SELECTED ARTIFICIAL INTELLIGENCE TECHNIQUES IN INFORMATION RETRIEVAL SYSTEMS RESEARCH.

SMITH, LINDA CHERYL
DEGREE DATE: 1979
**NOTICE:** Return or renew all Library Materials! The Minimum Fee for each Lost Book is $50.00.

The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University. To renew call Telephone Center, 333-8400

<table>
<