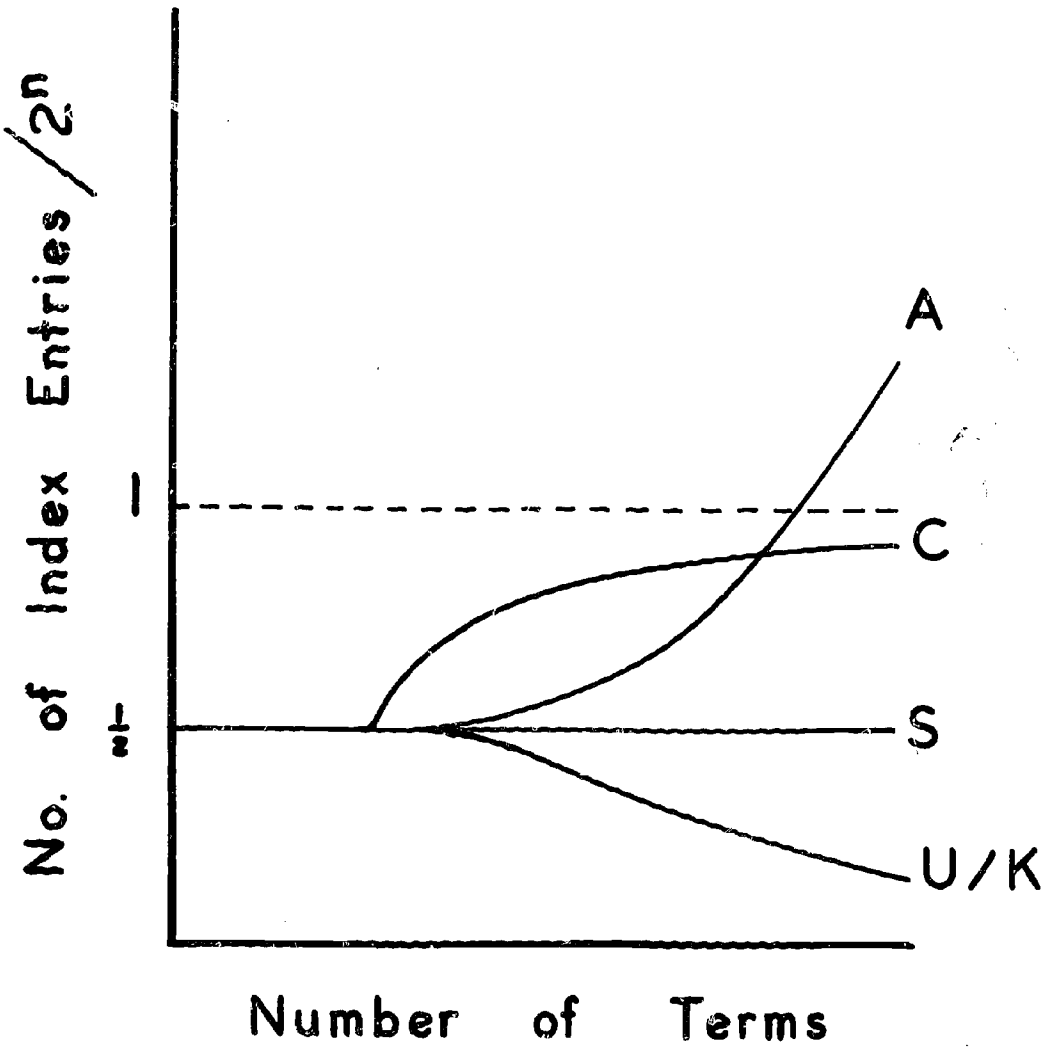


Figure 17.2: Relationship Between Real-World Indexes and the Theoretical Index.
(A = Articulated; C = Combinations; S = SLIC index; U/K = Uniterm or KWIC index; ... = real-world, ___ = theoretical)

is--if only for purposes of evaluation of operational systems.

17.2 Possible Real-World Index Improvements

Index size is a sensitive subject. By example, the size of a back-of-the-book index cannot grow, for reasons of economy, to a size equal to or larger than the size of the book! However, any increase over present day sizes would be beneficial in terms of efficient and accurate information retrieval. The only practical way of increasing the accuracy of a book index is to increase the number of index entries, and certainly to increase the number of multi-term entries. This would go a long way toward increasing both the number of relevant data elements and data-element relations. Unfortunately, most other forms of indexes suffer from a lack of "depth of indexing" [57] and would therefore benefit from an increase in the number of entries. By example, adding subject, classification or text enrichment terms to a document title will, in many cases, vastly increase its utility in a retrieval data base--especially when the title is used in a KWIC index.* The main problem is that depth of indexing is *not* solely associated with the number of index entries (*i.e.*, the number of keywords), but relies on the exactness of the specification of data-element relations. Accurate retrieval depends on the commonalty of data elements *and* relations. We shall consider the representation of data elements and relations by a finite state graph (really a model of the language component of $\mathcal{J}$).

---

* One unexplored possibility is the use of KWIC indexing to represent bibliographic citations. Indexes could be prepared not only for authors, but also for title terms, sources and dates. This would eliminate the tedious scanning of lengthly reference lists.

Figure 17.3 shows a very simple example of two data elements associated by means of three relations: A, B, C. By construction, all three relations fall into the same equivalence class (since they all link the same two data elements), however, each relation will serve to identify a different set of documents. Thus, the product of all relations between data element 1 and data element 2 will yield all of the documents in some way related to both data elements.

In this model, new relational equivalence classes can be easily defined, for example, by specifying a directional nature to the edges of the graph (*e.g.*, the "paradigmatic trees" in SYNTOL [58]). It is noteworthy that the larger the number of simultaneous relations in the query (*i.e.*, the larger the query set) the smaller the number of retrieved documents. Also, the longer the path between any two data elements (the more included data elements), the fewer the number of documents that will be retrieved.

The data-element/relation system (Figure 17.3) is completely defined (macroscopically) by the initial and final states (data-element 1 and data-element 2 in this case). However, the entropy of the "relational phase space" is greatly reduced by the actual identification of the specific relations A, B, C... This enables the indexing system to precisely define the relative "position" within the phase space of all the documents in the collection.

However, such a "structural" representation is not presently available in indexing systems. Data reaches the indexing system in the form of natural language strings, whose elements exhibit strong syntactic and semantic relations. We can infer that the number of such relations is significantly reduced after indexing, as evidenced by poor retrieval results. Paradigmatic

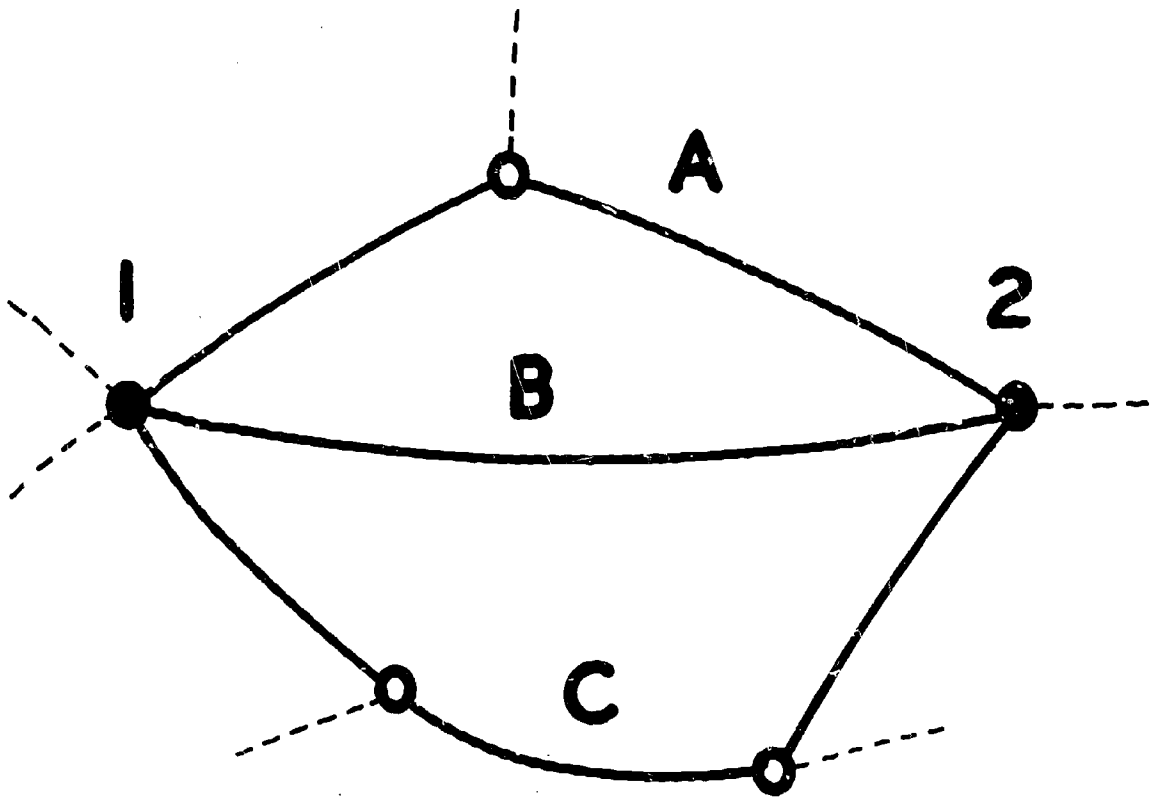


Figure 17.3: Data Element Relation Structure.
Circles represent terms and lines represent relations.

systems such as links and roles only serve to place the problem on a different level since they can provide only a limited number of syntactic or semantic markers for a data element. It is possible that a data element may not belong precisely to any such classes, and the potential misrepresentation of the syntax and semantics in a document can lead to false retrieval. Such difficulties may be avoided by preserving in the index the syntactic and semantic relations among data elements as given in the original document. This problem is at least partially resolved by a Case Grammar analysis of natural language strings.

Case Grammar was first described by Fillmore [59] in 1968 and presupposes causality and instrumentality in language. It is believed that the role and function (*e.g.*, "meaning") of words in deep structure is accurately portrayed by Case Grammar. We shall provide only a terse statement of the nature of Case Grammar since ample exposition is offered elsewhere [60]. Case Grammar (not to be confused with traditional notions of case) focuses on the pivotal role of the verb in natural language phrases.* Nouns are viewed as exhibiting a relationship with the coordinating verb. The relationships identified by Case Grammar include (and are denoted by the term *case*): agent, instrument, object, experiencer, possessive, source, time, location, manner and degree. The remaining words of the phrase (adjectives, adverbs, prepositions, etc.) are treated as facets of the case nouns.

The identification of the case grammar relations and the subsequent index entry generation involve six steps:

---

* Our analysis, following Cook [60], treats the clause as the basic "informational" unit in natural language discourse. "Clause" is defined as a word grouping containing one and only one predicate [62].

1) the input of text (title or sentence)

2) the identification of clauses

3) the identification of the verb (or auxiliary) within clauses

4) assignment of cases

5) facet isolation

6) index entry generation

   a) case index
   b) verb index
   c) facet index

Figure 17.4 shows a case grammar analysis of the title of the example document introduced in Section 11 of this Chapter. The words of the title are listed (in order of their occurrence) by case, verb or facet membership. The form of display is adapted from Cook [61]. Notice that the *case* and *verb* entries give the "essence" of the title: "effect-preventing-falls-tension-following-isoprenaline-subjects," while the facets provide the specifics. Figure 17.5 illustrates the index entries that would be created from this title. It is believed that the entries of the "Case Grammar Index" preserve the order of discussion and exhibit the organization of the underlying thought.

Finally, Figure 17.6 presents a structural representation* of the subject title. Content words are represented by capitalized letters and function words are represented by lower case letters. Connections in the structure represent the logical (relational) dependencies between the words. Notice the complete isomorphism between this structure and the corresponding case grammar assignments. The nodes with the highest connectivity correspond to the case grammar entries and the surrounding nodes correspond to the facets.

---

* Based upon that proposed by Rush [65].

Document # 820

N P D A A N P PRT
Effect of a selective beta-adrenergic blocker in preventing

N P A A N V N P
falls in arterial oxygen tension following isoprenaline in

A N
asthmatic subjects.

Syntax: N = noun D = determiner PRT = participle
P = preposition A = adjective V = verb

Case grammar: O = object (receiver of action)
LOC = location (place, extent, duration)

CASE	VERBAL	FACET
Effect (O)		blocker, beta-adrenergic, selective, a, of
	preventing	
falls (O)		
tension (LOC)		oxygen, arterial
	following	
isoprenaline (O)		
subjects (LOC)		asthmatic, in

Figure 17.4: Case Grammar Analysis of a Title.

Case Index

Effect [O], preventing falls 820
, blocker, beta-adrenergic, selective, of, 820

Falls [O], effect preventing 820

Isoprenaline [O], tension following 820

Subjects [LOC], asthmatic, in 820

Tension [LOC], following isoprenaline 820
, oxygen, arterial 820

Verb Index

Following, tension ___ 820
, ___ isoprenaline 820

Preventing, effect ___ 820
, ___ falls 820

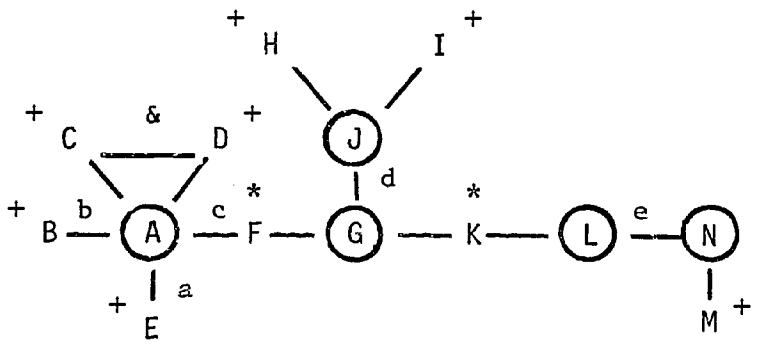
Facet Index

Asthmatic, in (subjects) 820
Arterial, oxygen (tension) 820
Beta-adrenergic, selective, a, of, blocker (effect) 820
Blocker, beta-adrenergic, selective, a, of, (effect) 820
Oxygen, arterial (tension) 820
Selective, a, of, blocker, beta-adrenergic (effect) 820

Figure 17.5: Index Entries Derived from Case Grammar Analysis.

Document # 820

A a b B C D E c F
Effect of a selective beta-adrenergic blocker in preventing
G d H I J K L e
falls in arterial oxygen tension following isoprenaline in
M N
asthmatic subjects.



◯ = case index entry

* = verb index entry

\+ = facet index entry

Figure 17.6: A Structural Representation of the Title Showing Isomorphism with the Case Grammar Analysis.

An indexing system would process and store this structure by reducing it to a connection-matrix representation (*e.g.*, by means of the Morgan [64] numbering algorithm) and one form of storage would be to represent the structure by the sequence of case nouns: A - - G (- J) - - L - N. Queries in this system would be effected by means of sub-structure searches.

18. The Index as a Tool of Inquiry

In this Chapter we have presented the basis for a comprehensive theory of Information Storage and Retrieval. Our thesis has been that this theory has its genesis in a theory of the indexing process. In other words, it is believed that the success of an IS&R system depends, primarily, on accurate and complete document representation, and that such representation is the goal of any indexing process. It has been contended that the index provides the necessary linkage between a multiplicity of sources and a single receiver. Conceptually, the indexing system is initially viewed as a black box that accepts documents as its inputs and produces the index as its only output. The various sources produce the documents which become the elements of the document space and the receiver produces queries which are matched against the index and, eventually, against the document store. Whether considering the source/document-space interface or the query/index interface, the elements of the underlying communication phenomena are the same: data elements and relations between data elements. Following the progression of schema presented in Figure 10.1, we first considered the necessary criteria for effective communication and concluded that the index provided the requisite common experience set between the source and the receiver. We then, more precisely, positioned the indexing system as intermediary between the communication channel and the receiver (searcher) and emphasized the role of

"noise" and feedback. Following a specification of the "position" of the black box or indexing system, we considered a theory of its operation. This theory, called the indexing process, defines the essential operation of the indexing system to be the creation of the representation of the document space. The analysis-document transformations and the final index-entry transformations were shown to be, respectively, a prerequisite to, and a function of, the document-space representation. Adequate examples of these transformations were provided through an analysis of the example document introduced in Section 11. Finally, the operating characteristics of the indexing system were modeled by means of the index space. From a different point of view, the concepts of error, organization, information and search were introduced through a consideration of the indexing process as a thermodynamic system. We could then postulate the existence of the "perfect" indexing system and the theoretical index as compared with their real-world counterparts.

We have cast the indexing process as a mechanical, well-defined set of operations; however, the use of the index data, by the receiver, presents an altogether different problem. As a consequence, a theory of the indexing process must also provide the means for the description of the process of searcher/index interaction. The modeling of the process of interaction is an admittedly "fuzzy" undertaking, but, it is believed that an understanding of the processes of index creation will provide the basis required for the analysis of search. Thus in this section we consider briefly the problem of the searcher-directed conversion between data and information and the concept of data element "value."

Following the discussion presented in Section 16, we must conclude that a query represents the searcher's hypothesis about the contents of the document space. However, only rarely would an initial hypothesis prove to be satisfactory with respect to the searcher's goal or "information need." It is believed that either he has an incomplete understanding of the organization of the system (the nature of the index entry and the indexing system representation) or he is unable to adequately formulate a hypothesis about its contents. Thus retrieval or search was modeled as a series of hypotheses and decisions which eventually end with goal achievement. Hopefully each interaction with the index leads to more precisely specified hypotheses and to hypotheses which are commensurate with the structure of the data base; such hypotheses will have a greater probability of yielding the desired goal.

We assume that the paths of maximal and minimal retrieval benefit, described in Section 16, have a small probability of occurrence. Consequently, most searcher/index interaction is adequately modeled by the intermediate case characterized by the alternation of hypotheses terminating with goal achievement. Hence,

Theorem 9.1 (Proof): The first data element retrieved, as a consequence of the first interaction with the index, will be of small benefit in goal achievement, since it will only provide meta-information leading to the formulation of a new hypothesis. This is a consequence of the high probability of occurrence of the intermediate case.

In the case of the maximal benefit, or the H - D - G path, we say that the data element that is retrieved, in response to the single hypothesis, has maximal utility or value since it immediately satisfies the information need

of the searcher. However if the same data element is retrieved after a series of hypotheses then its value, with respect to the information need, will have decreased or decayed. In the intermediate case, the retrieval of a given data element is dependent upon the prior sequence of hypotheses and data elements. Consequently, the decrease in utility of a data element is directly related to its position in a sequence of retrieved data elements (Thm. 9.2). The more hypothesis-testing and decision-making steps prior to the retrieval of a given data element then the smaller its utility--*i.e.*, the smaller its information content. Thus, any index/retrieval interaction longer than one operation sets up an n*th* order dependency between the n*th* retrieved data element and the n-1 ones previously retrieved.

We postulate that the value of a data element, with respect to goal achievement, is Poisson distributed over time. In Figure 9.1 the designated time intervals correspond to a succession of alternate hypotheses and consequent decisions. According to previous discussion, the greater the number of time intervals prior to the retrieval of a given data element then the smaller its utility or value. The downwards sloping curve is a direct consequence of the n*th* order dependency between successively retrieved data elements. The choice of a value for the parameter λ in the Poisson distribution

$$\frac{e^{-\lambda t}(\lambda t)^{v}}{v!}$$

controls the rate of decrease of value and is assumed to be characteristic of a given retrieval situation. It is possible that the value of λ depends on the experience of the searcher and that a decrease in the slope (over a sequence of sets of interactions) corresponds to the searcher learning the attributes that characterize the system and the index.

Assuming that the searcher is following the intermediate-path case then we postulate that although the value of each successively retrieved data element decreases, the rate of decay of the utility of newly retrieved data elements decreases with each new hypothesis. Figure 9.2 shows an envelop curve, E, which is the Poisson distribution of data element value. At each time, t_i, a new hypothesis is formed and a new data element is retrieved (of course, several data elements could be retrieved in response to a single hypothesis). The several curves, originating at each data element initial value, represent the decay in the utility of the data elements for goal achievement. The really significant observation is the number of data elements, in a given time interval, that are potentially of value for hypothesis testing and formulation. Thus, in the interval between t_4 and t_5 both of the data elements retrieved at t_3 and t_4 are useful in decision making and hypothesis formulation. We postulate that the rate of decay of the utility of a given data element is a function of the initial value (given by the curve E) and the necessity of forming the next hypothesis or making the next decision (characteristic of the problem solving situation).

The most obvious conclusion from this brief analysis of the search interface is the need for a "process of inquiry" characterization of the process of retrieval. We have argued that the indexing system must present data elements and relations to the searcher, but how are we to evaluate the effectiveness of this presentation--especially when comparing alternative systems? Possibly the greatest hindrance to such an understanding is the lack of precision associated with the concept of "information need." What does it mean to say that data elements satisfy an information need? Accordingly, this will be the major topic of discussion for Chapter 5. As we shall

see, an understanding of "information need" will resolve the apparent divergence between the concepts of retrieval effectiveness, the concept of search paths, hypothesis testing and the searcher's understanding of the organization of the document collection.

APPENDIX A

A Case History Illustrative of Index Failure *

During the closing stages of his doctoral research career, a friend was engaged in the study of certain chemical reactions which he characterized generally as follows:

$$RB(OH)_2 + 2\ Cu_2 + H_2O \rightarrow RX + Cu_2X_2 + HX + B(OH)_3$$

Believing that his searches of the literature had uncovered all available data on this type of reaction, he was somewhat chagrined to learn, near the end of his studies of the reaction, from a colleague that another document of (apparent) importance existed which he had failed to unearth. More than a little unsettled by this occurrence, he made an exhaustive effort to find the document in question through all available means. There were a number of obvious places one might look in an index for this document. Some of these are listed below without embellishment.

1. $RB(OH)_2$
2. CuX_2
3. RX
4. Cu_2X_2
5. Reaction of $RB(OH)_2$ with CuX_2
6. Production of RX from $RB(OH)_2$
7. Reaction mechanisims, of $RB(OH)_2$ with CuX_2
8. Specific compound names (of which there were potentially a large number)

* Names have been omitted to avoid unnecessary adverse criticism of specific persons or systems.

Since specific compounds reported in the document could not be anticipated, my friend had recourse only to those index entries of a more general nature, plus those specific compounds which has experience suggested he check. This rather exhaustive search failed to yield the desired document. So, my friend, having retrieved the document by means of the information supplied by the colleague, endeavored to find the index entries which corresponded with specific details in the document. The result obtained was that only the specific $RB(OH)_2$ compounds were indexed (without qualification) and that no index entries had been generated for any other part of the document.

The obvious conclusion which he drew was that a gross error had been committed in the indexing of this particular document. But this failure causes one to wonder whether other similar failures have gone undetected (where this one was detected only by chance). In any event, such occurrences certainly put the index user in an uneasy frame of mind, and, if opportunity exists, he will most likely turn to a different index rather than take a second chance with the one which has failed him.

References

1. J.S. Bruner, J.J. Goodnow and G.A. Austin, *A Study of Thinking*, John Wiley & Sons, Inc., New York, N.Y., 1956, 26.

2. C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, Illinois, 1964, 7.

3. L. Brillouin, *Science and Information Theory*, Academic Press, New York, N.Y., 1956, 159-161.

4. W. Heisenberg, *Nuclear Physics*, Philosophical Library, New York, N.Y., 1953, 30.

5. R.L. Collison, *Indexes and Indexing*, John deGraff, Inc., New York, N.Y., 1959, 20.

6. H. Garfinkel, *Studies in Ethnomethodology*, Prentice-Hall, Englewood Cliffs, N.J., 1967, 279.

7. D. Sorgel, "Mathematical Analysis of Documentation Systems--an Attempt to a Theory of Classification and Search Request Formulation", *Information Storage and Retrieval* 3, 1967, 129-173.

8. P.J. Cohen and R. Hersh, "Non-Cantorian Set Theory", *Scientific American* 217(6), 1967, 104-116.

9. C.E. Shannon and W. Weaver, *op. cit.*, 4.

10. C. Cherry, *On Human Communication*, The M.I.T. Press, Cambridge, Mass., 1966, 305.

11. *ibid.*, 7.

12. H. Garfinkel, *op. cit.*, 24-34.

13. B.C. Landry, N.M. Meara, H.B. Pepinsky, J.E. Rush and C.E. Young, "A Computer Assisted Study of Informative Display", Unpublished Manuscript, 1971, 2-8.

14. C. Cherry, *op. cit.*, 221-222.

15. D.M. Mackay, "The Place of 'Meaning' in the Theory of Information", in C. Cherry (ed.), *Information Theory: Proceedings of the 3rd Symposium*, Butterworths, London, 1955, 215-225.

16. B.C. Landry, N.M. Meara, H.B. Pepinsky, J.E. Rush and C.E. Young, *op. cit.*, 2.

17. E.E. Graziano, "On a Theory of Documentation", *American Documentation* 19(1), 1968, 85.

18. R.A. Fairthorne, *Toward Information Retrieval*, Butterworths, London, 1965, 70-77.

19. R.L. Collison, *op. cit.*, 19.

20. M.G. Mellon, *Chemical Publications: Their Nature and Use*, McGraw-Hill Book Co., New York, N.Y., 1965, 203.

21. M. Grems and S. Fisher, "Primigenial Indexing for Heuristic Retrieval", in B.F. Cheydleur (ed.), *Proceedings of the Colloquium on Technical Preconditions for Retrieval Center Operations, Philadelphia, Pa., 24-25 April 1964*, Spartan Books, Washington, D.C., 1965, 117-123.

22. M.G. Mellon, *op. cit.*, 34.

23. H. Skolnik, "Chemical Indexing: Management's Point of View", *Journal of Chemical Documentation* 1(1), 1961, 57-61.

24. L.C. Pauling, *The Nature of the Chemical Bond*, Cornell University Press, Ithaca, N.Y., 1940.

25. R. Geballe, "Matters of Concinnity", *Science* 172, 1971, 688-690.

26. *ibid.*, 690.

27. F. Levéry, "Les Problemes Posés par le Vocabulaire Documentaire et l'organization des Dictionaries et Thesaurus", *Proceedings of the AGARD Convention*, North Alantic Treaty Organization, 1968, 23-37.

28. F. deSaussure, *Course in General Linguistics*, Philosophical Library, Inc., New York, N.Y., 1959.

29. R.A. Fairthorne, "Content, Analysis, Specification and Control", in C.A. Cuadra (ed.), *Annual Review of Information Science and Technology 4*, Encyclopaedia Britannica, Inc., Chicago, Illinois, 1969, 79.

30. C.L. Bernier, "Correlative Indexes IX. Vocabulary Control", *Journal of Chemical Documentation* 4, 1964, 101.

31. M.C. Yovits and R.L. Ernst, "Generalized Information Systems: Consequences for Information Transfer", in H. B. Pepinsky (ed.), *People and Information*, Pergamon Press, Elmsford, N.Y., 1970, 1-31.

32. L. Brillouin, *op. cit.*, 289.

33. C.E. Shannon and W. Weaver, *op. cit.*

34. R. Carnap and Y. Bar-Hillel, "An Outline of a Theory of Semantic Communication", *M.I.T., Research Lab. Electronics, Tech. Rept. 247*, 1953.

35. M. C. Yovits and R. L. Ernst, *op. cit.*

36. C.E. Shannon and W. Weaver, *op. cit.*, 48-53.

37. R.A. Fisher, *Statistical Methods for Research Workers*, Hafner Publishing Co., New York, N.Y., 1946.

38. S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, Inc., New York, N.Y., 1959.

39. M.C. Yovits and R.L. Ernst, *op. cit.*, 4.

40. R.A. Fairthorne, *Towards Information Retrieval*, *op. cit.*, 69.

41. A.D. deGroot, "Perception and Memory versus Thought: Some Old Ideas and Recent Findings", in B. Kleinmuntz (ed.), *Problem Solving*, John Wiley & Sons, Inc., New York, N.Y., 1966, 19-50.

42. F.W. Lancaster, *Information Retrieval Systems*, John Wiley & Sons, Inc., New York, N.Y., 1968, 83.

43. M.E. Maron and J.L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval", *Journal of the Association for Computing Machinery* 7, 1960, 216-244.

44. C.T. Meadow, *The Analysis of Information Systems*, John Wiley & Sons, Inc., New York, N.Y., 1967, 14-63.

45. L. Schmerling, "Chemical Indexing: The Research Chemist's Point of View", *Journal of Chemical Documentation* 1(1), 1961, 46-51.

46. R.A. Fairthorne, *op. cit.*, 65.

47. M. Tribus and E.C. MacIrvine, "Energy and Information", *Scientific American* 224(3), 1971, 179-188.

48. C. Caratheodory, "Gründlagen der Thermodynamic", *Math. Ann.* 67, 1909, 355.

49. J. Rothstein, "Informational Generalization of Entropy in Physics", *Ohio State Univ., Computer and Information Science Research Center Technical Report 70-24*, 1970.

50. W. Heisenberg, *op. cit.*

51. J. Rothstein, "Toward a General Theory of Information Storage, Search and Retrieval", Unpublished Manuscript.

52. E.E. Reid, *Invitation to Chemical Research*, Franklin Publishing Co., Inc., Palisade, N.J., 1961, 295.

53. M.G. Mellon, *op. cit.*, 211.

54. J.E. Armitage and M.F. Lynch, "Articulation in the Generation of Subject Indexes by Computer", *Journal of Chemical Documentation* 7, 1967, 170.

55. J.R. Sharp, "The SLIC Index", *American Documentation* 17, 1966, 41-44.

56. A.E. Petrarca and W.M. Lay, "The Double-KWIC Coordinate Index: A new Approach for Preparation of High-Quality Indexes by Automatic Indexing Techniques", *Journal of Chemical Documentation* 9, 1969, 256-261.

57. J.C. Costello, Jr., "Indexing in Depth: Practical Parameters" in P.W. Howerton (ed.), *Information Handling: First Principles*, Spartan Books, Washington, D.C., 1962.

58. R.C. Cros, J.C. Gardin and F. Levéry, L'Automatisation des Recherches Documentaires--Un Modèl Générale -LE SYNTOL-, Gautier-Villars, Paris, 1964.

59. C.J. Fillmore, "The Case for Case", in E. Bach and R. Harms (eds.), *Universals in Linguistic Theory*, Holt, Rinehart & Winston, Inc., New York, N.Y., 1968, 1-88.

60. B.C. Landry, N.M. Meara, H.B. Pepinsky, J.E. Rush and C.E. Young, *op. cit.*, 9.

61. W.A. Cook, S.J., "Case Grammar Analysis of 'The Old Man and The Sea'", Unpublished Mansucript, 1971.

62. W.A. Cook, S.J., *Introduction to Tagmemic Analysis*, Holt, Rinehart & Winston, Inc., New York, N.Y., 1969.

63. *ibid.*, 65.

64. H.L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service", *Journal of Chemical Documentation* 5, 1967, 101-113.

65. J.E. Rush, "A New Approach to Indexing", presented before the Student Chapter of A.S.I.S., Case-Western Reserve University, Cleveland, Ohio, January 1969.

CHAPTER V. ON RELEVANCE AS A MEASURE FOR IS&R

> "You have seen the literary articles which have appeared at intervals in the Eatanswill Gazette in the course of the last three months, and which have excited such general--I may say such universal--attention and admiration?"
> "Why", replied Mr. Pickwick, slightly embarrassed by the question, "the fact is, I have been so much engaged in other ways, that I really have not had an opportunity of pursuing them."
> "You should do so, sir," said Pott, with a severe countenance.
> "I will," said Mr. Pickwick.
> "They appeared in the form of a copious review of a work on Chinese metaphysics, sir," said Pott.
> "Oh," observed Mr. Pickwick; "from your pen, I hope?"
> "From the pen of my critic, sir," rejoined Pott with dignity.
> "An abstruse subject I should conceive," said Mr. Pickwick.
> "Very, sir," responded Pott, looking intensely sage. "He crammed for it, to use a technical but expressive term; he read up for the subject, at my desire, in the *Encyclopaedia Britannica*."
> "Indeed." said Mr. Pickwick; "I was not aware that that valuable work contained any information respecting Chinese metaphysics."
> "He read, sir," rejoined Pott, laying his hand on Mr. Pickwick's knee, and looking round with a smile of intellectual superiority, "he read for metaphysics under the letter M, and for China under the letter C, and combined his information sir."
>
> Charles Dickens

1. Introduction

There is a timeless quality to the method used by Mr. Pott's critic. Indeed, Dickens has created a character who might be our contemporary. Information storage and retrieval systems have changed--they are larger and faster; but the same problems still exist--retrieval is still accomplished by the elementary combination of "information" on various subjects. Our methodology has changed but we are still as unsure of the result as was Mr. Pott. The problem of retrieval-system evaluation becomes of paramount importance. It is assumed that an understanding of the concept of *relevance* is essential to the solution of systems evaluation. Accordingly, this

chapter is directed toward the definition of the problem of relevance, to definitions and history of the relevance and evaluation concepts as applied to system performance, to a schema for IS&R systems evaluation and, finally, toward new directions in systems evaluation.

2. The Problem of Relevance

Wooster [1] has recently enumerated some of the many criteria available for the evaluation of the effectiveness of Information Analysis Centers. There criteria generally fall into five broad classes: need, use, cost, performance and benefit. Although his listing was specifically directed toward the *Analysis Center* concept, similar criteria are easily applied to the general information storage and retrieval evaluation problem. Wooster's exposition shows that current measures are diverse and exhibit little consistency of approach. Such conditions can only lead to confusion. Unless we throughly understand the problems of system evaluation, performance and benefit, efforts toward system description and comparison will merit Rees's [2] phrase: "...busy people spending large sums of money, designing--or attempting to design--phantom systems for non-existent people in hypothetical situations with unknown needs."

But then, what is *evaluation*? To Richmond [3]

> The very term *evaluation* suggests a qualitative procedure--making a value judgment. The quantification of evaluation is a matter of abstracting those factors that are not purely human, apparently such as performance and operation, and setting them aside to function as data upon which to make a value judgment.

Evaluation, then, reduces to "making a value judgment." But why? It is easy to say that a judgment is made of performance and operational data, but toward what end? This question is partially answered if we assume that a

value judgment serves as a two place predicate between a datum and a well-defined end or goal. We say, for example, "x *equals* y", "x *satisfies* the goal y", or "x *completes* the process y." In information storage and retrieval such a value judgment does exist--but between which "x" and which "y"? There are five candidates available from the elements of the retrieval process (*cf.* Chapter IV):

- The informational need (goal) of the user.
- The expression of this need (query).
- The corpus of documents.
- The system for retrieval.
- The set of retrieved documents.

As we shall see later, many value judgments can be made on the performance* of system components (indexing, abstracting, thesaurus, document acquisition, query processing, etc.) that are apt to be hidden from a user; but the value judgment of primary importance is that which speaks to how effectively, from the user's viewpoint, the objectives of the search are being met. This value judgment is the correlation between the expression of the need (the formulation of the query which represents the informational goal) and the set of retrieved documents (the result of the system's action). The most pertinent question, or judgment, from a user's point of view is whether the retrieved documents satisfy the goal requirement.

A logical extension of Richmond's view of evaluation is offered by Stevens [4]:

> The most generally accepted criterion for appraising the effectiveness of indexing [or systems in general] is that

---

* Performance is usually measured in terms of efficiency (retrieval time and cost parameters) and effectiveness (goal satisfaction).

> of *retrieval effectiveness*. But, in general, this is merely the substitution of one intangible for another, entailing a string of yet unanswerable or at least unresolved questions. Retrieval of what, for whom, and when? How can effectiveness be measured except by the relative question of relevance judgments? How can human judgments of *relevance* and *value* [italics added] be measured and quantified?

Thus, a value judgment between the retrieved documents and the expression of a need is *relevance*. The criterion of relevance is an expression of the connectivity or linkage between documents and a request. As Hillman [5] sees it:

> The problem is to describe a concept of relevance independent of, and logically prior to, any notion of relevance as determined by, and thus restricted to, a particular system of storage and retrieval.

This is, as we shall see, the most logical criterion for system evaluation, but the best measure, instrument and methodology have yet to be implemented. Accordingly, relevance measurement has a lengthy and conceptually fuzzy history in information storage and retrieval.

3. Definitions and Measures of Relevance

3.1 Definitions

The domain of Information Storage and Retrieval suffers from an overabundance of definitions of *relevance*. Consistency is difficult to maintain between studies because each study that is undertaken involves a different definition of relevance. In addition, the introduction of new terminology forces the construction of new definitions of relevance, or the modification of previous ones. In this review, for convenience of presentation, we have chosen to condense and enumerate the various definitions of relevance under four headings:

- Dictionary definition:

  -Relevance is a relation to the matter at hand.

- Communication definition:

  -Relevance is a measurement of information transfer.

  -Relevance is a phenomenon of communication indicating relations.

- Value definition:

  -Relevance is the amount of satisfaction in information transfer.

  -Relevance is the "appropriateness" of the document to the user.

  -Relevance is the "utility" of the document to the user.

  -Relevance is the "satisfaction" derived by the user.

- Connection definition:

  -Relevance is that fraction of the retrieved material that is actually relevant to the request.

  -Relevant documents are those that describe situations identical with that specified by the requester.

  -Relevance is a user decision about document/query match.

  -Relevance is the occurrence of each descriptor of the search profile of the request in that of the document.

Very few of these definitions are of immediate use in the quantification of the document-relevance/decision process. They are largely qualitative and should be interpreted as merely indicative of a *philosophy* of relevance.

It is premature to suggest another definition of relevance, but it would be advantageous to list some important attributes and desirable features of an improved definition of relevance. Some of these attributes are included in the conclusions reached by the *1958 International Conference on Science Information* [6]:

- Relevance is more than the operation of relating what is performed internally within systems;
- Relevance is not exclusively a property of document content;
- Relevance is not a dichotomous decision;
- There is such a thing as "user relevance" that can be judged.

This listing indicates directions for future *relevance* research.

3.2 Measures of Relevance

Bourne, in his review of the *Evaluation of Indexing Systems* [7], identified 31 terms used in relevance measures. Many of these terms differ only superficially, but their number certainly indicates the diffuseness of the state-of-the-art of measures of relevance. For this very reason, they are deemed worthy of enumeration:

recall	precision ratio	accuracy
recall factor	normalized precision	efficiency
recall ratio	sensitivity	snobbery ratio
relative recall	productivity	fallout ratio
normalized recall	relative productivity	discrimination
relevance	specificity	distribution
relative relevance	effectiveness	resolution
generality	hit rate	elimination
generality ratio	acceptance ratio	pertinency factor
precision	completeness	omission

Bourne's conclusion denies the existence of a current *science of relevance assessment*:

> The experimental work reported in the literature seems to have used almost completely different measures for every single experiment reported.

The results of these studies are not only frequently non-reproducible, they are non-comparable.

From these introductory remarks it seems pointless to dwell on these "measures". The interested reader is thus referred to several good reviews of the pertinent literature [8-11]. The conclusions to be drawn from this

literature are not encouraging: There are too many different measures of relevance; there are too many mathematical forms employed; there are too many variations in method; and, the results, if at all meaningful, are mostly system specific.

Early studies characterized relevance as a highly subjective measure that indicated the degree of match between retrieved documents and a query. It was generally agreed that the concept of relevance was not identical to the contents of documents, but was, rather, a form of conceptual relatedness between query and retrieved document. Efforts at quantification produced the measures of *recall* and *precision* that serve as the formal basis for most of the 31 terms listed above:

$$\text{recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents in the system}}$$

$$\text{precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}}$$

Suddenly (as Taube [12] argues), the subjective notion of relevance is given a spuriously precise, mathematical definition. Taube cites the Cranfield studies [13] and the Arthur D. Little report [14] as primary causes for what he terms *the pseudo-mathematics of relevance*. A series of questions may be asked of *these measures*: How does one know *a priori* the *number of relevant documents in a system*? Does recall apply only to contrived systems where the team of evaluators has total knowledge of the system's contents? (Or, if not, and total knowledge is available, then what is the excuse for a system yielding recall values below 100%?) What is meant by *relevant documents*? Is this a circular definition of relevance?

These measures are even less palatable when they are applied to the task of systems evaluation. Let us consider an example of the application

of the measures of recall and precision to the evaluation of a hypothetical system. First, we assume that this system has the mechanism by means of which a search (query formulation) can be performed over its data base. Second, we assume that there exists a facility to vary search strategies, through, for example, boolean combinations of terms, in order to increase, decrease or mix values of recall and precision. Finally, we assume that for each search a value of recall and precision can be calculated.* The details of a general search (A) and of successively more precise sub-searches (B-E) are as follows:

Search	Search Terms**
General search strategy (A)	M only
Sub-search (B)	M&N
Sub-search (C)	$M\&N_1$
Sub-search (D)	$M\&N_1\&O$
Sub-search (E)	$M\&N_1\&O_1$

Exhaustive analysis of the system's data base yields a recall/precision point-pair for each of the above searches. Their hypothetical values are plotted on a standard recall/precision graph (see Figure 3.2.1).

One may not argue the existence of these points. But there is no reason to assume that these separate data points can be joined by a curve. If we remember that a curve is an expression of a functional relation between data points, then it is not evident that such a relationship exists between

---

* This is an unwarranted assumption: in addition to the problem of deciding what is relevant, the total number of relevant references is unknown short of exhaustive system search.

** N is generic to N_1; O is generic to O_1; & represents logical AND.

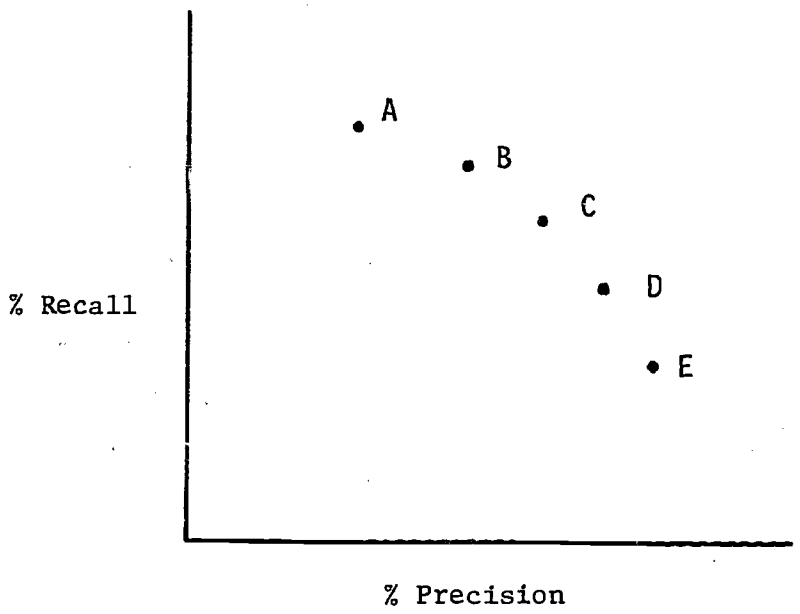


Figure 3.2.1: The Recall-Precision Graph.

successive determinations of recall/precision point-pairs. Recall/precision graphs are often used (seemingly, unknowingly) to indicate an unvalidated dependence between successive searches.

Viewed differently, the assumption of a functional relationship implies the existence of values intermediary between the five data points plotted in Figure 3.2.1. This also implies that there exist additional terms in the system (apart from other combinations of M,N,N_1,O,O_1) that can be used to vary the specificity of the search. The assumption of the existence of additional terms is faulty if the only index terms in the system are M,N,N_1,O,O_1.

Lancaster's [15] experience with MEDLARS* has shown a wide point-scatter in recall/precision plots obtained from his test search results. Indeed, the recall and precision points he obtained for the various test searches were near random in nature. Clearly this suggests the absense of any correlation or functional relationship between searches. It appears that the measures of recall and precision are simply indicative of a "fuzzy-zone" of system performance.

As if to destroy the utility of system performance curves, O'Hara [16] has demonstrated that of the eight possible boundary positions on a recall/precision plot (see Figure 3.3.2), two require contrived definitions to be meaningful, and two are clearly impossible (0% recall and less than 100% precision, 0% precision and less than 100% recall).

Extensions of the concepts of recall and precision have involved both micro and macro definitions [17], probability measures [18], decision table

---

* Medical Literature Analysis and Retrieval System.

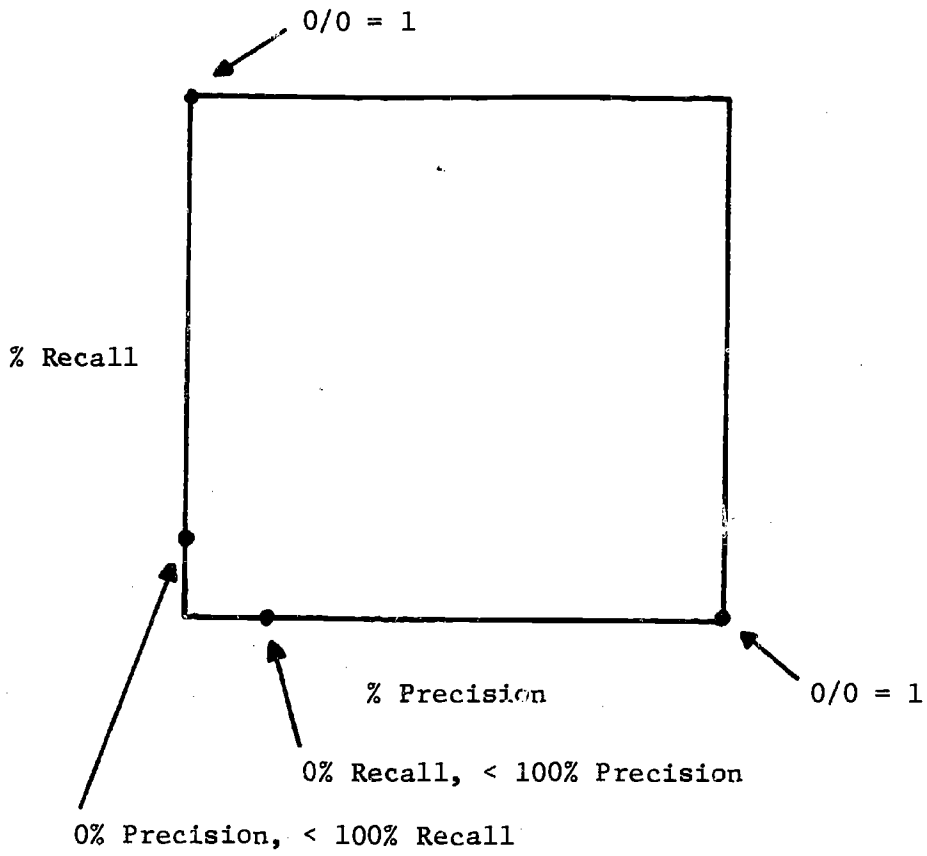


Figure 3.2.2: Limiting Cases Associated with a Recall-Precision Graph.

analysis [19,20], expected search length [21], and relevance feedback [22] to mention a few. In my opinion these studies rest upon faulty postulates--*i.e.*, recall and relevance. Informally and from a subjective viewpoint, these measures may have some value, but they cannot serve as a basis for formal analysis. Clearly, relevance has yet to be assigned a definition suited to quantification and mathematical manipulation. Without such a definition, empirically testable generalizations about systems performance are impossible.

4. A Schematic for IS&R Systems Evaluation

If it is difficult to find a consensus on a definition of relevance, it is as difficult to evaluate comprehensively studies of systems evaluation. The number of factors and interfaces examined in these studies are numerous. Cuadra [23] has identified some of the features of relevance judgments:

> Evidence has been developed that suggests that relevance judgments can be and are influenced by the skills and attitudes of the particular judges used, the documents and document set used, the particular information requirement statements, the instructions and settings in which the judgments take place, the concepts and definitions of relevance employed in the judgments, and the type of rating scale or other medium used to express the judgments.

Cuadra's [23] final report in *Experimental Studies of Relevance Judgments* identifies four broad classes of factors that influence the relevance judgment decision:

Document	Judgmental Conditions
-Subject matter	-Time of judging
-Diversity of content	-Order of presentation
-Difficulty level	-Size and breadth of document
-Scientific hardness	-Use of control judgments
-Amount of information	-Specification of task
-Level of condensation	-Definition of relevance
-Textual attributes	

Information requirement statement	The Judge
-Subject matter	-Knowledge/experience
-Difficulty level	-Intelligence
-Diversity of content	-Cognitive style
-Specificity of information	-Biases
-Functional ambiguity	-Judging experience
-Textual attributes	-Attitude
	-Distribution expectancy

To me, the problem lies in the number of different interfaces (or points of correspondence) in the information storage and retrieval process where a value judgment can be effected and measured. The following model is presented as an aid in the clarification and classification of the various relevance judgmental decisions.

4.1 The Model

Figure 4.1.1 shows the position of the index in the information storage and retrieval process; note the importance of the feedback process in the search operation. The model is divided into three units based on the fundamental operations of document creation, representation and retrieval. These operations are also depicted in Figure 4.1.2, which shows that the indexing operation encompasses document acquisition, representation and storage, while retrieval is represented by the exchange between a user's informational need and the expression of this need. The double arrow between these two components is indicative of the feedback that is essential to these activities. Noise has not been indicated but it could perturb any of the components or operations involved in the communication process. The seven dotted lines indicate the various types of judgmental decisions that are applicable to systems evaluation. Each is described briefly as follows:

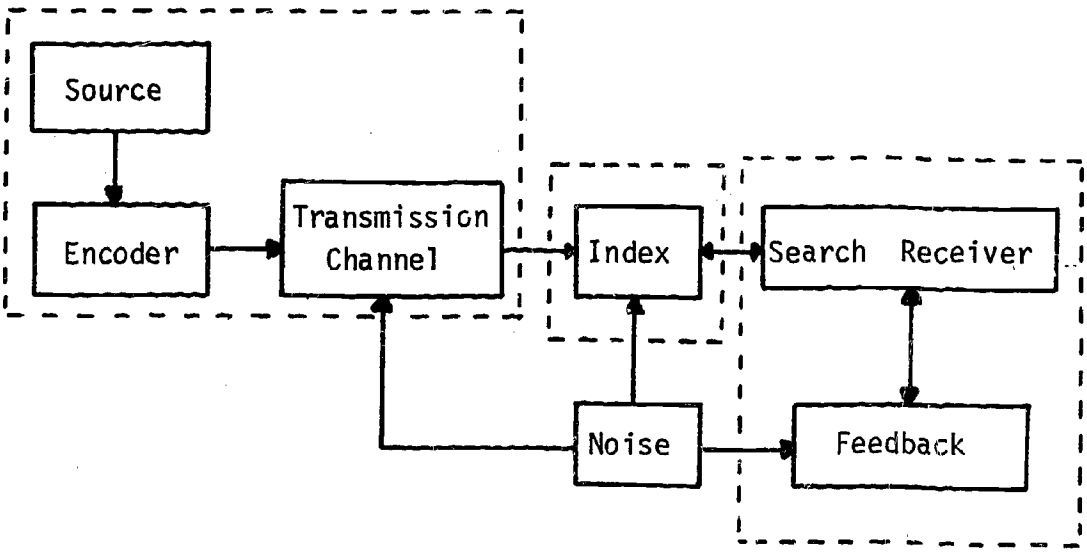


Figure 4.1.1: The Position of the Index in the IS&R Process.

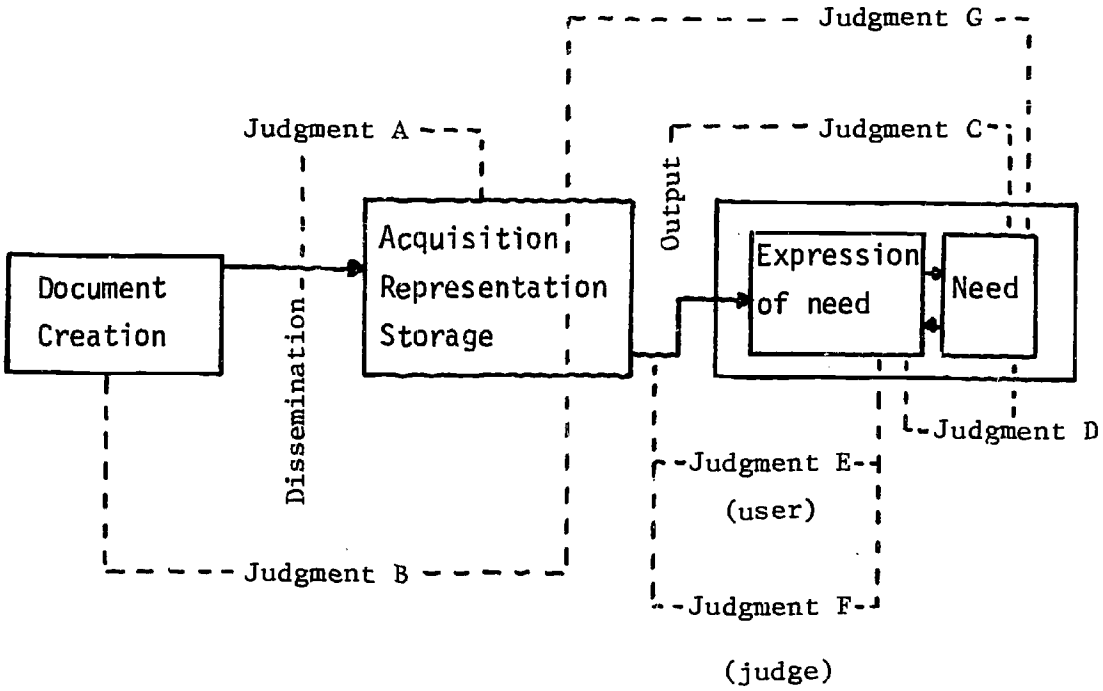


Figure 4.1.2: Evaluation Judgments.

Judgment A:

Judgment A is probably the least studied interface because of the difficulty of its measurement. However, to a first approximation, a system is only as good as the documents that it collects. This is why efforts should be directed toward evaluation of system input procedures. From a slightly different point of view, Paisley and Parker [24] have recognized that source deficiencies may severely limit system performance. To this author's knowledge no comprehensive studies have been undertaken dealing with judgment A.

Judgment B:

The "aboutness" judgment (as Fairthorne [10] puts it) receives the greatest amount of attention by researchers--possibly because it is the easiest to measure. As discussed in Chapter IV, accurate document representation is the most important system input function. Accordingly, the exhaustiveness of the indexing and the specificity of the indexing language are the most often studied. While Zunde [25] concludes that "documental" factors are the most important parameters in indexer consistency, St. Laurent [26], in a comprehensive survey, opines that no conclusion can be reached from the diverse studies of indexer consistency. One of the better summaries of this interface is proposed by Lay [27]. He defines three types of index terms (T) and relations (R) between terms based on their presence in the document, the index, or both (see Figure 4.1.3). It is believed that an effective study of interface B might employ such a decomposition as its underlying model.

	In Document	In Index
T_1	+	+
T_2	-	+
T_3	+	-
R_1	+	+
R_2	-	+
R_3	+	-

+ = present
- = absent
T = terms
R = relations

Figure 4.1.3: Term/Relation Analysis.

Judgments C and D:

In my opinion the most significant correlations for evaluation are those between the need and the expression of the need, and between need and the retrieval output. Unfortunately these distinctions are largely overlooked in the literature of relevance. Unless the user's *need* is satisfied the retrieval is *not* effective.

Judgments E and F:

Taulbee [9] characterizes these judgments by the decision on the relationship between the "information" need and a given document. (There is some debate as to whether this judgment should be on a corpus of retrieved documents rather than on a 1-1 basis-see Goffman [28]). These relevance judgments are often formalized by means of document and query vectors that permit facile comparisons. The reader is directed to reports on the SMART system [17] and of Ide's [22] "relevance feedback" analysis for details.

Although it is claimed that these judgments can be made by the user or by an independent judge (observer or mathematical criterion), O'Conner [29] takes a different view:

> The basic causes of relevance disagreements are differences in interpretation of requests or documents, rather than such factors as the education of the judges and what they take to be the purpose, environment and timing of the request.

Judgment G:

No studies have dealt with the difficult questions of the correlation between the user's need and the system's representation (we will not consider *selective dissemination* as representative of the essence of judgment G). Leslie [30] humorously depicts this difficult judgment:

> Somebody has defined indexing as a game involving two players--an indexer and a user. In this game, the first player (the indexer) tries to guess where the user will look for a particular

> record. The second player (the user) tries to guess where the indexer put it. The game gets a little complicated when the user tries to guess where the indexer guessed the user would guess the indexer guessed the user would look for it.

Many systems can, unfortunately, be described in these terms.

5. Directions

It has been argued that the *systems evaluation* problem reduces to the task of document *relevance* assessment. This conceptual reduction does not, however, yield a corresponding reduction in the difficulty of solution. Furthermore, the situation is worsened by the many "pseudo-" measures of relevance, all of which seem to rely on the less-than-satisfactory measures of *recall and precision*. Clearly, the task of relevance assessment is ripe for new directions.

The model presented above serves, mainly, for the enumeration of the many judgmental inferfaces that exist in the generalized information storage and retrieval process. Unfortunately, this enumeration has but increased our awareness of the difficulty of deciding just what is relevant, and on what basis.

The main inference to be drawn, apart from that of the chaotic state of relevance studies, is that *systems evaluation must be centered on the needs of its users*. That a user may not want *all* of the relevant references that a system can provide--perhaps the first one retrieved will be sufficient to satisify his *informational need*--should also be taken into consideration. This suggests that evaluation research should be directed not toward the correspondence between query and document, but toward the identification of the attributes of the user's goals. Judgment G of the model described in Section 4.1 must be an integral part of such research.

Cuadra [31] summarizes the problem:

> Nearly all studies purporting to evaluate the effectiveness of information retrieval systems have relied very heavily on the notion of a "relevant set of documents" identified by a particular set of judges. The relevance judging process in these studies has been treated largely as a "black box", with little serious effort to understand what happens inside the box or how variations in the judgments might lead to variations in the "relevant set of documents."

There is, then, a clear difference between relevance to a query and relevance to a need. This calls for the analysis of the searcher-receiver box of the model which represents the correlation between the *expression of the need* and the *need*. Figure 5.1 shows an adaptation of a model proposed by Kegan [32], and indicates some of the pertinent factors of this correlation. It is assumed that a more nearly complete representation and understanding of the above factors will aid in deciding what is *information* both to the user and to the corresponding relevance judgment.

6. Interregnum

The notable conclusion to be drawn from these introductory sections is that the field of Information Storage and Retrieval (IS&R) lacks a comprehensive theory of retrieval systems evaluation--an unfortunate circumstance since a solid theoretical foundation is essential for the characterization and evaluation of retrieval experience. Stated differently, an observer's experience in the real world has meaning only if he has a sound *predictive model*. The diffuseness of any existing predictive model (assuming that a model does exist) is indicated by several observations:

- The problem of the definition of "systems evaluation"--*e.g.*, which of several possible definitions is the most useful?
- The IS&R researcher is confronted by a wide range of definitions of relevance--which definition is of greatest value?

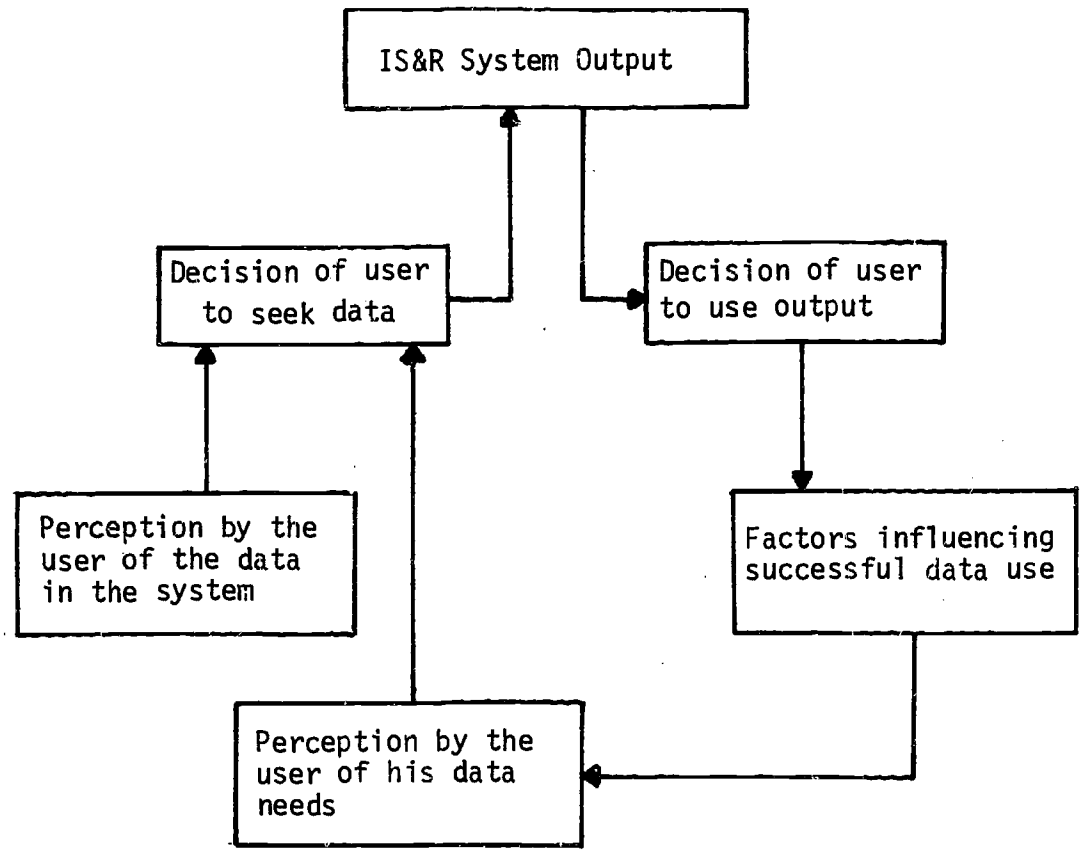


Figure 5.1: IS&R Decisions.

- There is an equally extensive collection of "measures" of relevance--which of these measures is meaningful, and in what sense?

- One is plagued by indecision and confusion when presented with the various possible forms of relevance judgment--what is the meaning of all of these "judgments"?

The most important observation that can be made, at present, is that *evaluation is a judgmental relation*, which is best characterized as a two place predicate or relation between a datum (or data) and a goal. One should not be constrained to think of this relation as a simple operator in a formal calculus, rather, it is hypothesized that the evaluation relation is a place-holder for a process or an algorithm connecting data and goals. Evaluation is viewed as a process which is best characterized by judgments C and D, as depicted in Figure 4.1.2. More explicitly: evaluation is a relational algorithm for measuring the strength of connection between the informational need and the *retrieved documents* plus the *expression of the need* (see Figure 6.1). Retrieved documents are the retrieval results from an IS&R system and the expression of the need is the query presented to the system by the user. The main problem with this definition of evaluation is that the term "information need" is a fuzzy concept. The purpose of the following sections is to concretize the concept of information need.

The Kegan model (see Figure 5.1) is a representation of the processes employed in the interaction (represented by the double arrow (↔) in Figure 4.1.2) between the need and the expression of the need. Three important attributes of interaction are identified which are worthy of enumeration:

- The user's identification of the factors required for successful data use.

- User decisions.

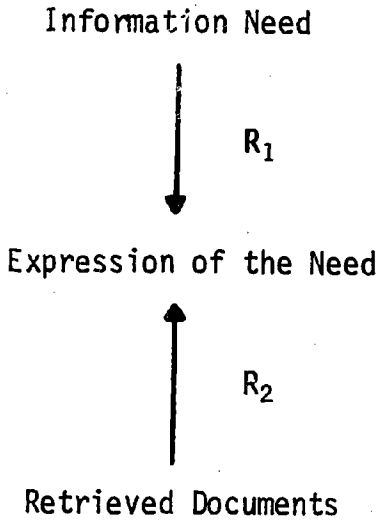


Figure 6.1: Interaction Between the Need and the Expression of the Need.

- The perception by the user of what it is that he needs and his perception of what the system has to offer.*

These attributes are not only useful in the description of this specific interface, but, as we shall see, they are also instrumental in the analysis of *information need*. System attribute identification, user decision making and user perception(s) are all terms that will assume increasing importance as we strive to better understand the need/expression-of-the-need interface. But it is already clear that attribute identification, decision making and perception represent complex and highly dynamic activities.

Furthermore, inquiry may be represented as the progression, or iteration, of queries presented to the system. Thus, there is no reason to assume, *a priori*, that the nature (form, status, mode) of the above-listed activities is invariant through several inquiry iterations with an IS&R system. And, if we can accept the idea that information need is a dynamic concept, then we must ask how these activities arise, how they interact among themselves and how they contribute to the identification and use of "relevant" data.

In what follows, an attempt will be made to characterize information need and system interaction through a consideration and development of the following topics: observation and measurement, experience, the central role of the process of inquiry, the goal as information need, the reason for the the existence of the goal, hypothesis testing, decision making, estimation of probability and information gain.

---

* It is important to realize that these perceptions are not necessarily the same. These perceptual differences are the source of considerable error in retrieval system interaction.

7. Information Need

7.1 The Problem Posed to IS&R

The standard IS&R-oriented definitions of "information need" are nicely enumerated by O'Connor [33]. He identifies three broad meanings of the statement "satisfying a requester's information need":

- Request negotiation good! [*e.g.*, interactive procedures with the system have been successful--documents are provided].
- System provides the user with information helpful to his work.
- System provides the user with "documents that he is glad to get."

Although these statements are simplistic, they accurately characterize current thinking in IS&R system evaluation. Most measures and relevance j gments either have their origin in, or ultimately reduce to, a measure having one of these three "meanings".

It should be clear that these three statements are procedurally, or operationally, oriented--the problem of information need is not directly addressed. I choose, rather, to replace these general statements with a series of questions directed toward basic issues, the answers to which will yield a definition of information need:

- Why is the user seeking information?
- What creates a need for information?
- How is the process of inquiry related to information need?
- What is meant by information?

7.2 Observation and Measurement

Evaulation and relevance are assumed to be basic concepts essential in studying the interaction of man with his environment.

Man lives in a world of interaction and communication. Thus, we may accept as a basic premise that men and/or systems that exist in total isolation from their environments are without interest. Indeed, it may be argued that systems* of any kind exist within environments and must interact with them. An observer of such interactions may record data that are transferred to or from the system and in that sense, may be said to observe actions taken or caused by the system. But in what sense does the observer actually *observe* these interactions?

We note that a system tends to maintain *equilibrium*** with its environment. Now, if a system remains in equilibrium with its environment, then a new and uninitiated observer, who is told to "observe" the system for the first time, will have great difficulty in separating it from its environment. The point of this argument is that *disequilibrium is essential to the observation of interaction.* That is, if, for some reason, a system is in disequilibrium with its environment, then it must effect potentially observable change to re-establish equilibrium. In effect, the change from disequilibrium to equilibrium corresponds to a system's entropy reduction operation, that is, the entropy of an effective system must be lower than the entropy of its environment. Since systems in which we are interested are presumed to be finite and to contain a finite number of states, the process of equilibration must involve changes in a finite number of states. These changes in state are what the observer is *privileged* to *observe*. Consensus about the reality of any such observations must depend, in the final analysis,

---

* A *system* is defined as that portion of the universe chosen for observation.

** *Equilibrium* is used analogously to its use in Thermodynamics.

upon the acceptability to others of methods by which matters of truth, principle, or any other justifiable grounds for shared belief are made evident. Under a fundamental rule by which Western scientists make the empirical world evident to each other, nothing exists that does not exist in some amount and is not, therefore, measurable. Granted this additional premise of shared belief about the world of science, then the *ability* of the observer to *perceive* a change of state depends on the precision of the measuring device(s) available to him.

While measurement in these terms involves observation, it must also involve an operation--*i.e.*, the assignment of a value to what was perceived. So grounded, each uniquely perceptible element of an observation may be assigned a unique number. Measurement, viewed in this way, becomes analogous to the action of a *random variable*. A random variable effects a one-to-one mapping between the event space (all possible states of the system) and the real line. Thus, a random variable is a measuring device (*cf.* Def. 2.1, Chapter IV).

In principle, the observer can observe any changes of state of the system under scrutiny (moments of equilibrium being infrequent). By this argument, he should be able eventually to develop a probability function that assigns to each measured observation a probability of occurrence. The probability function that is developed is the observer's *subjective estimation of probability*, since the observer is usually unable to observe the system for the extremely long periods of time required by objective probability.

7.3 Interpretation and Extension of Experience

To facilitate the previous discussion, it was assumed that the system and the observer were separate entities. Let us now consider the system and

the observer to be the same, so that attention can be directed to the analysis of the system's experience with its environment. Of primary concern is how a human system (man) goes about creating order out of the apparent chaos that confronts it. Perception, observation, measurement, and the estimation of probability-of-occurrence are, as I see it, the initial factors required in this ordering process. Caws [34] has outlined the subsequent steps that man must take:

- Step from a specific experience to correlation with prior experience.
- Step from prior experience to knowledge of one's own particular world.
- Step from knowledge of one's own world to knowledge of a world shared with other men.

Knowledge is defined by Caws as "the ability to make true statements and defend them as true" [35]. Thus, the statement: "system X contains documents giving the boiling point of water at 10,000 feet of altitude" is not knowledge since it has not been defended as true--*i.e.*, there is no indication that the system has been searched to find at least one of these documents. A statement indicating knowledge of the system would be similar to: "system X contains documents giving the boiling point of water at 10,000 feet of altitude, and document # 973 contains those data."

The third step is probably the most important because it gives rise to common experience. The validation of common experience is the first step in scientific activity, and the specialized inductive inference exhibited in the first two steps is manifested in the inductive inference of *scientific method*. Science assumes that a logical progression of inquiry will yield answers that asymptotically converge to an explanation of an "unknown".

Scientific activity and scientific method, represented by a presumed correspondence between empirical propositions and general theoretical propositions, are a logical extension or, perhaps mirror of basic human activity.

Theory, as a by-product of scientific method, may be defined as a logical, probabilistic structure that is created from empirical propositions. But empirical propositions are the result of measurement which is, in turn, dependent upon a defensible identification of observables. Questions about theory (structure), thus defined, ultimately reduce to questions about observables, which are the elements of experience.

7.4 Science as the Generation of Hypotheses

Inquiry, as I have described it, results from observation and measurement. An alternate way of viewing the progression of inquiry is as a process of hypothesis generation and testing. A hypothesis about observables serves as a formal representation of the observer's subjective estimation of the probability of an event (or group of events). We shall define a hypothesis, then, as any verifiable proposition that is not itself an observational statement. Consequently, the decision problem associated with adducing support for or rejecting a hypothesis amounts to finding a method or algorithm for discovering whether a well-formed statement (a meta-observational formula) is refutable.

Prior to the formulation of a hypothesis, an *observational sentence* and an *empirical generalization* must have been stated (or else must be susecptible of construction on demand). An observational sentence is defined as a sentence of observational terms joined by grammatical and logical connectives; an empirical generalization is of the form: all X's are Y's.

A *hypothesis* may, therefore, be defined as follows: "a sentence which has as a consequence at least one empirical generalization, but whose contradictory does not have the form of a protocol [an observational] sentence" [36]. The following are examples which illustrate the meaning of these three terms:

Observational Sentence:	This retrieved document from system X satisfies my information need.
Empirical Generalization:	All retrieved documents from system X satisfy my information need.
Hypothesis:	A retrieved document from system X will satisfy my information need.

The cyclical process characteristic of the scientific method can now be modelled as indicated in Figure 7.4.1. The closing of the cycle is provided by the observables that result from the testing of hypotheses.

The *verifiability theory of meaning* [37], popularized by the Vienna Circle of the 1920's and 1930's, contended that a sentence was empirically meaningful only if it was verifiable. Thus, a sentence (*e.g.*, a hypothesis) is relevant to some thing only if it asserts or denies something with respect to it. It is easy to infer that the relevance of a sentence to and knowledge about a thing are closely tied, if not identical. A sentence about a thing is irrelevant if knowledge of it cannot be obtained.

The observational cycle is now complete. Observations of one's environment constitute measurement; measurement permits the formulation of empirical propositions (observational sentences) and these propositions yield hypotheses which, when tested, yield new observables or knowledge (or both).

7.5 Information Acquisition Through Hypothesis Testing

The observation/measurement/hypothesis-testing cycle also has an informational counterpart or explanation. Let us assume that the system under

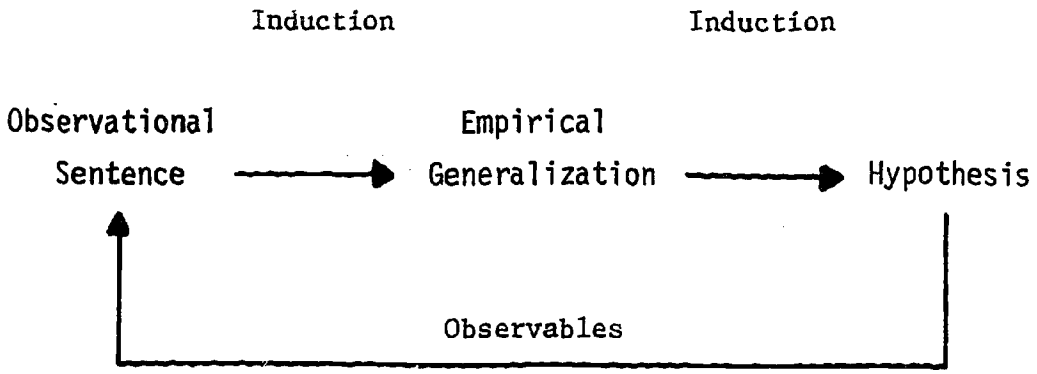


Figure 7.4.1: The Scientific Method.

observation can be in any of n possible states $\beta = \{E_1,\ldots,E_n\}$. The observer observes the system because he is uncertain about the disposition of its component states (*e.g.*, which states are in existence). Prior to his interaction with a system,* the observer must allow for the existence of any one of n different states (n could be infinite). The observation of a change in the state of the system effects a partitioning of the initial states into those states which are observed to exist, and those states that have not yet been observed. Expressed differently, the reception of data by a system yields a reduction in the size of the set of *a priori* alternative states of the environment:

$$|\beta| > |\beta'|$$

This reduction in the number of alternatives, or increase in certainty about the entity under observation, *is* information in the classical Maxwell/Boltzmann sense.

Thus, information is a reduction** of ignorance through an n-fold polychotomy of β. Or,

$$\text{Ignorance } (\beta') < \text{Ignorance } (\beta)$$

Information acquisition also amounts to the observer's certainty about the existence of any given state. This certainty is manifested in the subjective estimation of the probability of a state E_i, $P_{sub}(E_i)$. Given a datum d_ν (*i.e.*, one which occurs during the νth observation of a state) the credibility

---

* We have already pointed out that interaction is effected through a change of state which yields observables.

** This reduction in ignorance can only be effected through an expenditure of energy or negentropy (N). We assume, following Brillouin [38], that $\Delta I - \Delta N \leq 0$.

of the hypothesis that event E_i will occur is measured by $P_{sub}(E_i)$. This credibility is updated through Bayes' theorem:

$$P^{(\nu)}(E_i) = \frac{P^{(\nu-1)}(E_i)\ P(d_\nu|E_i)}{\sum_k P^{(\nu-1)}(E_k)\ P(d_\nu|E_k)}$$

Watanabe [39] defines an *inductive entropy* $U^{(\nu)}$ which is a measure of the uncertainty of the validity of the probability hypothesis:

$$U^{(\nu)} = -\sum_{i=1}^{n} P^{(\nu)}(E_i) \log P^{(\nu)}(E_i)$$

His *inverse H theorem* states that $U^{(\nu+1)} \leq U^{(\nu)}$, or that the average uncertainty of the probability of a given state monotonically decreases with each new observation. This reduction of uncertainty may be defined as generalized learning.

The continuous decrease of U and the convergence of $P_{sub}(E_i)$* provides support for the observer's hypothesis that $P(E_i)$ is the true value. If $P(E_i)$ does not converge, then the observer must adopt a new estimation $P'(E_i)$ of the probability of occurrence of state E_i. In effect, he must formulate a new hypothesis in the face of conflicting data.

The observer's information processing and hypothesis testing activities may be explained as an attempt to reduce the undertainty about, and the number of attributes (states) of, the system under observation which must be processed. Knowingly or unknowingly, that is, the observer attempts to eliminate redundancy through "information" reduction**. As will be shown, efficient information processing is assumed to be a prerequisite to the

---

* The convergence of P_{sub} means that $P^{(\nu+1)}(E_i) - P^{(\nu)}(E_i) < \varepsilon$, where ε is small.

** This is analogous to Posner's [40] information reduction, conservation and transformation.

observer's cognitive and recognitive tasks.

7.6 Hypotheses and Hypothesis Testing

To this point we have been concerned with the role of hypothesis testing in observation, information processing, inquiry, induction and in the estimation of probability. But we have not investigated how hypothesis formulation and testing forms an integral part of man's problem solving behavior. Let us state the fundamental postulate of this section, and, indeed, of generalized information need, namely that *all action and thought are based on the testing of hypotheses.** This is a bold assumption, but it will be shown to be instrumental in the understanding of "information need."

Minsky gives a first clue to the nature of an observer's *model* of the environment:

> The problem solving abilities of a highly intelligent person lies [sic] partially in his superior heuristics for managing his knowledge structure and partially in the structure itself. These are probably somewhat inseparable. In any case, there is no reason to suppose that you can be intelligent except through the use of an adequate, particular knowledge or model structure. [41]

> A man's model of the world is a distinctly bipartite structure: one part is concerned with matters of mechanical, geometrical, physical character, while the other is associated with things like goals, meaning, social matters. [42]

He assumes that a model of the world is based on an individual knowledge structure, or *cognitive structure*. Minsky finds it convenient to partition this structure into two parts, each part (or, perhaps, each process) dealing

---

* One may make a case for a distinction between purposeful and non-purposeful (involuntary) action. While such a distinction may ultimately be helpful, it will not be pursued here.

with a different aspect of the environment. The *physical* processing part of the structure deals with the raw data inputs from the environment, while the *goal-directed* part is man's reaction to, and interaction with, the environment. The former process we choose to call the *concurrent hypothesis structure* and the latter the *actively hypothesizing structure*. Thus, there are postulated to exist two distinct levels of hypothesis testing within man's model of the world.

7.6.1 Concurrent Hypotheses About the Perceived World

Information that the observer obtains from the environment (*e.g.*, the polychotomy of β, see Section 7.5) permits him to estimate the probability of occurrence of any observed state. This perceptual information we assume to be manifested in the creation of "constancy" hypotheses about the states of the environment. These hypotheses are either the observer's estimation of probability or, if $P_{sub}(E_i) = 1$, his observation of a continuously occupied state. Since no action is required of the observer in these cases, that is, since he is in *equilibrium* with the environment, with respect to the states in question, these "constancy" hypotheses are relegated to non-attentive processing. In other words, as long as sensed data appear to support the hypotheses, no action, or conscious attention, is required of the observer.

The reception (acquisition) of negative data elements, or data elements that do not support one of the set of concurrent hypotheses, throws this set of hypotheses out of equilibrium. If this is the case, the observer must either change the probability associated with the hypothesis, formulate an alternative hypothesis, or both. In any event, *information need* is an expression of the observer's need to acquire more data, to create new

hypotheses and to observe the environment with renewed attention. This view is analogous to Harmon's [43] interpretation of information need: "Information needs might be viewed as products of change within a system of personal constructs."

7.6.2 Active Hypothesis Testing

Human intellectual activity, according to my argument, reduces to the active creation and subsequent testing of hypotheses. I assume hypothesis testing to be a central mechanism for updating cognitive structure. The terms "cognition" and "structure" are employed to suggest the mental processes of data acquisition and ordering.

Cognitive structure is viewed as an ordered collection of *data elements* and of *relations* between data elements. For purposes of analogy, cognitive structure is taken to be similar to Quillian's [44] data structure where nodes are words and linkage indicates a relationship between words. (A similar idea was earlier proposed by Bernier [45]). Although information is assumed to be obtained from the partitioning of the event space, information is also postulated to exist as a context-sensitive structure (see Ernst and Yovits [46]) that represents both an imposition of order upon things "known" to exist and an order of observation of data from the environment. One part of cognitive structure is thus assumed to be a representation of what the observer has observed in the environment. We may conclude that cognitive structure is an observational *theoretical index* of the perceived environment.

In addition to the indexing function, another function of the cognitive structure is presumed to provide a site for active hypothesis testing. This hypothesis testing takes the form of assumptions about the existence of as yet unobserved data elements and of relations between them. Two cases for

hypothesis generation arise:

case 1: A relational structure exists but the value of a data element locus is unknown. A hypothesis is formed about the existence of such a data element, and the environment is observed--this is inferred to be a manifestation of *information need.*

case 2: A hypothesis is formed about the existence of a relation between data elements, or about the existence of a specific, unobserved data element in relation to a known data element--this is also inferred to be a manifestation of *information need.*

In both cases, information need is interpreted to be the expression of a *need to provide support for* a hypothesis. Negative support will create a disequilibrium in the existing collection of hypotheses and will require continued data acquisition and new hypothesis formulation on the part of the observer. Positive support *may* lead to an end-state, or *goal*, of the ongoing process of inquiry.

In both the above cases, we assume an observer and decision maker who acts "rationally" according to our model. However, experience indicates that quite often "irrationality" (or alternative "rationality") prevails--*e.g.*, the acceptance of a hypothesis is based on a definition of the situation that our model does not prescribe. A "favored" hypothesis persists, despite what we should expect, as a complex union (or intersection) of simple hypotheses. One explanation of this phenomenon is that credibility of the favored hypothesis is achieved through a form of transitive logic over the sub-hypotheses. For example, hypotheses A,B, and C have been observed to be supported; hence, D, the favored hypothesis, a compound of hypotheses A, B, and C, is also assumed to be true. The "irrational" processor, who thus fails to act as a perfect information processor in our view, will continue

to accept hypothesis D* in the light of what we "know" to be conflicting data--usually until some of the component hypotheses are demonstrated to be false. "Error" of this type is viewed as a temporary deviation from intellectual hypothesis testing as our model prescribes for it.

7.7 Some Information about "Information Need"

We are now ready to answer the four questions posed in Section 7.1. The answers to these questions provide a convenient summary of the material presented in the preceeding sections.

- Why is the user seeking information?

  The user (or IS&R system observer) seeks information for the testing of hypotheses that he has about the data elements contained in the system.

- What creates an information need?

  Information need is created by either a) hypothesis disequilibrium, or b) active intellectual hypothesis testing.

- How is the procedure of inquiry related to information need?

  The process of inquiry is the scientific method, a cyclic progression through observation and measurement, generalization and hypothesis testing. Problem solving behavior is effected through hypothesis testing.

- What is meant by information?

  Information is defined as the reduction of uncertainty derived through partitioning of the event space. Information is also defined as acquisition of data suitable for hypothesis testing--*e.g.*, data of value in decision making.

---

* That is, if A: System X contains data on metling points,
B: System X contains data on titanuim compounds,
then =≠> System X contains data on melting points of titanuim compounds.

The following sections will deal with the role of information need and hypothesis testing in human behavior, with a formal model of hypothesis testing in information retrieval, and, finally, with a reconsideration of the concept of relevance.

8. Problem-Solving and Decision-Making Behavior

8.1 Introduction

In the initial sections of this chapter, it was pointed out that a thorough understanding of the often used phrase "IS&R Systems Evaluation" demanded a prior and careful consideration of the term *evaluation*. It was postulated that evaluation was a judgmental, or correlational, relation between retrieved data and the user's information goal. Although *retrieved data* is quantifiable and is amenable to analysis, an *informational goal* is recognized to be a qualitative, subjective concept. The argument was presented that progress toward quantification of an informational goal could be achieved through a detailed analysis of the concept of the user's *information need*.

Information need is characterized as the impetus to a *process of inquiry* involving measurement, observation, information acquisition, the investigation cycle, hypothesis testing and decision making. Although the theoretical discussions that have been presented appear to be consistent, experimental testing of the derived hypotheses is required. The sections that immediately follow point toward such testing by means of an analysis of both models and behavorial investigations of human problem solving.

8.2 Problem Solving as Inquiry

The presentation of the scientific method, as outlined in Section 7.4, involved the description of a progressive shift from an individual's personal

experience to experience that can be generalized. This shift, or transition, is the essence of Caws' three-step model, which terminates with the acquisition of *knowledge* about an observer's environment. The cyclic process depicted in Figure 7.4.1, which is by itself the formal definition of a hypothesis, is also a description of the transition from individual experience to shared knowledge. The feedback of observables from the testing of a hypothesis closes the investigative loop of the scientific method and implies that the scientific method, as a hypothesis testing cycle, is an open ended process. This posited hypothesis-testing cycle also provides a convenient representation of problem-solving behavior.

Recall that observation and measurement are effected through the perception of disequilibria in the observer's environment. It follows that the hypothesis-testing cycle represents a progression from an observer's initial response to disequilibrium, through a series of if-then relationships (hypotheses) to a solution. This form of problem solving was recognized by Dewey in his principle of *continuum of enquiry*: "The conclusions reached in one inquiry become means, material and procedural, of carrying on further inquiries" [47]. The idea is often overlooked, however, that the feedback of observables itself may contribute to a structuring or patterning of the inquiry. Such patterning helps to account for the existence of the relationship between the object of the investigation and the manner of the inquiry. This is recognized in Russell's "structural postulate" of Scientific Inference [48] or in Whitehead's "grouping of occasions" [49]. This form of inquiry and behavior is also identifiable in the vocabulary of some psychologists as a *Gestalt*.

In the Gestalt view behavior is a pattern or configuration of action and the linkage of ideas associated with problem solving is progressive inquiry. With such an emphasis on the organization of ideas, it is believed that "psychological organization" tends to move toward the state of Prägnanz--*i.e.*, toward the *good Gestalt*.* This view of things is consistent with the idea of a continuum of enquiry as an entropy-reducing operation (see Section 7).

Trace theory, often associated with Gestalt formulations posits a stochastic representation of the subject's past in the characterization of his present. Information gained from the test of a current hypothesis is assumed to be mediated by information obtained from previous hypotheses. This is one elaboration upon the idea that what is information (and, by implication, meaning) to a user is highly context sensitive.

Achievement of a "good" Gestalt implies that problem solving behavior is directed toward an end situation which brings closure with it. The person's desire for completion of a task emphasizes the importance of his acting as if inquiry had an attainable goal. According to Dewey: "The nature of the problem fixes the end of thought, and the end controls the process of thinking" [50]. Although I do not propose to view hypothesis-testing and problem-solving behavior as *Gestalten*, this idea of things is useful in placing emphasis on the organization and structure of inquiry, the utility of acquired information, and the relevance of goal definition to a sense of closure when an end state has been achieved.

---

* Characterized by the laws of similarity, proximity, closure and continuity.

8.3 Problem Solving Models

The first modern model of progressive inquiry in problem solving was developed by Dewey. He believed that a problem should not be characterized as a crisis or by the obtained solution, but, rather, as an inquiry sequence. This view is reflected in Dewey's five problem-solving steps:

- a difficulty is sensed,
- the difficulty is located and defined,
- possible solutions are suggested,
- consequences are considered, and
- a solution is accepted.

Notice that *information need*, hypothesis testing and decision making are implicit in all five steps.

Recently Guilford [51] has presented an extension of Dewey's conception of problem-solving. This model, depicted in modified form in Figure 8.3.1, is a process model and exhibits the sequential ordering of data inputs, attention, cognition, production, evaluation and memory. The initial input and attention operation corresponds to Dewey's first step; the first cognitive* operation corresponds to his second step; the production of a tentative answer corresponds to the third step; the second cognitive operation is analogous to the fourth step; and, the final production may be equated with Dewey's fifth step. It is important to realize that the final production (and adopted solution) may only be achieved after several cognitive/productive iterations.

---

* Guilford [52] defines cognition as awareness, immediate discovery or rediscovery, or recognition of information in various forms.

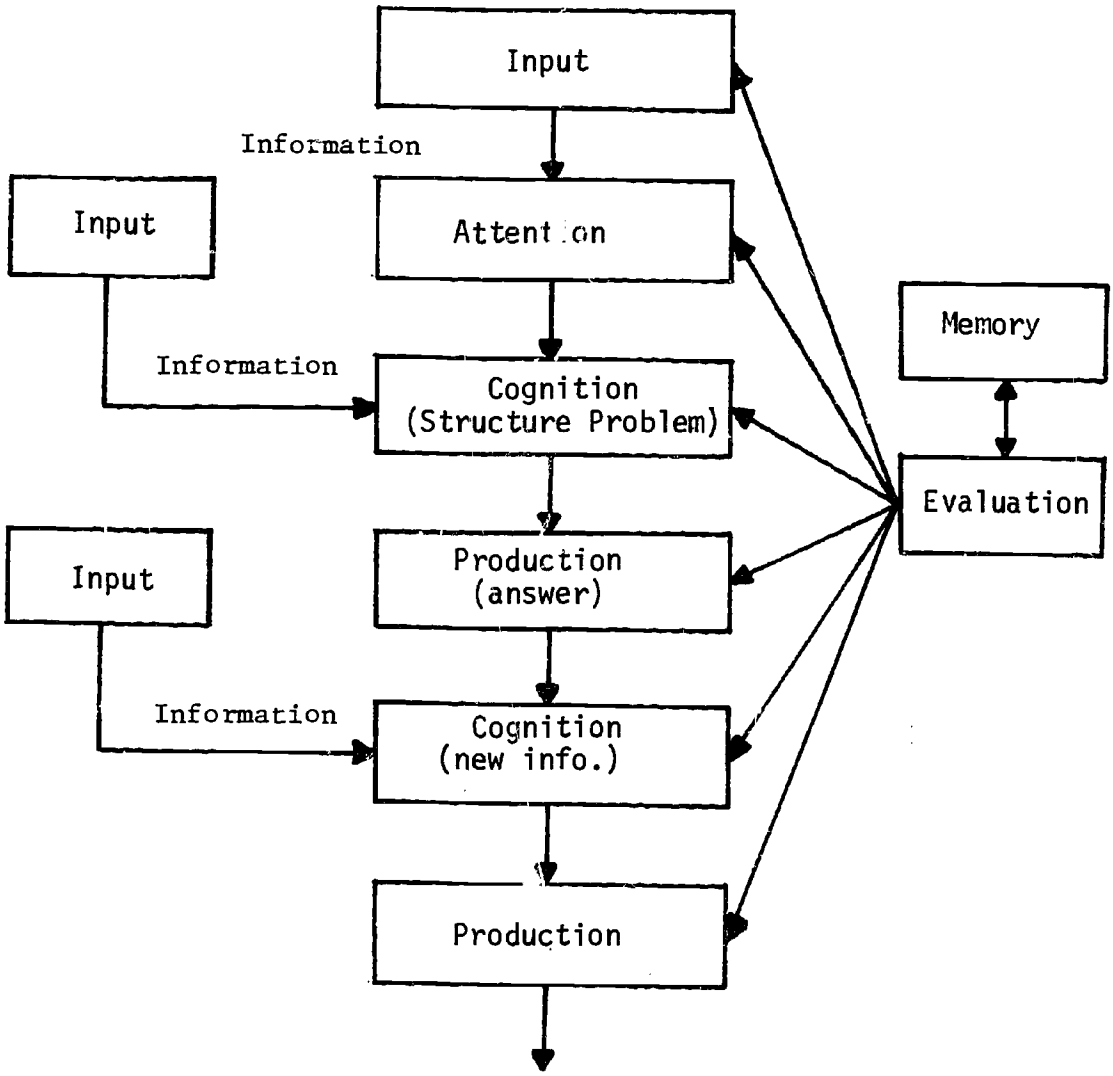


Figure 8.3.1: A Problem-Solving Model.

The initial attention operation (in conjunction with the evaluation/ memory operation) corresponds to what we have chosen to call the hypothesis disequilibrium, while the several cognitive operations and the evaluation operation correspond to active hypothesis testing. If we conceptually separate these two forms of hypothesis testing, and remember that each one is tied to information need, Guilford's model can be reduced to the three stage model depicted in Figure 8.3.2. The recursive nature of this model, and its application to the analysis of Information Storage and Retrieval problem solving, will be considered later in this chapter. Attention is now directed to the analysis of hypothesis-testing behavior.

Investigation of problem-solving behavior is difficult because the central postulates remain, as yet, untried. Bourne [53] provides us with a hint of this problem:

> We can only infer that a decision-making process exists in problem-solving - [we] search for something only rumored to exist, but so indescribable that we cannot even tell when [that something] occurs.

One of the central postulates of the hypothesis or, as they are sometimes called, process theories is that in a problematic situation, a person (subject) entertains at least one hypothesis. Thus the stimuli that the subject receives (be they controlled or random environmental inputs) provide a test of the hypothesis(es) under consideration. Observed data, following validation with respect to the hypothesis leads to acceptance, rejection, or revision of the hypothesis.

It should be clear that active hypothesis testing, especially as depicted in Figure 8.3.2, involves input both from the exemplar (stimulus) and from the subject's environment. This input provides the data, subsequently

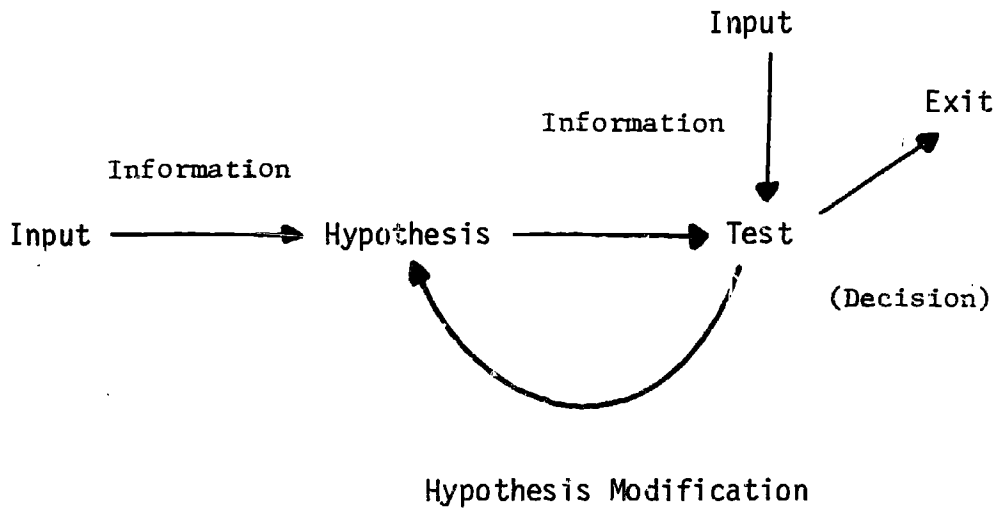


Figure 8.3.2: A Three-Stage Problem-Solving Model.

information, recognized by the problem solver. Hypothesis-testing behavior thus defined is an attempt to reduce disorder that is sensed in the problem, through a reduction in the number of degrees of freedom associated with the problem. It is inferred that the information that is obtained, either as input or as the result of hypothesis testing, serves to partition the set of alternative hypotheses and, consequently, to provide further information about the problem. This partitioning results in the information gain discussed in Section 7.5.

The number of hypotheses that are adopted (*i.e.*, the number of iterations through the model of Figure 8.3.2) is assumed to be a function of the number of attributes associated with the problem. This view of problem solving implies that the problem solver is able to identify correctly the pertinent attributes of the problem. In hypothesis testing, the subject's major freedom comes from the rich domain of hypotheses from which he can choose. However, the problem usually structures the order of occurrence of the instances (*e.g.*, data inputs) encountered. As we shall see, efficient problem solving demands not only a person's choice of hypotheses but his complete identification of the attributes involved. Experimental data on attribute identification and information processing in problem solving will be useful for the characterization of IS&R retrieval operations.

8.4 Attribute Identification and Problem-Solving Strategies

In the preceeding sections, we have emphasized the importance of the act of inquiry in a problem solving situation. Several models have been presented that define inquiry as a multi-step process. Common to all of the steps involved in these processes is the process of hypothesis testing. Thus, all data and information that the problem solver directly encounters

are presumed to involve either implicit or explicit hypothesis testing. As has been previously emphasized, both forms of hypothesis testing are dependent upon an *information need* which creates a demand for new data, a new hypothesis, and so on...

How the problem solver handles and processes newly encountered data depends on several factors. These include the previous data encountered, hypotheses that have been tested, and information that has been gained. This means that the interpretation of new instances is a function of previous conceptualizations. One aspect of *interpretation* is posited to be the learning of a *rule* that can be applied to the analysis of successive data inputs. The other aspect of *interpretation* is assumed to be an ability to identify the attributes, or variables, that characterize the problem at hand. The Bruner, Goodnow and Austin [54] definition of *attribute* is adopted for this discussion: "an attribute is any discriminable feature of an event that is susceptible to some discriminable variation from event to event."

When a problem has been solved (or a concept attained, in the Bruner sense) we shall say that the subject has identified those *attributes* and *rules* which enable him to classify and act upon any future instances encountered which are pertinent to the problem situation. There are many ways in which a problem solver might go about obtaining a solution to his problem. We shall describe these various methods of solution as strategies, *i.e.*, a sequence of decisions involving the acquisition and utilization of information for the achievement of a well-defined *goal*. A strategy is presumed to be adopted for several reasons: to minimize the number of iterations through the model in Figure 8.3.2, to minimize the subject's

"information overload", and to minimize error in decision making. Strategies may be evaluated by their efficiency in eliminating alternative hypotheses concerning the attributes which are pertinent to the solution of the problem. Within these constraints, the number of iterations required to reach a solution is dependent on the number of attributes that must be correctly identified.

Bruner, *et al.*, [55] have identified four basic problem solving strategies:

- Simultaneous scanning
- Successive scanning
- Conservative focusing
- Focus gambling

Briefly, *simultaneous scanning* may be defined as the simultaneous testing of several hypotheses about attribute importance; *successive scanning* is the testing of a new hypothesis with each successive instance encountered; *conservative focusing* is the orderly testing of attributes, and involves the use of only one attribute as the independent variable for each hypothesis tested (this strategy is, on the average, an optimal strategy); *focus gambling* reduces to the adoption of a "favored hypothesis", as described in Section 7.6.2.

In the following section we will consider the nature and importance of attributes in an Information Storage and Retrieval environment. Attention will be directed toward the identification of the problem solving steps in the retrieval/search interface as a prelude to the description of an extension of the classical Bruner conjunctive-concept experiment.

8.5 Attributes and the Retrieval Interface

The retrieval interface of an IS&R system presents a unique problem solving situation. The searcher (system user) has as his predefined goal the extraction of a subset of the data elements in the system data base in order to satisfy an information need. The searcher brings, initially, only two capabilities to the task: 1) his own cognitive structure and reasoning ability; 2) his perception of the search interface. These two features are presumed to determine the particular behavior pattern that the searcher will display. However, as Reitman [56] cautions, the observed behavior (in terms of strategy and results) will be less than ideal:

> In real conflict and cooperation problems, the conditions for game theoretic solutions are rarely met. We know neither the full set of alternatives, the states of the world, nor our opponent's (e.g., the IS&R system's) perceptions and evaluation of them. Bluff, deceit, and efforts to influence and persuade are possible because and only because this is so. If the relevant facts about the world, the alternatives, and the payoffs were known, there would be nothing to deceive or persuade about.

In such a situation, the controlling variable appears to be the precision of measurement implicit in the user's hypothesis (cognitive) structure - *e.g.*, the precision of measurement associated with the definition of the component *data elements*. This precision is dependent on both previous experience and on general knowledge about the subject area encompassing the information need. Saracevic, in his relevance measurement studies, has shown that the more desperate a user is for information, the more relevant everything becomes for him [57] - *e.g.*, the user has a low precision of measurement demanded by with his data element definition. This type of observation places emphasis on the value of the user's *a priori* subject knowledge (what the user

brings to the search interface); indeed: "The less we know, the more everything becomes relevant, and the more we know the more stringent we become in our judgment [58]."

In addition to a lack of subject knowledge (which is an expression of his information need), the searcher has a less-than-perfect knowledge of the system's *index space*. This means that a searcher may not understand the system's operating characteristics in terms of the *transmission decoding*, *language* and *vocabulary* variables. By example, and considering only the language variable, a searcher may be able to find data elements by reference to a specific transmission decoding element (*i.e.*, a subject heading), but he is not knowledgeable of the set of relations (*i.e.*, the search language) available for modification and direction of a search.

Although we have discussed some of the attributes essential to efficient IS&R problem solving, all attributes can be grouped into four broad classes:

- The elements of the index space
- The number of access points to the data
- The rules of the cross-reference language
- The range of data elements (and associated documents) that are available

It is postulated that effective search requires that the user obtain, as soon as possible, knowledge and mastery of these attributes.

The problem-solving steps involved in the search interface are assumed to be essentially those described in Figure 8.3.2. Satisfaction of the searcher's information need is termed the goal of the problem-solving activity; his query, as presented to the system is actually a hypothesis

about the nature of the data elements in the system. The data elements that are retrieved serve both to test this hypothesis and to permit the searcher to decide whether his goal has been achieved. The retrieval and testing of data elements may serve as a basis for modification of the user's hypothesis structure. This modification may be inferred by an observer from a new query by the searcher; it may also be inferred that he has acquired a different expectation of the types of data elements to be retrieved. Saracevic has confirmed the existence of this from of behavior [59]: "Items judged as non-relevant tend to remain as such; items judged as relevant are subject to change following iterations with the system."

Search feedback is essential to the solution of any retrieval problem and to the satisfaction of a searcher's information need. Feedback enables a user to obtain information about the system. This information takes the form of data that enable him to decide which system attributes and user hypotheses will be effective for the achievement of his goal. Frequently, information obtained will create, through a modification of the user's hypothesis structure, an alteration of his information need. However, it is not clear just how effectively the searcher can process feedback information. Some evidence is needed for how efficiently human problem solvers can identify and utilize attributes in the solution of a problem. A "relevant" experimental investigation is described in the next section.

8.6 Experimental Investigation of Attribute Processing

The subject in the Bruner Concept Attainment Experiment [60] is tested for his ability to achieve a fixed *conjunctive concept*. A conjunctive concept is defined as: "The joint presence of the appropriate value of several

attributes." The simplest case would be when the concept was composed of only one attribute--*e.g.*, apples, squares, etc. A variation of this basic experiment could center about the achievement of a fixed *disjunctive concept*--*e.g.*, attribute(1) *or* attribute(2) *or*, ...

As depicted in Figure 8.6.1, the subject is provided with an array of *instances* (data elements characterized in terms of *attributes* and *attribute values*) to be tested in order to attain a concept. With each instance encountered or identified, the subject must decide whether the instance is an example of the concept sought. A brief examination of Figure 8.6.1 will show that the subject is presented an ordered array of 81 instances constructed from four attributes (border, color, number, shape) each of which may take on three different values.* After an initial exemplar of the concept has been presented by the experimenter, the subject is instructed to use a question-answer technique in an effort to discover the chosen concept.

It should be noted that in the Bruner studies the subject is given the array of instances as a problem-solving aid. Because the array is systematically ordered by attribute values, the subject is more likely to perceive the set of attributes involved than when the instance array is not so well ordered. Finally, the subject is informed both of the definition of a conjunctive concept and of the procedural rules of the experiment.

We have previously noted that the studies of Bruner, *et.al.*, have identified four general problem-solving strategies. The usually optimal strategy (conservative focusing) relies on the subject's systematic variation of one attribute while holding the remaining three constant. In such a case,

---

* 1,2, or 3 borders; red, green, black; 1,2, or 3 objects; square, circle, cross.

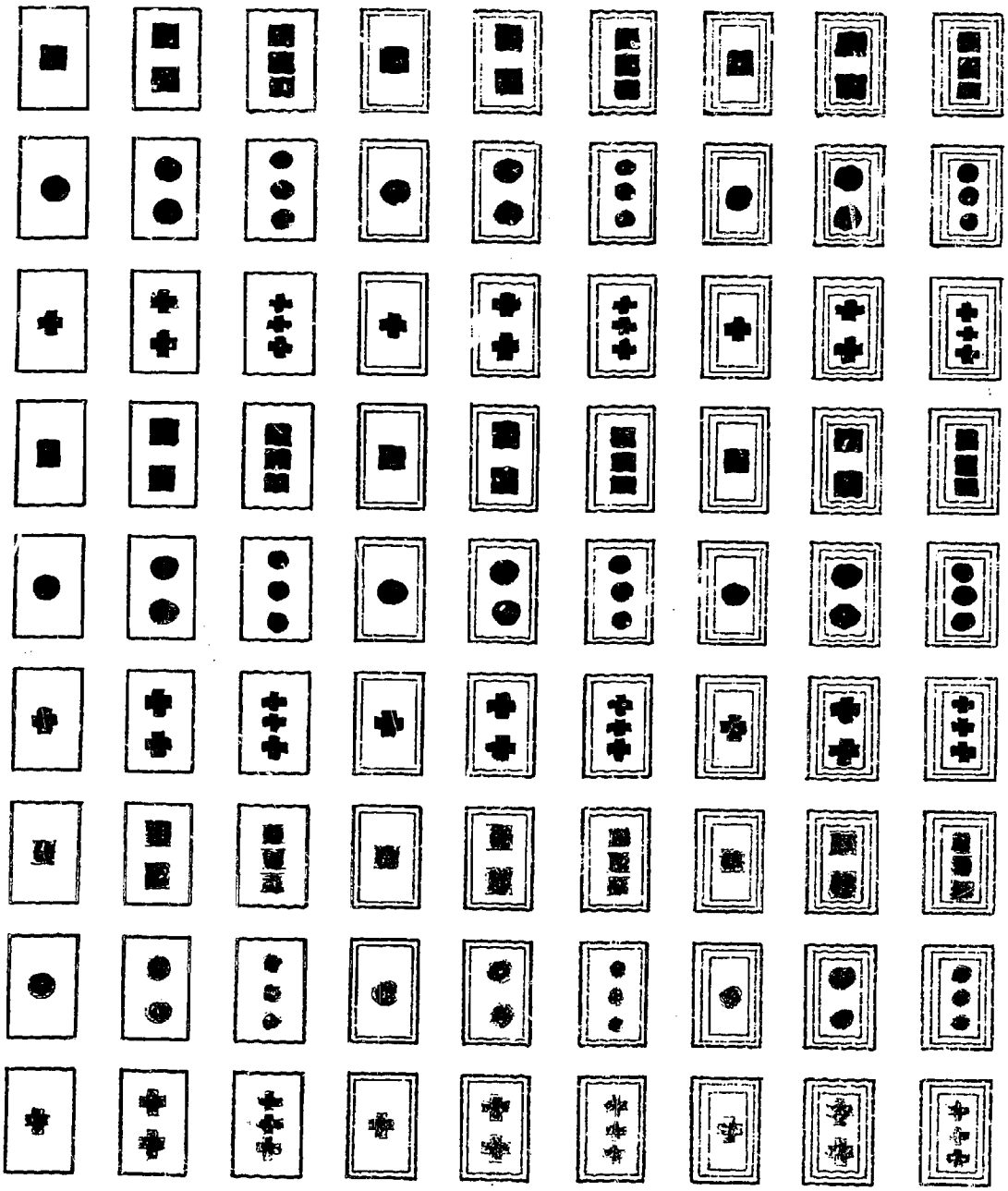


Figure 8.6.1: The Bruner Instance Array.

the information received for testing successive hypotheses about which attribute is the concept is maximal since it permits a minimal partitioning of the space of instances. Four general results have been obtained from these information processing studies [61]; they are listed below.

- Strategies can be described both in terms of goal and by the problem-solving steps.
- In the absence of new information the subject will fall back to the testing of previously useful cues.
- Subjects may fail to use information arising out of negative instances or indirect tests.
- Subjects frequently fail to assimilate as much information as is potentially available from the testing of an instance.

Unfortunately, the real-world is seldom structured in the way problem situations are structured in the Bruner experiment. Very seldom does one possess complete knowledge of the collection of attributes involved in a problem-solving task with which he is confronted. Thus, an initial step in a systematic solution (and prior to the adoption of an ideal strategy) of a problem is the identification of the set of variables or attributes involved in the problem. It is believed that whether the problem is the identification of a concept, the location of a book in a library, or the retrieval of data from an IS&R system, the problem-solving steps are essentially the same. Thus, one of my goals has been to see if the general conclusions about problem-solving behavior of subjects in the Bruner, *et.al.*, studies (see above) arise when a subject is supplied an unstructured task, characterized by the attainment of a conjunctive concept as goal, together with a randomized Bruner instance array (with each instance separated from the others, and a minimum of procedural information. Given an unstructured task (which is not

dissimilar in nature to that of an IS&R system interactive interface) I want to know whether a subject is able to 1) perceive the attributes involved, 2) attain the concept, 3) utilize a recognizable strategy, and 4) perceive the information content of the "informative displays" supplied to him during the course of the (attempted) task achievement.

8.6.1 The Extended Bruner Experiment

I have previously emphasized that, in everyday life, one rarely has complete knowledge of the attributes associated with a given problem. Consequently, to solve a problem, one must first identify its pertinent attributes. Early and accurate attribute identification is prerequisite to obtaining an efficient solution to any IS&R problem. The purpose of the experiment described below was to determine if potential problem solvers (Ss) could identify the attributes associated with a problem and use them to achieve problem solution. This experiment is based upon several of the central assumptions of Hypothesis Theory [62]. Six governing assumptions are:

- Ease of problem solution is directly related to the degree of structure (*i.e.*, number of attributes, lack of ambiguity, and clarity of attribute presentation) associated with the problem.
- In a problem-solving setting, a subject's behavior results from his hypothesis testing when concept attainment is requisite to problem solution.
- A subject (S) enters any situation with some preconceived hypotheses (Hs) about what is to occur and about what behavior he is expected to exhibit.
- S has an initial set of Hs from which he samples until he selects the correct H (or else gives up).
- S modifies his initial hypothesis with increasing experience.
- If the situation is unstructured (*i.e.*, the problem is presented in an ambiguous or unclear fashion) S will attempt to impose a working structure (*i.e.*, impute logical relationships among elements considered to be pertinent attributes) based on his own experience.

The reader will recall that in the Bruner Conjunctive Concept Game, S was provided an ordered array (see Figure 8.6.1) of instances to assist him in solving the problem. The S was given the definition of a conjunctive concept and ample procedural rules for the conduct of the game. The play was started with selection, by E, of an exemplar of the concept which was then presented to S. S would then select (following an initial hypothesis concerning the nature of the concept) an instance from the array and ask E if it contained the concept; E would then answer YES or NO. In the extended Bruner experiment the subject was presented with minimal structure, sparse procedural definitions, and little information as to how he was to behave. Instead of presenting S with an ordered array, each of the 81 instances was placed on a separate card and the resulting deck of cards was randomized. Thus, in the extended Bruner experiment, the exemplar and the array were presented, respect[illegible] card and as a randomized deck.

8.6.1.1 The [illegible]ions and Conduct of the Experiment

In this experiment the S's only introduction to the problem consisted of the following statement read by E:

- I am interested in how people solve problems. I am particularly interested in the processes people use. In fact, I am more concerned with what you do in trying to find the answer, than with whether you are able to find it. I have manufactured a problem for you and assume you have as your goal the solution of this problem.

E then reads and demonstrates the following instructions:

- Here is a card that has some objects on it [E places the exemplar card before S]. I have an object in mind [the concept was "square"

throughout the experiment]. Your job is to identify what I have in mind.

- Here is a deck of cards [E places the deck of cards before S]. Each card in this deck is similar to the card I have shown you. You may use the deck of cards to help you identify and name what I have in mind.
- I will record what you do in solving the problem, but will not help you in any way. Over here, however [E points to his assistant], I have installed a helping machine. The helping machine will answer any question you have with a YES card, a NO card or a BLANK card.*

Finally, E places the instructions beside S and begins to record data. The recorded data included: the question asked by S; the helping machine's response; the number of times S read the instructions; the number of times S sequentially scanned the deck of cards. The experiment was terminated when (1) the S realized he had obtained the solution, (2) the subject announced he had quit attempting to solve the problem or (3) after fifteen minutes of play had elapsed.

8.6.1.2 Results and Discussion

Experimental results obtained from 48 subjects (a mixture of senior-high school, undergraduate and graduate university students) are presented in Table 8.6.1.2.1. Column A shows the overall average number of questions asked by the Ss, the average number of yes, no and blank answers, the average

* Any non-compound question about an attribute of the deck was answered YES or NO; any other question was answered by a BLANK card.

COLUMN A	COLUMN B
Sample Size: 48 Ss	14/48 = 29% Had a Theoretical Solution
Average values	Average values (up to theoretical)
Number of Questions: 20 Number of Yes's : 3 Number of No's : 5 Number of Blanks : 12 Number of times instructions read : 3 Number of times deck sequentially scanned : 2	Number of Questions: 20.5 Number of Yes's : 3 Number of No's : 7 Number of Blanks : 10.5 Number of times instructions read : 2 Number of times deck sequentially scanned : 1.6
Number of Solutions: 10/48 = 21%	Number of Solutions: 4/14 = 28%

Strategy:

Focus Gambling: 6/10
Other (unidentified): 4/10

No Theoretical Solution Achieved: 34/48

Average number of attributes eliminated: 0.6

22 Ss eliminated 0 attributes
5 Ss eliminated 1 attribute
5 Ss eliminated 2 attributes
2 Ss eliminated 3 attributes
--
34

Table 8.6.1.2.1: Results of the Extended Bruner Experiment.

number of times S read the instructions, and the average number of times that the deck was sequentially scanned by S. Twenty-one percent of the Ss actually obtained (realized that they had) the solution to the problem. It should be noted that 60% of the Ss who obtained the solution utilized a focus gambling strategy.

Column B presents data on those Ss (14/48) who exhibited a *theoretical solution*. Theoretical solution is an analytical construct that indicates a point in the S's protocol when, by means of the information potentially available from the questions and answers, he has eliminated all but the correct attribute. The word "theoretical" is used because S did not realize that the solution could be obtained from the available information. It is interesting to note that the average values up to the point of theoretical solution are very close to the average values of the entire experimental group (column A). In other words, those Ss exhibiting a theoretical solution had an above-average total number of questions and answers. Any information acquisition activities beyond the theoretical solution represents redundant data processing on the part of the S. Furthermore, only twenty-eight percent of those Ss exhibiting a theoretical solution actually obtained the correct solution. Finally, the 34 Ss that did not exhibit a theoretical solution, on the average, eliminated less than one attribute.

Although some of the Ss obtained the solution, the majority of the Ss failed to properly identify the necessary attributes. Following de-briefing sessions, it became obvious that although many Ss could name the attributes of the deck, they failed to understand the relationships between what they observed in the deck and the solution of the problem. The failure to relate what they observed in the deck to the instructions given by E for problem

solution, partially accounts for the fact that so few subjects actually attained the solution. It is concluded, that in such an unstructured problem-solving situation where the stimuli appear ambiguous to the S, the concept of strategy has little meaning. With an incomplete understanding of the nature of the attributes involved in the problem, a subject is hard pressed to adopt an "optimal" strategy for the methodical testing and elimination of attributes. Although it was clear that Ss were repeatedly testing hypotheses from some predefined set, their apparent lack of problem structure permitted only a focus gambling strategy.

A subject's failure to identify the pertinent attributes, adopt a useful strategy and attain the concept suggest that in a relatively unstructured task Ss, with few exceptions, were unable to extract all of the potential information contained in the informative displays (questions and answers) of the play. The fact that only four of fourteen Ss who displayed theoretical solutions achieved actual solutions supports this observation. Furthermore, for many Ss, the successive modification of hypotheses seemed to be confused by a recurring uncertainty of the procedural rules of the interaction. This was reflected in the proportionally greater number of blank responses, than of yes-no responses given. Frequently, Ss would fall back to previously confirmed hypotheses (by asking the same questions) in an apparent attempt to validate the consistency of the helping machine, or to explore the conditions of a yes or no response when confronted with a perplexing and non-useful sequence of blank responses.

Although this investigation was not designed to represent completely an IS&R interface, it is believed that the experiment does reflect and highlight some of the important features of information retrieval problem-solving

interaction. The results of this experiment support the hypothesis that effective interaction, information acquisition and processing and subsequent goal achievement, are related to the amount of problem solving structure and aids available to the subject or user. Some subjects exhibit problem solving capabilities even in the most unstructured and ambiguous situations, however good systems design must provide the necessary structure to stimulate "natural" problem-solving capabilities.

9. A Hypothesis Structure Model

9.1 Introduction

Most studies in man/machine interaction assume, as a starting point, that the use of a computer will enhance human creativity by providing significant insights into the solution of the problem under consideration. This form of interaction is often referred to, in biological metaphor, as a symbiosis. With such an optimistic view of man/computer interaction, one would expect the IS&R search interface to reflect these desirable qualities. Unfortunately, experience provides little data to support this expectation.

Sackman's studies of man/computer problem solving [63] have revealed that only 10 percent of the total problem-solving time is spent at the system interface. Users come to the system armed with preconceived assumptions about the disposition of the system's data, and spend just enough time with the system to test their assumptions. New hypotheses are not formulated on-line, but are developed during subsequent periods of isolation from the system. From Sackman's evidence and that obtained in the extended Bruner experiment, it appears that the time spent by a user in direct on-line interaction with a system is best represented by the testing and feedback operations of the problem-solving process discussed earlier.

A considerable portion of this chapter has been devoted to a discussion of the problem-solving process, and I have assumed that this process arises as a consequence of a user's *information need*. *Information need* was defined, in Section 7, as the result of either hypothesis disequilibrium or active hypothesis testing. The consideration of problem-solving behavior, presented in Section 8, has emphasized that, while frequently adopting a strategy, a problem solver is apt to derive considerably less than maximal benefit from the information available to him, even when he adopts an identifiable strategy. This failure may be attributed largely to the subject's lack of perception of the full range of alternatives involved in the problem. Consequently, the design of effective systems must take into account these human behavior patterns. Similarly, the extent to which such behavorial considerations are accounted for in IS&R systems design provides a basis for evaluation of such systems. As will be shown, the necessary measure of value is defined through a formalization of the user's *hypothesis structure*. This formalization is found in a reconsideration of the hypothesis-testing model briefly presented in Section 8 of this chapter.

9.2 The Hypothesis Structure Model

One of the major tenets of this chapter is that, despite the apparent complexity of most IS&R systems, the problem of system performance evaluation reduces to the problem of formulating an adequate description of a single interface--*i.e.*, a man/machine decision-making interface. We shall characterize this user/IS&R-system interface as the communication between two cognitive structures. The reader is referred to Figure 9.2.1 for a conceptualization of this interface (after Carbonell [64]). In this figure,

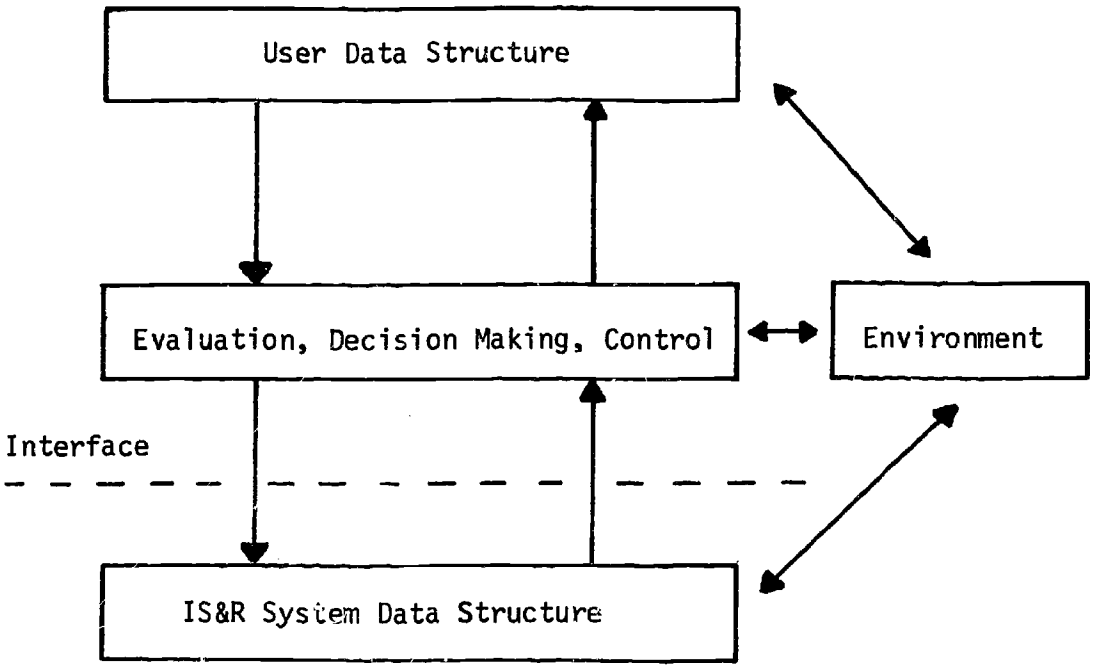


Figure 9.2.1: The Man/Machine Interface.

the user's data structure models the goal (the satisfaction of his *information need*) that he is attempting to achieve, while the IS&R data structure embodies all of the data in the system that may be of value in goal achievement. Figure 9.2.1 also suggests that the user must successively evaluate the system and its outputs, exercise control over the system and, finally, act as a decision-maker concerning the *relevance* of the data to his goal.

Following the discussion in Section 7 of this Chapter, I assume that hypothesis testing, data-element acquisition and data-element ordering are the essential aspects of cognition in respect to use of an IS&R system. In this case, a cognitive data structure may be updated through a process of active hypothesis testing. Two distinct forms of hypothesis testing have been identified: 1) that concerning the existence of a specific data-element value of a node in the cognitive structure, and 2) that concerning the occurrence of data elements conforming to a particular relational pattern. The need to acquire data for the testing of such hypotheses creates what I have chosen to call *information need*.

Clearly defined use of an IS&R data base demands a prior and well-defined information need. Examples of an information need are: "What is the boiling point of water?" or "What is the melting point of Titanium Oxide at various pressures?" Presumably, a user comes to an IS&R system with the intent of discovering data that will satisify some such information need. It is postulated that the user's interaction with the IS&R system will be guided by a hypothesis he has formulated about how the data are to be retrieved. A user's hypothesis about a system's data is presumed to be dependent upon

his information need, which in turn is assumed to be dependent on the user's cognitive structure.

An initial hypothesis about data to be retrieved from an IS&R system data base could be phrased as: "There are documents in the system dealing with the *melting properties* of Titanium Oxide."* Data elements retrieved by the searcher will serve either to support or refute the hypothesis under consideration. It is possible, however, that the retrieved data are insufficient to test the initial hypothesis, and that a new, more appropriate (*i.e.*, likely to be supported) hypothesis will need to be formulated. More appropriately restrictive hypotheses might be: "There are documents dealing with the *properties* of Titanium Oxide", "There are documents dealing with Titanium Oxide." If the retrieved data do serve to test the hypothesis, and if the outcome is positive, then a decision will have to be made as to whether the hypothesis satisfies the information need. If the information need has not yet been satisfied, then a new hypothesis must be formulated and the data acquisition process resumed.

The cyclic nature of this process is further formalized in Figure 9.2.2. The information need, as determined by the user's own cognitive structure, is assumed to define a hypothesis structure which may be interpreted as a representation of all data which have been or are expected to be retrieved. Data elements from this hypothesis structure (H.S.), in conjunction with the *language* parameter of the system's index space, define the formulation of

---

* This example illustrates the use of simultaneous scanning and, as the discussion subsequently shows, such a strategy often leads to a more extended interaction with the system than would a conservative focusing strategy. Strategy is further considered in Section 9.3.

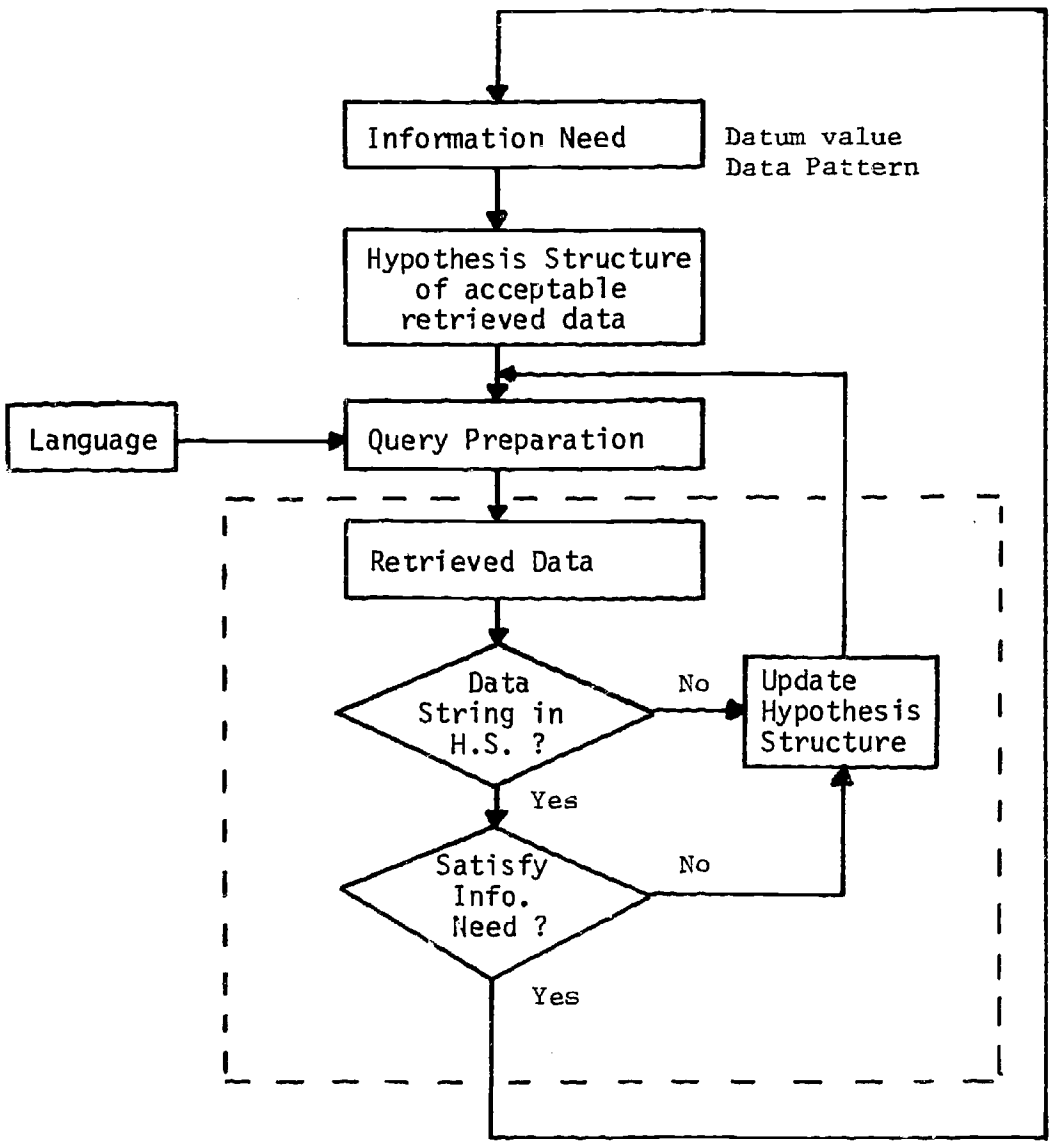


Figure 9.2.2: The Cyclic Nature of Information Retrieval.

of the query. The retrieved data are sampled until a sequence of data elements acceptable in terms of the H.S. is obtained. If the data elements of this sequence are identical with the data elements demanded by the information need, then the sampling process is temporarily suspended until a new information need arises. However, if the retrieved data elements are insufficient to satisfy the H.S., then it is assumed that the H.S. and the attendant query itself must be modified (*e.g.*, through the incorporation of new data elements into the set of acceptable strings defined by the H.S.). The reader, at this point, should realize that the steps enclosed within the dotted line of Figure 9.2.2 are a re-statement of the process-of-inquiry model previously depicted in Figure 8.3.2. I shall now represent this process in terms of a generalized machine that processes data elements as inputs.

The hypothesis formulation model is viewed as a sequence of two finite deterministic Rabin-Scott automata--see Figure 9.2.3. I assume the existence of a finite input alphabet Σ of data elements which correspond to the *transmission decoding* elements of a system's index space. A string input to the user's hypothesis-structure automaton is, then, a finite sequence of data elements from Σ: $d_1 d_2 d_3 \ldots d_n$. The hypothesis structure may therefore be described as a finite automaton over Σ,

$$H.S. = (S, TRANS, s_0, F)$$

where S is a finite set of states of H.S., TRANS (a binary transition matrix) is a mapping of $S \times \Sigma$ into S and s_0 and F are the initial and final states respectively.

In this model the states represent the data elements that the user believes to exist in the system. The binary transition matrix, TRANS,

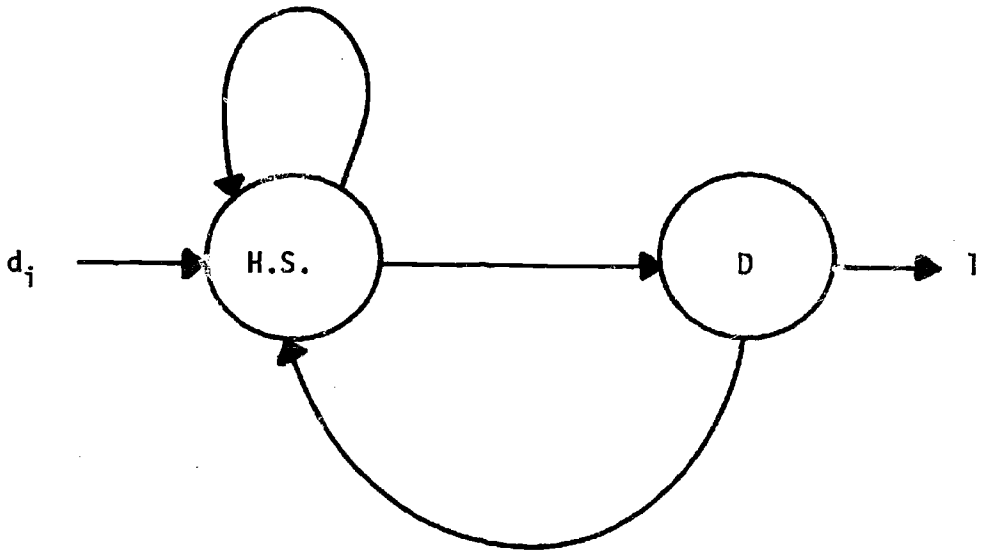


Figure 9.2.3: The Hypothesis Structure Automaton. (d_i = i th data element; H.S. = hypothesis structure; D = decision automaton; 1 = stop state)

$\alpha_{ij} \in$ TRANS, $\alpha_{ij} = 0,1$, indicates by means of a 1 that data element d_i can be followed by data element d_j. This serves to define the set of possible strings of data elements $d_{i_1}, d_{i_2}, d_{i_3}, \ldots, d_{i_r}$ that can potentially be associated with a given hypothesis. A cut-off value, λ, is applied to the sampling, so that if the sample size (the number of data elements observed from the retrieval operation) exceeds the cut-off λ, a new hypothesis structure is selected involving a new transition matrix TRANS' and a new s_0 and a new set F. This process is continued until a string

$$d_1 d_2 d_3 \ldots d_f$$

$f \in F$, $f \leq \lambda$ is obtained. A string accepted by the hypothesis structure is then input to the decision automaton, D.

The string input to the decision automaton is one of the set of possible strings $d_{i_1}, d_{i_2}, \ldots, d_{i_r}$ defined by the hypothesis-structure automaton. Thus, the decision automaton is a finite automaton defined over TRANS (or TRANS'),

$$D = (S^{\ddagger}, TRANS^{\ddagger}, s_0{}^{\ddagger}, F^{\ddagger})$$

where $S^{\ddagger}$ is a set of states representing the data elements of the information need, $TRANS^{\ddagger}$ defines the exact information need/data element sequence, and $s_0{}^{\ddagger}$ and $F^{\ddagger}$ (a singleton set) represent the initial and final data elements of the information need.

If the information need involves just a single datum then $\{S^{\ddagger}\} = s_0{}^{\ddagger} = F^{\ddagger}$. However, if the information need is a pattern of data elements (exhibiting an ordering relationship) then $s_0{}^{\ddagger}$, $F^{\ddagger}$ and $TRANS^{\ddagger}$ define a specific string of data elements. If, upon completion of the scan of the input tape, the decision automaton is not in its final state, then the operation of the

H.S. automaton is re-initiated; if the final state is reached, then the decision automaton outputs an "1" to terminate the data gathering process.

This IS&R interaction and decision-making model assumes that the data inputs are stochastically invariant. Thus, the user's information about the IS&R data structure changes with each cycle in the model--including the two feedback loops (H.S. into itself and D to H.S.). In this model the obvious user-controlled variables are λ, TRANS, s_0 and F. We assume that the definition of D is controlled by the specification of the information need. It is interesting to note that the subject's probability estimations of the transitions in the TRANS matrix are conveniently modeled by the type of Bayesian estimator employed in the information acquisition discussion of Section 7.5 of this Chapter. While such considerations tend to idealize human behavior, Peterson and Beech [65] and Schum [66] have shown that persons tend to adopt the conservative strategy (optimal) in their interential probability estimates--especially when faced with an increase in the amount of data to be evaluated. This is consistent with the observations, made in Section 8, that subjects may not readily accept the value of negative evidence. The null hypothesis is always present--indeed, it is reflected in the choice of λ. Generally, lack of user confidence in the current hypothesis structure is reflected by the assignment of a small value to the cut-off λ. This serves to place additional weight on the probability of a different hypothesis structure from the set of alternative hypothesis structures prior to subsequent interaction with the system.

This hypothesis-testing, decision-making model appears to provide the necessary quantification for the testing and evaluation of systems performance.

However, before a final consideration is made of the concept of relevance, let us examine an example of the hypothesis-testing cycle that has been developed.

9.3 An Example of the Hypothesis Structure

The theoretical notion of a hypothesis structure, which I have developed in the previous section, is susceptible of exemplification as I shall show in this section. The illustration which is offered is derived from a sample query given in the *CAS Preparation of Search Profiles* manual [67]. The query, as originally formulated, is shown in Figure 9.3.1, using set notation rather than the form employed by CAS in constructing profiles (queries). Several inferences regarding the user's knowledge of the retrieval system and of its data base can be drawn by inspection of the query in the form shown in Figure 9.3.1. For instance, the use of truncation, as for TOXIC* (a search term which would "match" TOXIC, TOXICITY, TOXICOLOGICAL, etc.), to produce generalized search terms* indicates that the user hypothesizes that the CAS search system is capable of handling such specifications.

One may argue similarly that the use of logic (AND and OR at least) as well as the use of alphanumeric** characters to describe the search terms indicates a good understanding of the attributes of the system on the user's part, or else high expectations on the part of the user as to the capabilities of the CAS search system. Although one is tempted to explore further these

---

* For details see, for example, Colombo and Rush [68].

** A term signifying alphabetic (Roman) characters, punctuation and the Arabic numerals.

{ pharmacol* or toxic or poison or analy* } and { artificial or sweeten* or saccharin* }

Figure 9.3.1: The Original Query, as taken from the *CAS Preparation of Search Profiles*.

considerations, my immediate purpose will be served if our attention is confined to inferences about the user's hypotheses concerning the contents of the data base upon which the search system operates.

The user is assumed to have come to the system with an information need expressed as a composite of three hypotheses, *viz.*,

> The system contains documents dealing with the pharmacology, toxicology and/or analysis of artificial sweeteners.

Taking this expression, augmented by the terms SACCHARIN and POISON, and employing truncation, we obtain seven data elements which constitute the initial query. These seven data elements form the names of the rows and columns of the user's TRANS matrix, as illustrated in Figure 9.3.2. For convenience of display, the TRANS matrices used throughout this example have been limited to two dimensions. In the generalized case, however, multidimensional arrays would have to be employed to account for the many possible permutations of strings of data elements which would potentially satisfy the conditions specified in the query. The TRANS matrix is entered by means of the first recognized data element, s_0, in the search output. A 1 in a cell of TRANS indicates a term in the i^{th} row of TRANS may be followed by a term in the j^{th} column of TRANS. The symbol 1_f signifies that the transition (from i to j) results in the attainment of a final state (*i.e.*, the attainment of an $f \in F$). This means that the document containing the sequence of data elements leading to a final state is accepted by the hypothesis structure (H.S.) automaton. Examples of strings of data elements which would be accepted by the H.S. automaton (TRANS matrix) of Figure 9.3.2 include:

s_0	pharmacol*	toxic*	analy*	artificial	sweeten*	saccharin*	poison
pharmacol*				1	1_f	1_f	
toxic*				1	1_f	1_f	
analy*				1	1_f	1_f	
artificial					1_f		
sweeten* f							
saccharin* f							
poison*					1_f	1_f	

Figure 9.3.2: The Initial Hypothesis Structure (TRANS matrix).

s_o = initial term
f = final state
* denotes truncation of suffix or prefix
λ = 5

- PHARMACOL* ARTIFICIAL SWEETEN*
- ANALY* SACCHARIN*
- SWEETEN*
- SACCHARIN*

Such sequences of data elements are not intended to carry the implication that any one specific logical or functional relationship exists between these data elements as they occur in a source document. Rather, a particular sequence merely represents the user's expected order of occurrence of the data elements in the document.

With a sampling cut-off value of five (*i.e.*, $\lambda = 5$) for the initial query, as represented in Figure 9.3.2, the document whose title is given below was retrieved (and was accepted by the H.S. automaton).

200 Analytical methods of artificial sweeteners. Determination of sodium cyclamate.

The string of data elements

ANALYTICAL - ARTIFICIAL - SWEETENERS

supports the hypothesis that there are documents in the data base which contain data on the analysis of artificial sweeteners. However, the other two hypotheses of the original composite hypothesis remain unsupported. It should be noted that the sequence of data elements presented to the H.S. automaton (with $\lambda = 5$) namely

ANALYTICAL METHODS OF ARTIFICIAL SWEETENERS

Contains two non-query data elements,

- METHODS
- OF

These data elements correspond to null states in the TRANS array. The remain-

ing three non-query data elements in the retrieved document--DETERMINATION - SODIUM- CYCLAMATE--are not considered by the H.S. automaton because $\lambda = 5$, however they may serve as meta-information in the formulation of a subsequent query.

The second query (or search iteration) represents an attempt, on the part of the user, to obtain clear support or refutation of the remaining two hypotheses. However the user has already obtained *some* information about these two hypotheses. The reader should recall that the absence of data provides both meta-information with respect to the information need and information with respect to the process of inquiry. Thus, the new query is modeled by an updated matrix, TRANS' (see Figure 9.3.3), in which two new terms, taken from the data elements associated with document number 200, have been added to those of TRANS:

- DETERMIN*
- CYCLAMATE*

With λ increased to 20 (document # 200 required a sampling of 5 data elements to be retrieved) the following documents are retrieved and accepted by the H.S. Automaton:

# 200	Analytical methods of artificial sweeteners. Determination of sodium cyclamate.
# 100	Mechanism of the laxative effect of sodium sulfate, sodium cyclamate and calcium cyclamate.
# 350	Rapid method for the estimation of impurities in saccharin and sodium saccharin.
# 50	Peptide synthesis with mixed anhydrides from N-acyl amino acids and saccharin.
# 39	Distribution and excretion of carbon-14-cyclamate sodium in animals.

s_o	pharmacol*	toxic*	analy*	artificial	sweeten*	saccharin*	poison*	determin*	*cyclamate*
pharmacol*		1	1	1	1_f	1_f			
toxic*			1	1	1_f	1_f			
analy*				1	1_f	1_f			1_f
artificial					1_f				
sweeten* f									
saccharin* f									
poison*					1_f	1_f			
determin*				1	1_f				1_f
cyclamate f									

Figure 9.3.3: The Updated Hypothesis Structure (TRANS' matrix). $\lambda = 20$

Documents 100, 350 and 39 support the remaining two hypotheses. These documents, furthermore, would satisfy the information need if the user's D automaton* provided an explicit association relationship between the retrieved data elements:

MECHANISM, LAXATIVE, IMPURITIES, EXCRETION

and the search terms *pharmacology* and *toxicology*. If the D automaton failed to accept any or all of documents 100, 350 and 39, this would become clear (to an observer) by the observation of the user effecting a third search iteration with the system. Although we shall not pursue the example further, a possible subsequent H.S. modification would include the assignment of a value of $\lambda = 10$ (to eliminate document # 50) and the addition of the new data elements

- PHYSIOLOG*
- IMPURIT*
- EXCRETION*

to the TRANS' matrix.

10. A Reconsideration of the Concept of Relevance

> "How is bread made?"
> "I know that" Alice answered eagerly. "You take some flour --"
> "Where do you pick the flower?" The White Queen asked. "In a garden, or in the hedges?"
> "Well, it isn't picked at all," Alice explained. "It's *ground*--"
> "How many acres of ground?" said the White Queen. "You mustn't leave out so many things."
>
> Lewis Caroll

Although I have, in this chapter, neither specified not resolved all possible sources of doubt and have no doubt overlooked many important topics,

---

* The structure of this automaton is analogous to the association table of related terms mentioned by Kochen *et.al.* [69].

I hope the material presented has contributed to a clearer understanding of the concepts of evaluation and *relevance*. Relevance assessment has been described as an integral part of an algorithm that embodies the process of inquiry which is characteristic of an interactive retrieval interface. Discussion has so far relied heavily on the acceptance of the concepts of information need, inquiry, problem solving, hypothesis testing, attribute identification and, implicitly, data-element relevance. Let me summarize briefly what has been said about relevance and evaluation in this context.

I have accepted as a premise that the problem of IS&R systems evaluation is both paramount to effective systems design and a corollary to the theoretical considerations that have been developed in the previous chapter. The arguments that have been presented have assumed, further, that evaluation is best described as a value judgment or formal correlation between data retrieved by a user and his information need. There remains the task of quantifying this correlation. All correlational measures, in IS&R applications, presumably reduce, ultimately, to variations of the measures of *recall* and *precision*, which in turn are based on a user's precise determination of document relevance. Unfortunately, relevance remains a subjective, fuzzy concept. I have chosen to attack the problem through a detailed analysis of the concept of *information need*. It is believed that the testing of hypotheses concerning the user's environment (analogous to the Scientific Method) is consequent upon his information need. It has been useful to describe the hypothesis → information-need → hypothesis cycle in terms of problem solving behavior that focuses on the user's identification of attributes and the employment of optimizing strategies. This conceptualiz-

ing on my part about a user's thinking and approach has led to the development of a hypothesis structure model which I believe to be descriptive of an IS&R-system user's data-element acceptance abilities. This model, as we shall see directly, provides for the necessary quantification of the user's relevance decision.

10.1 Relevance

Cooper [70] describes the action of an IS&R system in response to a query as the establishment of a "ranking among the documents in the collection." The rank-ordered documents are then examined, one-by-one, and a decision is made concerning their utility in the satisfaction of the information need. However, the previous discussion implies that utility decisions are made upon data elements--not documents. The reader will recall that the inputs to a hypothesis-structure automaton are strings of data elements, and the concept, or framework, of a document was treated as purely coincidental to these inputs. We assume that strings of data elements are the essential components of the concept of relevance. Two distinct, *relevant* strings of data elements are identified:

- Data elements accepted by the H.S. automaton.
- Data elements accepted by the H.S. automaton and matching the user's information need.

The complexity of the strings of data elements are postulated to range from a specific datum to complex, prescribed patterns of data elements. To me, the important observation is the recognition that *relevance* refers either to the acceptance or to the matching of data elements. Thus, three forms of data-element match are identified:

- specific data element value (datum value) *e.g.*, 27 $\pm$ 0.1 feet high
- total data element pattern match (*e.g.*, Patent Office novelty and anticipation searches)
- general pattern or equivalence class of data elements (*e.g.*, a genus of organism)

Regardless of the type of match that is required in a search, the probability that rank-ordered data elements will be relevant (*i.e.*, accepted) is assumed to depend on the degree to which the user's query (hypothesis structure) can be embedded in the IS&R data structure. Embedding is interpreted as structural similarity. The extent of the similarity between a query and the system's data base is measured in terms of the degree of overlap of the syntax, data elements and relationships present in both the user's H.S. and the IS&R system.

The two forms of relevance are shown in Figure 10.1.1, which itself is a depiction of the various processes that, potentially, are utilized in an interactive retrieval environment. The correlation between the query and the retrieved data elements is not shown because it is assumed always to be perfect. In any event, the two forms of relevance are believed to be an integral part of a generalized inquiry algorithm. In such a context, relevance is a measure of the precision of the measurement and perception employed by the algorithm. Various forms of such an algorithm can be listed by increasing precision:

- Scientific Method
- Process of Inquiry
- Problem Solving
- Hypothesis Testing

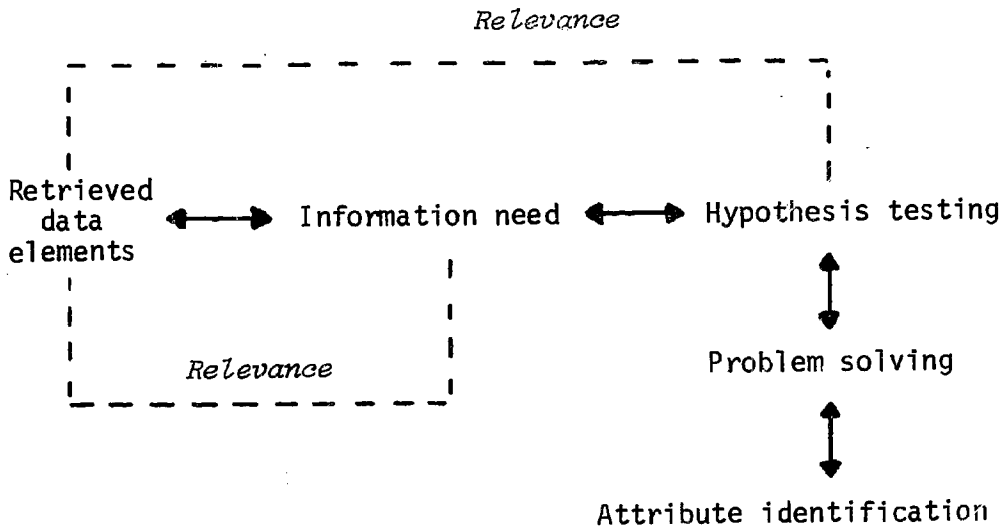


Figure 10.1.1: The Two Forms of Relevance.

- Actribute Identification

Even though the algorithms associated with the first two levels of precision may be esoteric, they are still accountable in terms of the aforementioned hypothesis structure model.

Interesting support for the particular view of relevance which is presented in this Chapter can be found in the results of experiments 8 and 12 (analysis of precision and recall failures) of the recent *United Kingdom Chemical Information Science* (UKCIS) report [71]. Briefly, the most significant factors influencing "precision" failures were found to be [72]:

- items retrieved by terms in the wrong context
- wrong correlation of terms

These conclusions indicate an observer-deduced failure in the searcher's hypothesis structure configuration--*i.e.*, the user's initial hypothesis about the form of the data base is interpreted by an outside observer as incorrect. Corrective search iterations are required [illegible] hypotheses (updating of the transition values in the hypothesis structure matrix - TRANS). The UKCIS study also identified the following factors, which were inferred to influence the user's "recall" failures:

- profile narrower than interest (concepts [illegible])
- inadequate concept expansion
- too restrictive logic or weighting
- input error

These results support the assertion made in Section 8 of this chapter: as viewed by an outside observer, users generally fail to identify and use pertinent system attributes. Part of this problem is believed to stem from

the user's inability to perceive the system's data structure and is reflected in his incorrect hypothesis structure. Attribute identification is no small problem, since most systems emphasize the importance of a "middle man" who identifies the "relevant" system attributes for the user. It becomes apparent that either user problem-solving abilities need to be improved, or IS&R systems must match their operation to *classes* of problem-solving skills manifested by their users.

10.2 Evaluation

Since the quantification of relevance reduces to the identification and description of strings of data elements accepted by the H.S. automaton, the problem of IS&R systems evaluation is resolved through an analysis of the variables associated with the hypothesis structure model. Six major variables are identified.

- User's system confidence value as indicated by his choice of cut-off value, λ. The user's view of the reliability of the system is seen in the inverse relationship that exists between λ and confidence. A high λ-value implies a potentially lengthy sampling and a low confidence in the system's rank ordering.
- The number of hypotheses tested.
- The time between iterations.
- The time devoted to problem solving.
- The number of decisions made.
- The number of redundant data elements that have to be examined before a pattern can be perceived. This amounts to a paraphrasing of Cooper's *Average Search Length* [73]: The number of irrelevant data elements one would expect to search through before finding as many relevant data elements as needed.

Measures of IS&R system evaluation, by implication, should not be averaged over all systems or all users. Rather, evaluation (as it has been defined

above) may be employed by an outside observer to describe system performance with respect to well-defined classes of users (this implies, also, the possibility of inducing well-defined classes of queries). Sackman touches on this problem [74]:

> An empirically derived taxonomy of man-computer tasks should be developed based on demonstrated differences in human-problem solving style rather than on a confusing welter of strictly descriptive characteristics.

The notion "classes of users" is posited to be identically similar to the notion of the Bruner, *et.al.*, strategies of problem solving discussed in Section 8. Thus, a system's performance may be evaluated on its ability to satisfy the information needs of well-defined classes of problem-solving users. The establishment of such "well-defined classes" must be a prerequisite [illegible] of evaluation.

In summary, I have attempted to provide in this chapter a workable definition of relevance, which is consistent with theoretical indexing and human behavioral considerations. Possibly the most significant feature of the previous discussion about relevance is its attempt to provide an integrated, cross-disciplinary approach to relevance assessment. The retrieval interface is not studied *in vacuo*, but as a complex symbiotic process. A model of retrieval problem solving has been developed that is assumed to be amenable to quantification. This quantification is achieved through a definition of the H.S. and D automata and, of course, the concept of the data element. From this point of view, evaluation has been presented as a measure defined over classes of users and their queries. Finally, it should be emphasized that prior imprecision in reference to the term "relevance" has been reduced to the subjective selection by the outside observer of transition values in the H.S. automaton.

References

1. H. Wooster, "An Information Analysis Center Effectiveness Chrestomathy," *Journal of the American Society for Information Science* 21(2), 1970, 149-159.

2. A.M. Rees, "Information Needs and Patterns of Usage," *Information Retrieval in Action*, The Press of the Western Reserve University, Cleveland, 1963, 17-23.

3. P.A. Richmond, "The Final Report of the Comparative Systems Laboratory: A Review," *Journal of the American Society for Information Science* 21(2), 1970, 160-162.

4. M.E. Stevens, "Automatic Indexing: A State-of-the-Art Report." *National Bureau of Standards Monograph 91*, 1965.

5. D.J. Hillman, "The Notion of Relevance (I)," *American Documentation* 15(1), 1964, 26-34.

6. *Proceedings of the International Conference of Science Information*, National Academy of Sciences, Washington, D.C., 1959.

7. C.P. Bourne, "Evaluation of Information Systems," in *Annual Review of Information Science and Technology* Vol. 1, C. Cuadra,(ed.), John Wiley and Sons, Inc., New York, N.Y., 1966, 171-190.

8. J.R. Sharp, "Content Analysis, Specification, and Control," in *Annual Review of Information Science and Technology* Vol. 2, C. Cuadra, (ed.), John Wiley and Sons, Inc., New York, N.Y., 1967, 87-122.

9. O.E. Taulbee, "Content Analysis, Specification, and Control," in *Annual Review of Information Science and Technology* Vol. 3, C. Cuadra, (ed.), Encyclopaedia Britannica, Chicago, Illinois, 1968, 105-136.

10. R.A. Fairthorne, "Content Analysis, Specification, and Control," in *Annual Review of Information Science and Technology* Vol. 4, C. Cuadra, (ed.), Encyclopaedia Britannica, Chicago, Illinois, 1969, 73-110.

11. M.M. Henderson, "Evaluation of Information Systems: A Selected Bibliography with Informative Abstracts," *National Bureau of Standards Publication Number 297* PB-188-657, 1967.

12. M. Taube, "A Note on the Pseudo-Mathematics of Relevance," *American Documentation*, 16(2), 1965, 69-72.

13. F.W. Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The Aslib-Cranfield Project," *American Documentation* 15(1), 1964, 4-13.

14. Arthur D. Little, Inc., "Centralization and Documentation," *Final Report to the National Science Foundation*, 1963.

15. F.W. Lancaster, *Information Retrieval Systems: Characteristics, Testing and Evaluation*, John Wiley and Sons, Inc., New York, N.Y., 1968.

16. F.M. O'Hara, Jr., "The Corners and Edges of the Precision-Recall Square," *Journal of the American Society for Information Science* 21(2), 1970, 166.

17. G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill, New York, N.Y., 1968.

18. S.M. Pollock, "Measures for the Comparison of Information Retrieval Systems," *American Documentation* 19(4), 1968, 387-397.

19. J.A. Swets, "Effectiveness of Information Retrieval Methods," *American Documentation* 20(1), 1969, 72-89.

20. I.J. Good, "The Decision-Theory Approach to the Evaluation of Information Retrieval Systems," *Information Storage and Retrieval* 3, 1967, 31-34.

21. W.S. Cooper, "Expected Search Length--A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," *American Documentation* 19(1), 1968, 30-41.

22. E.R.C. Ide, "Relevance Feedback in an Automatic Document Retrieval System," *ISR Report #ISR-15*, Cornell University, PB-184-246, 1969.

23. C.A. Cuadra, "Experimental Studies of Relevance Judgments: Final Report," *Systems Development Corporation Report TM-3520/001-003*, 1967.

24. W.J. Paisley and E.B. Parker, "Information Retrieval as a Receiver-Controlled Communication System," in *Education for Information Science* L. Heilprin, (ed.), Spartan Books, 1965, 23-30.

25. P. Zunde and M.E. Dexter, "Factors Affecting Indexing Performance," *Proceedings of the American Society for Information Science* 6 1969, 313.

26. M.C. St. Laurent, "Studies in Indexing and Retrieval Effectiveness," *Thesis*, University of Chicago, PB-174-395 , 1967.

27. M. Lay, Personal Communication, 1969

28. W. Goffman, "On Relevance as a Measure," *Information Storage and Retrieval* 2, 1964, 201-203.

29. J. O'Connor, "Relevance Disagreements and Unclear Request Forms," *Institute for the Advancement of Medical Communication*, Philadephia, Pa., 1966.

30. P. Leslie, "Indexing Philosophies," in *Practical Problems of Library Automation*, Special Libraries Association, 1967, 7-22.

31. C.A. Cuadra, "Experimental Studies of Relevance Judgments: Second Progress Report," *Systems Development Corporation Report TM-3068/000/000*, 1966.

32. D.L. Kegan, "Measures of the Usefulness of Written Technical Information to Chemical Researchers," *Journal of the American Society for Information Science* 21(3), 1970, 179-189.

33. J. O'Connor, "Some Questions Concerning 'Information Need'", *Journal of the American Society for Information Science* 19(2), 200-203 (1968).

34. P. Caws, *The Philosophy of Science: A Systematic Account*, D. Van Nostrand Company Inc., Princeton, New Jersey, 1965, 17.

35. *ibid.*, 15.

36. *ibid.*, 79.

37. M. Schlick, "Meaning and Verification," in Feigl and Sellars, *Readings in Philosophical Analysis*, New York, 1949, 157.

38. L. Brillouin, *Science and Information Theory*, Academic Press Inc., New York, N.Y., 1956, 152-161.

39. S. Watanabe, *Knowing and Guessing: A Quantitative Study of Inference and Information*, John Wiley & Sons, Inc., New York, N.Y., 1969, 112.

40. P.M. Fitts and M.I. Posner, *Human Performance*, Brooks/Cole Publishing Co., Belmont, California, 1968, 5.

41. M. Minsky (ed.), *Semantic Information Processing*, The MIT Press, Cambridge, Mass., 1968, 27.

42. *ibid.*, 427.

43. G. Harmon, "Information Need Transformation During Inquiry: A Reinterpretation of User Relevance," *Proceedings of the ASIS Annual Meeting* 7, 41 (1970).

44. M.R. Quillian, "The Teachable Language Comprehender: A Simulation Program and Theory of Language," *Communications of the ACM* 12(8), 459-476 (1969).

45. C.L. Bernier and K.F. Heumann, "Correlative Indexes III. Semantic Relations Among Semantemes--The Technical Thesaurus," *American Documentation* 8, 1957, 211-220.

46. R.L. Ernst and M.C. Yovits, "Information Science as an Aid to Decision-Making", Computer and Information Science Research Center, The Ohio State University, Tech. Report 69-13, 1969, 17.

47. J. Dewey, *Logic: The Theory of Inquiry*, Holt, Rinehart & Winston, New York, N.Y., 1938, 140.

48. B. Russell, *Human Knowledge*, Simon & Schuster, New York, N.Y., 1948, 487.

49. A.N. Whitehead, *Adventures of Ideas*, Mentor, New York, N.Y., 1933,205.

50. J. Dewey, *How We Think*, Heath, Boston, 1933, 15.

51. J.P. Guilford, *The Nature of Human Intelligence*, McGraw-Hill, New York, N.Y., 1967, 315.

52. *ibid.* 203.

53. L.E. Bourne, Jr., "Concept Learning and Thought: Behavior, not Process" in *Approaches to Thought*, J.F. Voss, (ed.), Merrill, 1969, 186.

54. J.S. Bruner, J.J. Goodnow and G.A. Austin, *A Study of Thinking*, John Wiley & Sons, Inc., New York, N.Y., 1956, 26.

55. *Ibid*, 83.

56. W.A. Reitman, *Cognition and Thought*, John Wiley and Sons, Inc., New York, N.Y., 1965.

57. T. Saracevic, "On the Concept of Relevance in Information Science," Ph.D. Dissertation, Case Western Reserve University, 1970, 111.

58. *ibid*, 117.

59. *ibid*, 92.

60. Bruner, *et al.*, *op. cit.*, 42.

61. *ibid*, 237.

62. M. Levine, "Hypothesis Theory and Non Learning Despite Ideal S-R-Reinforcement Contingencies," *Psychological Review* 78(2), 1971, 130-140.

63. H. Sackman, *Man-Computer Problem Solving*, Auerbach, New York, N.Y., 1970.

64. J.R. Carbonell, "On Man-Machine Interaction: A Model and Some Related Issues," *IEEE Transactions on Systems Science and Cybernetics*, SSC-5(1), 1969, 16.

65. C.R. Peterson and L.R. Beach, "Man as an Intuitive Statistician," *Psychological Bulletin*, 68, 1967, 29-46.

66. D.A. Schum, I.L. Goldstein, W.C. Howell and J.F. Southard, "Subjective Probability Revisions Under Several Cost-payoff Arrangements," *Organizational Behavior and Human Performance*, 2, 1967, 84-104.

67. Chemical Abstracts Service, *Preparation of Search Profiles*, American Chemical Society, Columbus, Ohio, 1967, 32-38.

68. D.C. Colombo and J.E. Rush, "Use of Word Fragments In Computer-Based Retrieval Systems," *Journal of Chemical Documentation* 9(47), 1969, 47-50.

69. R. Tagliacozzo, D. Semmel and M. Kochen, "Written Representation of Topics and the Production of Query Terms," *Journal of the American Society for Information Science* 22(5), 1971, 345.

70. W. S. Cooper, "On Deriving Design Equations for Information Retrieval Systems," *Journal of the American Society for Information Science* 21(6), 1970, 387.

71. F.H. Barker, A.K. Kent and D.C. Veal, "Report on the Evaluation of an Experimental Computer-Based Current Awareness Service for Chemists," The Chemical Society, Burlington House, London, 1970.

72. *ibid*, 34.

73. Cooper, *op. cit.*, 387.

74. Sackman, *op. cit.*, 240.

VI. SUMMARY AND RESEARCH DIRECTIONS

> The reason why there is no Table or Index added hereunto is, that every page is so full of signal remarks that were they couched in an Index it would make a volume as big as the book, and so make the Postern Gate to bear no proportion to the building.
>
> Howell

The material contained in this brief concluding chapter is presented in three sections. First, a summary of the previous five chapters is presented. This is followed, secondly, by a discussion of the several ways in which indexing theory models information storage and retrieval. Finally, directions for future research in indexing theory are outlined.

1. Summary

We began this inquiry into indexing theory with a brief consideration of some of the causes of the phenomenon commonly referred to as the "information explosion." It was concluded that present day shortcomings in information retrieval are the result of a failure to properly contend with the problem of data representation. Science does not suffer from a lack of accumulated knowledge but, rather, it suffers from the inability to efficiently communicate what has been previously discovered. As a consequence of the difficulty of data communication within and between the Sciences, there has been a growth in the number of specialized areas of investigation. In many respects this proliferation of specialties has only served to further hamper effective data and information retrieval.

Information storage and retrieval was initially characterized as a communication interface between the data of the Sciences and a diverse population of users. The object of any interaction with IS&R systems is to

develop a high level of shared agreement or common understanding between the storage scheme of the system and both the information need and the resulting search techniques employed by the user. It was postulated that the effectiveness of this interaction was, primarily, dependent on the fidelity of document representation. Furthermore, it was assumed that the indexing operation was a prime exemplar of the process of document representation.

The field of IS&R suffers from the absence of a unifying theory of document (data) transfer. Consequently, one is embarrassed by the difficulty of effectively evaluating the many IS&R systems that are currently in operation (and the many more that are still in the planning stages). It was concluded that the primary goal of theory development is the creation of sound evaluation measures. Furthermore, it was argued that a theory of indexing could serve as an adequate model for many of the processes of information storage and retrieval. If properly developed, this theory would provide the basis for the systematic analysis of both indexing procedures and of resultant indexes, and it would provide the conceptual basis for the development of evaluation techniques. Unfortunately, initial investigation is hampered by a long history of considering indexing as an "artful" practice. Thus, a theory of indexing must first turn to a consideration of the following fundamental questions: why index at all; what should be indexed; what is the role of indexing "aids" in the process of indexing and; how are indexes to be evaluated?

Previous indexing theories can be faulted for not providing answers to these fundamental questions. Two related theories were reviewed, mainly for the purpose of building in the reader an appreciation of their general tone

and direction of approach. Heilprin's theory raised several interesting points which have been pursued in the current work toward a theory of indexing. These include the conceptualization of indexing systems as closed systems; the importance of the effect of noise in the indexing operation; the concept of a search path and, finally, the use of an indexing region as a means of system characterization.

Chapter four presented the basis for a comprehensive theory of information storage and retrieval. The fundamental thesis was that this theory had its genesis in a theory of the indexing process. In other words, as has been previously emphasized, it is believed that the success of an IS&R system depends primarily on accurate and complete document representation, and that such document representation is the goal of any indexing process. It was contended that the index provides the necessary linkage between a multiplicity of sources and a single receiver. Conceptually, the indexing system is initially viewed as a black box that accepts documents as its inputs and produces the index as its only product (output). Various sources produce the documents which become the elements of the document space and receivers produce queries which are matched against the index and, eventually, against the document store. Whether considering the source/document-space interface or the query/index interface, the elements of the underlying communication phenomena are the same: sets of documents, sets of attributes and sets of relations expressing a connection between documents and attributes. I have chosen to represent these attributes by the concept of the data element. Following the progression of schema presented in Figure 10.1 of Chapter IV, we first considered the necessary criteria for effective communication and

concluded that the index provided the requisite common experience set between the source and the receiver. We then more precisely positioned the indexing system intermediary between the communication channel and the receiver (searcher) and emphasized the role of "noise" and feedback. Following a specification of the "position" of the black box or indexing system in communication, we considered a theory of its operation. This theory, called the indexing process, defined the essential operation of the indexing system to be the creation of a representation of the document space. The analysis-document transformations and the final index-query transformations were shown to be, respectively, a prerequisite to, and a function of, the document space representation. Examples of these transformations were provided through the analysis of a sample document. Finally, the operating characteristics of the indexing system were modeled by means of the index space. From a different point of view, the concepts of error, organization, information and search were introduced through a consideration of the indexing process as a thermodynamic system. Thus, indexing was viewed as an order-increasing operation that identifies common data elements and relations between data elements present in the input document stream. The existence of both the "perfect" indexing system and the theoretical index were then postulated and compared with their real-world counterparts. Several suggestions for real-world indexing improvements (with the idea of emulating the theoretical index) were presented and, finally, it was argued that the value of each newly retrieved data element was a function of the order of retrieval.

In Chapter V attention was directed toward applying the current indexing theory to the problem of IS&R systems evaluation. It was postulated that

successful systems evaluation is based on an understanding of why certain retrieved documents are judged to be relevant to the searcher. Specifically, I wanted to know more about data element relevance. Unfortunately, the concept of relevance (and evaluation, for that matter) has had a rather confused history in the field of IS&R. Thus, I have chosen to investigate the nature of data element relevance by means of a reconsideration of the concept of information need. It was argued that an information need resulted from two alternative forms of hypothesis testing. Consequently, the process of satisfying an information need involves the utilization of problem solving strategies and selective decision-making criteria. Following a discussion concerning the processing of attributes, the results of a brief experimental investigation were presented that indicated that the success of problem solving strategies, and of hypothesis testing, was directly related to the level of structure associated with the problem solving setting. Finally, a hypothesis structure automaton was presented as a model of how a searcher evaluates the relevance (acceptability) of retrieved strings of data elements.

2. Indexing Theory as a Model of IS&R

In Chapter II we broadly characterized information storage and retrieval as serving to provide the basis for document data-element representation and searching. More specifically, the diverse operations of document acquisition, data-element representation, document storage, query preparation, data-element searching, document retrieval and retrieval evaluation were singled out, in Figure 2.1 of Chapter II, as prerequisites for successful information retrieval. That figure also showed the document representation and storage chain merging with the query/information-need processing chain

at the operation of the data-base search. One of the major conclusions to be drawn from the presentation in Chapter IV is that the indexing operation (data-element/document representation) is the controlling factor in the success of the search operation. That is to say, efficient storage and search algorithms are meaningless, with respect to satisfying the information need, if an accurate and complete document representation, *i.e.*, indexing, is not provided initially.

Thus, we have modeled IS&R with a theory of indexing which amounts to a theory of IS&R's most crucial operation. Like the characterization of the index, the theory is itself bi-directional in nature. The first part of the theory has dealt with the problem of the representation of data elements in a multiplicity of documents; the second part of the theory has been concerned with the user viewing the index as a tool for the resolution of an information need. It was concluded that one must speak about the use of the index when discussing the theory of its construction since an index is surely created to be used. Furthermore, it was concluded that the manner of index construction (and the form of the resultant index) specifies the class of queries that are acceptable to the retrieval system. Thus, a theory of index construction is, implicitly, a theory about index search and, consequently, a theory of information retrieval.

Both portions of the theory are amenable to evaluation. First, indexing viewed as a process for the representation of data elements and relations between data elements was modeled by a set of transformations which are applied by the indexing system to the input documents to create the index as an end product. The effectiveness of the transformations, and the

specificity of the index (as a representation of the input documents) are evaluated by comparing the resultant index with the theoretical index. Second, indexing viewed as the creation of a search interface places emphasis on the association structure created by the indexing process. We have postulated that the utility of the index can be measured following an understanding of how people go about using the index. Consequently, the evaluation of whether a system is able to provide an acceptable string of data elements is obtained through the observation of the rate of convergence of user decisions and hypotheses toward the satisfaction of the information need.

3. Directions for Future Research

'But I should like to know...' Pippin began. The information presented in this dissertation has only begun to satisfy this author's inquisitiveness about the processes of indexing and information storage and retrieval. As a consequence of this investigation it is possible to identify three separate, but conceptually related, directions for future research.

- Further studies in the theoretical representation of indexing.

It has been beneficial, from a theoretical viewpoint, to characterize the operations of the indexing system both by means of generalized transformations and by means of the index space. The next step is to develop these transformations and representations for specific operational indexing systems. It is hypothesized that such a detailed analysis will indicate the degree of ordering effected by alternative indexing systems. Also, it is believed that such a detailed analysis will show what types of data elements are preserved or discarded in the specific indexing process.

- The index as the search interface.

Further studies, following the lines of the extended Bruner experiment, should be undertaken to determine how users go about the identification of attributes in a retrieval setting. It is hypothesized that an understanding of how users differ with respect to the identification and the utilization of attributes (and structure) will be helpful in the design of improved indexes. From the point of view of evaluation, the hypothesis structure atuomaton suggests the development of a simulation model for predicting the retrieval behavior (and information-need satisfaction) of classes of users under varying retrieval requirements.

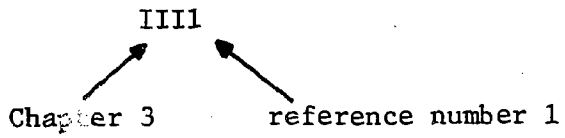
- The Case Grammar Index.

It is believed that this approach to the analysis of document content will yield an index that accurately represents data elements and relations between data elements. Such an assertion will have to be tested by the indexing of a sample document space.

This author hopes to have the opportunity of continuing work along the lines indicated above, and it is expected that various aspects of this research will be continued in these Laboratories.

KWIC INDEX OF CUMULATED REFERENCES

The following pages contain a KWIC (Key-Word-In-Context) index of the cumulated references from the preceeding chapters. Using a stop list of 36 common and function words, 147 bibliographic records generated 922 index entries. Document source entries are listed on pages 241-245; document author entries are listed on pages 245-251 and; document title words are listed on pages 252-271. The four-character accession number that accompanies each entry serves to indicate the chapter and citation number of the reference in question:



POSNER, M.I., HUMAN PERFORMANCE, BROOKS/COLE, 1968.= FITTS, P.M., V 40
ROCEEDINGS OF THE 3RD SYMPOSIUM, BUTTERWORTHS, 1955.=+ C., INFORMATION THEORY: P IV15
NG IN THE THEORY OF INFORMATION, BUTTERWORTHS, 1955.=+, D.M., THE PLACE OF MEANI IV15
., TOWARD INFORMATION RETRIEVAL, BUTTERWORTHS, 1965.= FAIRTHORNE, R.A IV18
ELEVANCE IN INFORMATION SCIENCE, CASE WESTERN UNIV., 1970.=+ ON THE CONCEPT OF R V 57
LAN, A., THE CONDUCT OF INQUIRY, CHANDLER, 1964.= KAP II14
PREPARATION OF SEARCH PROFILES, CHEMICAL ABSTRACTS SERVICE, 1967.= V 67
CE AS AN AID TO DECISION MAKING, CISRC, 69-13, 1969.=+S, M.C., INFORMATION SCIEN V 46
ND RE-INDEXING SIMULATION MODEL, CISRC, 69-14, 1969.=+ANDRY, B.C., AN INDEXING A III8
ALIZATION OF ENTROPY IN PHYSICS, CISRC, 70-24, 1970.=+N, J., INFORMATIONAL GENER IV49
PROGRAM AND THEORY OF LANGUAGE, COM. ACM 12(8),1969.=+OMPREHENDER: A SIMULATION V 44
., MEANING AND MISUNDERSTANDING, CONTEMP. PSYCH. 14(7),1969.= WEICK, K I 01
THE NATURE OF THE CHEMICAL BOND, CORNELL UNIV. PRESS, 1940.= PAULING, L.C., IV24
, P., THE PHILOSOPHY OF SCIENCE, D. VAN NOSTRAND, 1965.= CAWS V 34
, P., THE PHILOSOPHY OF SCIENCE, D. VAN NOSTRAND, 1966.= CAWS II17
STUDIES IN COORDINATE INDEXING, DOCUMENTATION INC. 1965.= TAUBE, M., II10
ORMATION SCIENCE AND TECHNOLOGY, ENCYCLOPAEDIA BRITANNICA JOHN WILEY.=+ OF INF I 14
GILYAREVSKII, R.S., INFORMATICS, FID PUBL. 435, 1969.=+V, A.I., CHERNI, A.I., I 06
INVITATION TO CHEMICAL RESEARCH, FRANKLIN, 1961.= REID, E.E., IV52
--UN MODEL GENERALE-- LE SYNTOL, GAUTIER-VILLARS, 1964.=+CHERCHES DOCUMENTAIRES IV58
AL METHODS FOR RESEARCH WORKERS, HAFNER, 1946.= FISHER, R.A., STATISTIC IV37
DEWEY, J., HOW WE THINK, HEATH, 1933.= V 50
., LOGIC: THE THEORY OF INQUIRY, HOLT, RINEHART & WINSTON, 1938.= DEWEY, J V 47
LLMORE, C.J., THE CASE FOR CASE, HOLT, RINEHART & WINSTON, 1968.= FI IV59
TRODUCTION TO TAGMEMIC ANALYSIS, HOLT, RINEHART & WINSTON, 1969.=+COOK, W.A., IN IV62
+ MODEL AND SOME RELATED ISSUES, IEEE TRANS. SYSTM. SCI. CYBER. SSC-5(1),1969.=+ V 64
, W., ON RELEVANCE AS A MEASURE, INFORM. STORAGE & RET. 2,1964.= GOFFMAN V 28
ALUATION OF INFORMATION SYSTEMS, INFORM. STORAGE & RET. 3,1967.=+ROACH TO THE EV V 20
ALYSIS OF DOCUMENTATION SYSTEMS, INFORM. STORAGE & RET. 3,1967.=+MATHEMATICAL AN IV07

NTERPRETATION OF USER RELEVANCE,
AN AS AN INTUITIVE STATISTICIAN,
S-R-REINFORCEMENT CONTINGENCIES,
FOR THE PREPARATION OF INDEXES,
XING METHODS AND SEARCH DEVICES,
H, R., NON-CANTORIAN SET THEORY,
E, E.C., ENERGY AND INFORMATION,
ALLE, R., MATTERS OF CONCINNITY,
UDGMENTS SECONDPROGRESS REPORT,
LEVANCE JUDGMENTS FINAL REPORT,
RUSSELL, B., HUMAN KNOWLEDGE,
SLIF, P., INDEXING PHILOSOPHIES,
PROBLEMS OF LIBRARY AUTOMATION,
FOR RETRIEVAL CENTER OPERATIONS,
NDEXING FOR HEURISTIC RETRIEVAL,
UCATION FOR INFORMATION SCIENCE,
CONTROLLED COMMUNICATION SYSTEM,
IN DEPTH: PRACTICAL PARAMETERS,
TION HANDLING: FIRST PRINCIPLES,
AWARENESS SERVICE FOR CHEMISTS,
ACTICE IN INFORMATION RETRIEVAL,
STORAGE AND RETRIEVAL SYMPOSIUM,
MATICAL THEORY OF COMMUNICATION,
ED STUDY OF INFORMATIVE DISPLAY,
ARD BASIC CRITERIA FOR INDEXERS,
ION NEEDS AND PATTERNS OF USAGE,
E GENERATION OF SUBJECT INDEXES+
+ BRUNER, J.S., GOODNOW, J.J.,
+ BRUNER, J.S., GOODNOW, J.J.,
NGUISTIC THEORY, HOLT, RINEHA+

PROC. ASIS 7,1970.=+ATION DURING INQUIRY: A REI V 43
PSYCH. BULL. 68,1967.=+N, C.R., BEACH, L.R., M V 65
PSYCH. REVIEW 78(2),1971.=+RNING DESPITE IDEAL V 62
RAND. CORP. ,1965.= HARRIS, E.T., A GUIDE II07
SCARECROW, 1964.=+ER, F., INDEXING THEORY, INDE III3
SCI. AMER. 217(6),1967.= COHEN, P.J., HERS IV08
SCI. AMER. 224(3),1971.= TRIBUS, M., MACIRVIN IV47
SCIENCE 172,1971.= GEB IV25
SDC REPORT # TM-[illegible]8/000/00, 1966.=+RELEVANCE J V 31
SDC REPORT # TM-[illegible]20/001-003,1967.=+UDIES OF RE V 23
SIMON & SCHUSTER, 1948.= V 48
SLA, 1967.= LE V 30
SLA, 1967.= PRACTICAL V 30
SPARTAN BOOKS, 1965.=+ TECHNICAL PRECONDITIONS IV21
SPARTAN BOOKS, 1965.=+FISHER, S., PRIMIGENIAL I IV21
SPARTAN BOOKS, 1965.= HEILPRIN, L., ED V 24
SPARTAN BOOKS, 1965.=+ RETRIEVAL AS A RECEIVER- V 24
SPARTAN, 1962.= COSTELLO, J.C., INDEXING IV57
SPARTAN, 1962.= HOWERTON, P.W. (ED), INFORMA IV57
THE CHEMICAL SOCIETY (G.B.), 1970.=+SED CURRENT V 71
THE SOCIAL IMPACT OF INFO. RETR. ,1970.=+AND PR II04
THE UNIV. OF MARYLAND, 1971.=+ ACM INFORMATION II02
UNIV. OF ILLINOIS PRESS, 1964.=+, W., THE MATHE IV02
UNPUBLISHED MANUSCRIPT, 1971.=+ COMPUTER ASSIST IV13
USA STANDARDS INST. ,1968.= USA STAND II06
WESTERN RESERVE UNIV., 1963.=+S, A.M., INFORMAT V 02
ARMITAGE, J.E., LYNCH, M.F., ARTICULATION IN TH IV54
AUSTIN, G.A., A STUDY OF THINKING, JOHN WILEY, IV01
AUSTIN, G.A., A STUDY OF THINKING, JOHN WILEY, V 54
BACH, E. (ED), HARMS, R. (ED), UNIVERSALS IN LI IV59

HOLT, RINEHART & WINSTON, 1969+
ASURE OF RETRIEVAL EFFECTIVENES+
NFORMATION RETRIEVAL SYSTEMS, +
AMETERS, SPARTAN, 1962.=
TISATION DES RECHERCHES DOCUMEN+
SCIENCE AND TECHNOLOGY, ENCYC+
JUDGMENTS FINAL REPORT, SDC +
JUDGMENTS SECONDPROGRESS REPOR+
GHT: SOME OLD IDEAS AND RECENT +
PHILOSOPHIC LIB., 1959.=

RINEHART & WINSTON, 1938.=
ANCE, PROC. ASIS + ZUNDE, P.,
AS AN AID TO DECISION MAKING, +
NSEQUENCES FOR I+ YOVITS, M.C.,
N AND CONTROL, ARTIST, 1969.
N AND CONTROL, ARIST, 1969.=
BUTTERWORTHS, 1965.=
YSIS.=
ART & WINSTON, 1968.=
ORKERS, HAFNER, 1946.=
ETRIEVAL, SPARTAN+ GREMS, M.,
BROOKS/COLE, 1968.=
N IN CHEMISTRY, ANGEWANDTE CH+
RECHERCHES DOCUMEN+ CROS, R.C.,
ENTICE-HALL, 1967.=
ENTICE-HALL, 1967.=
2,1971.=
+MIKHAILOV, A.I., CHERNI, A.I.,
. STORAGE & RET. 2,1964.=

COOK, W.A., INTRODUCTION TO TAGMEMIC ANALYSIS, IV62
COOPER, W.S., EXPECTED SEARCH LENGTH-A SINGLE ME V 21
COOPER, W.S., ON DERIVING DESIGN EQUATIONS FOR I V 70
COSTELLO, J.C., INDEXING IN DEPTH: PRACTICAL PAR IV57
CROS, R.C., GARDIN, J.C., LEVERY, F., L'AUTOMA IV58
CUADRA, C.A. (ED), ANNUAL REVIEW OF INFORMATION I 14
CUADRA, C.A., EXPERIMENTAL STUDIES OF RELEVANCE V 23
CUADRA, C.A., EXPERIMENTAL STUDIES OF RELEVANCE V 31
DEGROOT, A.D., PERCEPTION AND MEMORY VERSUS THOU IV41
DESAUSSURE, F., COURSE IN GENERAL LINGUISTICS, IV28
DEWEY, J., HOW WE THINK, HEATH, 1933.= V 50
DEWEY, J., LOGIC: THE THEORY OF INQUIRY, HOLT, V 47
DEXTER, M.E., FACTORS AFFECTING INDEXING PERFORM V 25
ERNST, R.L., YOVITS, M.C., INFORMATION SCIENCE V 46
ERNST, R.L., GENERALIZED INFORMATION SYSTEMS: CO IV31
FAIRTHORNE. R.A., CONTENT ANALYSIS, SPECIFICATIO IV29
FAIRTHORNE, R.A., CONTENT ANALYSIS, SPECIFICATIO V 10
FAIRTHORNE, R.A., TOWARD INFORMATION RETRIEVAL, IV18
FEIGEL, SELLARS, READINGS IN PHILOSOPHICAL ANAL V 37
FILLMORE, C.J., THE CASE FOR CASE, HOLT, RINEH IV59
FISHER, R.A., STATISTICAL METHODS FOR RESEARCH W IV37
FISHER, S., PRIMIGENIAL INDEXING FOR HEURISTIC R IV21
FITTS, P.M., POSNER, M.I., HUMAN PERFORMANCE, V 40
FUGMANN, R., THEORETICAL ASPECTS OF COMMUNICATIO I 07
GARDIN, J.C., LEVERY, F., L'AUTOMATISATION DES IV58
GARFINKEL, H., STUDIES IN ETHNOMETHODOLOGY, PR IV06
GARFINKEL, H., STUDIES IN ETHNOMETHODOLOGY, PR I 12
GEBALLE, R., MATTERS OF CONCINNITY, SCIENCE 17 IV25
GILYAREVSKII, R.S., INFORMATICS, FID PUBL. 43+ I 06
GOFFMAN, W., ON RELEVANCE AS A MEASURE, INFORM V 28

SUBJECTIVE PROBA+ SCHUM, D.A., GOLDSTEIN, I.L., HOWELL, W.C., SOUTHARD, J.F., V 66
EVALUATION OF INFORMATION SYSTE+ G, JOHN WILEY,+ BRUNER, J.S., GOOD, I.J., THE DECISION-THEORY APPROACH TO THE V 20
G, JOHN WILEY,+ BRUNER, J.S., GOODNOW, J.J., AUSTIN, G.A., A STUDY OF THINKIN V 54
AND THE COMMUNITY OF DISCIPLINE+ AMER. DOC. 19(1),1968.= GOODNOW, J.J., AUSTIN, G.A., A STUDY OF THINKIN IV01
AMER. DOC. 19(1),1968.= GORN, S., THE COMPUTER AND INFORMATION SCIENCES I 10
HEURISTIC RETRIEVAL, SPARTAN+ GRAZIANO, E.E., ON A THEORY OF DOCUMENTATION, I 15
, MCGRAW-HILL, 1967.= GRAZIANO, E.E., ON A THEORY OF DOCUMENTATION, IV17
NG INQUIRY: A REINTERPRETATION + GREMS, M., FISHER, S., PRIMIGENIAL INDEXING FOR IV21
HOLT, RINEHA+ BACH, E. (ED), GUILFORD, J.P., THE NATURE OF HUMAN INTELLIGENCE V 51
EXES, RAND. CORP. ,1965.= HARMON, G., INFORMATION NEED TRANSFORMATION DURI V 43
SPARTAN BOOKS, 1965.= HARMS, R. (ED), UNIVERSALS IN LINGUISTIC THEORY, IV59
AD-136-477, 1957.= HARRIS, E.T., A GUIDE FOR THE PREPARATION OF IND II07
IB., 1953.= HEILPRIN, L., EDUCATION FOR INFORMATION SCIENCE, V 24
MS A SELECTED BIBLIOGRAPHY WIT+ HEILPRIN, L.B., MATHEMATICAL MODEL OF INDEXING, III4
. 217(6),1967.= COHEN, P.J., HEISENBERG, W., NUCLEAR PHYSICS, PHILOSOPHIC L IV04
RELATIONS AMON+ BERNIER, C.L., HENDERSON, M.M., EVALUATION OF INFORMATION SYSTE V 11
R. DOC. 15(1),1964.= HERSH, R., NON-CANTORIAN SET THEORY, SCI. AMER IV08
SCH , D.A., GOLDSTEIN, I.L., HEUMANN, K.F., CORRELATIVE INDEXES-III. SEMANTIC V 45
PRI IPLES, SPARTAN, 1962.= HILLMAN, D.J., THE NOTION OF RELEVANCE-(I) AMF V 05
DOCUMENT RETRIEVAL SYSTEM, PB+ HOWELL, W.C., SOUTHARD, J.F., SUBJECTIVE PROBAB V 66
D SEARCH DEVICES, SCARECROW, + HOWERTON, P.W. (ED), INFORMATION HANDLING: FIRST IV57
TERMINOLOGY AND INDEXING METHO+ IDE, E.R.C., RELEVANCE FEEDBACK IN AN AUTOMATIC V 22
ED THEORY OF INDEXING, AD-1+ JONKER, F., INDEXING THEORY, INDEXING METHODS AN III3
64.= JONKER, F., OUTLINE OF A GENERAL THEORY OF INDEX III2
EN TECHNICAL INFORMATION TO CHE+ JONKER, F., THE DESCRIPTIVE CONTINUUM, A GENERAL III1
N OF AN EXPERIME+ BARKER, F.H., KAPLAN, A., THE CONDUCT OF INQUIRY, CHANDLER, III4
EY, 1965.= KEGAN, D.L., MEASURES OF THE USEFULNESS OF WRITT V 32
1969.= KENT, A.K., VEAL, D.C., REPORT ON THE EVALUATIO V 71
KLEINMUNTZ, B. (ED), PROBLEM SOLVING, JOHN WIL IV41
KNIGHT, G.N., TRAINING IN INDEXING, MIT PRESS, III1

ION NEED, JASIS 19(2),1968.=
SION-RECALL SQUARE, JASIS 21(+
VAL AS A RECEIVER-CONTROLLED CO+
R-CONTROLLED CO+ PAISLEY, W.J.,
CORNELL UNIV. PRESS, 1940.=
ERGAMON PRESS, 1970.=
P. TE+ LANDRY, B.C., MEARA, N.,
E STATISTICIAN, PSYCH. BULL. +
DINATE INDEX: A NEW APPROACH FO+
FORMATION RETRIEVAL SYSTEMS, +
1968.= FITTS, P.M.,
ER: A SIMULATION PROGRAM AND TH+
AGE, WESTERN RESERVE UNIV., 1+
RANKLIN, 1961.=

IVE SYSTEMS LABORATORY: A REVIE+
THE 1971 ACM INFO+ MINKER, J.,
NTROPY IN PHYSICS, CISRC, 70-+
ATION STORAGE, SEARCH AND RETRI+
+., MEARA, N., PEPINSKY, H.B.,

ETRIEVAL, THE SOCIAL IMPACT O+
SED RETRIEVAL S+ COLOMBO, D.C.,
, 1948.=
RBACH, 1970.=
ND RETRIEVAL, MCGRAW-HILL, 19+
FORMATION SCIENCE, CASE WESTE+

CHEMIST'S POINT OF VIEW, J. C+
OUTHARD, J.F., SUBJECTIVE PROBA+

O'CONNOR, J., SOME QUESTIONS CONCERNING INFORMAT V 33
O'HARA, F.M., THE CORNERS AND EDGES OF THE PRECI V 16
PAISLEY, W.J., PARKER, E.B., INFORMATION RETRIE V 24
PARKER, E.B., INFORMATION RETRIEVAL AS A RECEIVE V 24
PAULING, L.C., THE NATURE OF THE CHEMICAL BOND, IV24
PEPINSKY, H.B. (ED), PEOPLE AND INFORMATION, P IV31
PEPINSKY, H.B., RUSH, J.E., YOUNG, C.E., A COM IV13
PETERSON, C.R., BEACH, L.R., MAN AS AN INTUITIV V 65
PETRARCA, A.E., LAY, W.M., THE DOUBLE-KWIC COOR IV56
POLLOCK, S.M., MEASURES FOR THE COMPARISON OF IN V 18
POSNER, M.I., HUMAN PERFORMANCE, BROOKS/COLE, V 40
QUILLIAN, M.R., THE TECHABLE LANGUAGE COMPREHEND V 44
REES, A.M., INFORMATION NEEDS AND PATTERNS OF US V 02
REID, E.E., INVITATION TO CHEMICAL RESEARCH, F IV52
REITMAN, W.A., COGNITION AND THOUGHT.= V 56
RICHMOND, P.A., THE FINAL REPORT OF THE COMPARAT V 03
ROSENFELD, S., INTRODUCTION AND PERSPECTIVES FOR II02
ROTHSTEIN, J., INFORMATIONAL GENERALIZATION OF E IV49
ROTHSTEIN, J., TOWARD A GENERAL THEORY OF INFORM IV51
RUSH, J.E., YOUNG, C.E., A COMPUTER ASSISTED S+ IV13
RUSH, J.E., A NEW APPROACH TO INDEXING.= IV65
RUSH, J.E., THEORY AND PRACTICE IN INFORMATION R II04
RUSH, J.E., USE OF WORD FRAGMENTS IN COMPUTER-BA V 68
RUSSELL, B., HUMAN KNOWLEDGE, SIMON & SCHUSTER V 48
SACKMAN, H., MAN-COMPUTER PROBLEM SOLVING, AUE V 63
SALTON, G., AUTOMATIC INFORMATION ORGANIZATION A V 17
SARACEVIC, T., ON THE CONCEPT OF RELEVANCE IN IN V 57
SCHLICK, M., MEANING AND VERIFICATION.= V 37
SCHMERLING, L., CHEMICAL INDEXING: THE RESEARCH IV45
SCHUM, D.A., GOLDSTEIN, I.L., HOWELL, W.C., S V 66

ECHNICAL INFORMATION TO CHEMICAL RESEARCHERS, JASIS 21(3),1970.=+SS OF WRITTEN T V 32
+.J., PARKER, E.B., INFORMATION RETRIEVAL AS A RECEIVER-CONTROLLED COMMUNICATION+ V 24
+ ON TECHNICAL PRECONDITIONS FOR RETRIEVAL CENTER OPERATIONS, SPARTAN BOOKS, 19+ IV21
+ARCH LENGTH-A SINGLE MEASURE OF RETRIEVAL EFFECTIVENESS BASED ON THE WEAK ORDERI+ V 21
T, M.C., STUDIES IN INDEXING AND RETRIEVAL EFFECTIVENESS, PB-174-395, 1967.=+RAN V 26
A., EFFECTIVENESS OF INFORMATION RETRIEVAL METHODS, AMER. DOC. 20(1),1969.=+, J. V 19
+971 ACM INFORMATION STORAGE AND RETRIEVAL SYMPOSIUM, THE UNIV. OF MARYLAND, 19+ II02
EEDBACK IN AN AUTOMATIC DOCUMENT RETRIEVAL SYSTEM, PB-184-246, 1969.=+ELEVANCE F V 22
VICKERY, B.C., ON RETRIEVAL SYSTEMS THEORY, ARCHON, 1968.= III7
A+ LANCASTER, F.W., INFORMATION RETRIEVAL SYSTEMS CHARACTERISTICS,TESTING AND EV V 15
D ON THE WEAK ORDERING ACTION OF RETRIEVAL SYSTEMS, AMER. DOC. 19(1),1968.=+BASE V 21
OR THE COMPARISON OF INFORMATION RETRIEVAL SYSTEMS, AMER. DOC. 19(4),1968.=+ES F V 18
WORD FRAGMENTS IN COMPUTER-BASED RETRIEVAL SYSTEMS, J. CHEM. DOC. 9(4),1969.=+F V 68
DESIGN EQUATIONS FOR INFORMATION RETRIEVAL SYSTEMS, JASIS 21(6),1970.=+DERIVING V 70
LANCASTER, F.W., INFORMATION RETRIEVAL SYSTEMS, JOHN WILEY, 1968.= IV42
INFORMATION STORAGE, SEARCH AND RETRIEVAL.=+STEIN, J., TOWARD A GENERAL THEORY OF IV51
THORNE, R.A., TOWARD INFORMATION RETRIEVAL, BUTTERWORTHS, 1965.= FAIR IV18
ILISTIC INDEXING AND INFORMATION RETRIEVAL, J. ACM 7,1960.=+ON RELEVANCE, PROBAB IV43
TIC INFORMATION ORGANIZATION AND RETRIEVAL, MCGRAW-HILL, 1968.=+LTON, G., AUTOMA V 17
IMIGENIAL INDEXING FOR HEURISTIC RETRIEVAL, SPARTAN BOOKS, 1965.=+FISHER, S., PR IV21
+ORY AND PRACTICE IN INFORMATION RETRIEVAL, THE SOCIAL IMPACT OF INFO. RETR. ,1+ II04
NCYC+ CUADRA, C.A. (ED), ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY, E I 14
OMPARATIVE SYSTEMS LABORATORY: A REVIEW, JASIS 21(2),1970.=+INAL REPORT OF THE C V 03
+D, J.F., SUBJECTIVE PROBABILITY REVISIONS UNDER SEVERAL COST-PAYOFF ARRANGEMENTS+ V 66
LER, M.T., INDEXING: PRINCIPLES, RULES AND EXAMPLES, N.Y. STATE LIB. ,1957.=+HEE II09
+ AND NON LEARNING DESPITE IDEAL S-R-REINFORCEMENT CONTINGENCIES, PSYCH. REVIEW+ V 62
1956.= BRILLOUIN, L., SCIENCE AND INFORMATION THEORY, ACADEMIC PRESS, V 38
1956.= BRILLOUIN, L., SCIENCE AND INFORMATION THEORY, ACADEMIC PRESS, IV03
+), ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY, ENCYCLOPAEDIA BRITANNI+ I 14
+.L., YOVITS, M.C., INFORMATION SCIENCE AS AN AID TO DECISION MAKING, CISRC, 6+ V 46

RTHORNE, R.A., CONTENT ANALYSIS, SPECIFICATION AND CONTROL, ARIST, 1969.= FAI V 10
RTHORNE. R.A., CONTENT ANALYSIS, SPECIFICATION AND CONTROL, ARTIST, 1969. FAI IV29
TAULBEE, O.F., CONTENT ANALYSIS, SPECIFICATION AND CONTROL, ARIST, 1968.= II01
BAXENDALE, P., CONTENT ANALYSIS, SPECIFICATION AND CONTROL, ARIST, 1965.= II13
ND EDGES OF THE PRECISION-RECALL SQUARE, JASIS 21(2),1970.=+ F.M., THE CORNERS A V 16
ARDS INST. ,1968.= USA STANDARD BASIC CRITERIA FOR INDEXERS, USA STAND II06
C. 3(1),1962.= MARKUS, J., STATE OF THE ART OF PUBLISHED INDEXES, AMER. DO II03
+NS, M.E., AUTOMATIC INDEXING A STATE-OF-THE-ART REPORT, NAT. BUR. OF STANDARD+ V 04
R, 1946.= FISHER, R.A., STATISTICAL METHODS FOR RESEARCH WORKERS, HAFNE IV37
BEACH, L.R., MAN AS AN INTUITIVE STATISTICIAN, PSYCH. BULL. 68,1967.=+N, C.R., V 65
BACK, S., INFORMATION THEORY AND STATISTICS, JOHN WILEY, 1959.= KULL IV38
+ES FOR THE 1971 ACM INFORMATION STORAGE AND RETRIEVAL SYMPOSIUM, THE UNIV. OF + II02
A GENERAL THEORY OF INFORMATION STORAGE, SEARCH AND RETRIEVAL.=+STEIN, J., TOWARD IV51
MACHINE DESCRIPTION FOR CHEMICAL STRUCTURES, J. CHEM. DOC. 5 ,1967.=+A UNIQUE IV64
NC. 1965.= TAUBE, M., STUDIES IN COORDINATE INDEXING, DOCUMENTATION I II10
7.= GARFINKEL, H., STUDIES IN ETHNOMETHODOLOGY, PRENTICE-HALL, 196 IV06
7.= GARFINKEL, H., STUDIES IN ETHNOMETHODOLOGY, PRENTICE-HALL, 196 I 12
PB-174-395+ ST.LAURANT, M.C., STUDIES IN INDEXING AND RETRIEVAL EFFECTIVENESS, V 26
POR+ CUADRA, C.A., EXPERIMENTAL STUDIES OF RELEVANCE JUDGMENTS SECONDPROGRESS RE V 31
DC + CUADRA, C.A., EXPERIMENTAL STUDIES OF RELEVANCE JUDGMENTS FINAL REPORT, S V 23
+NG AND GUESSING: A QUANTITATIVE STUDY OF INFERENCE AND INFORMATION, JOHN WILEY+ V 39
+OUNG, C.E., A COMPUTER ASSISTED STUDY OF INFORMATIVE DISPLAY, UNPUBLISHED MANU+ IV13
GOODNOW, J.J., AUSTIN, G.A., A STUDY OF THINKING, JOHN WILEY, 1956.=+R, J.S., IV01
GOODNOW, J.J., AUSTIN, G.A., A STUDY OF THINKING, JOHN WILEY, 1956.=+R, J.S., V 54
+TICULATION IN THE GENERATION OF SUBJECT INDEXES BY COMPUTER, J. CHEM. DOC. 7 + IV54
+ HOWELL, W.C., SOUTHARD, J.F., SUBJECTIVE PROBABILITY REVISIONS UNDER SEVERAL C+ V 66
N THEORY: PROCEEDINGS OF THE 3RD SYMPOSIUM, BUTTERWORTHS, 1955.=+ C., INFORMATIO IV15
NFORMATION STORAGE AND RETRIEVAL SYMPOSIUM, THE UNIV. OF MARYLAND, 1971.=+ ACM I II02
NTAIRES --UN MODEL GENERALE-- LE SYNTOL, GAUTIER-VILLARS, 1964.=+CHERCHES DOCUME IV58
AN AUTOMATIC DOCUMENT RETRIEVAL SYSTEM, PB-184-246, 1969.=+ELEVANCE FEEDBACK IN V 22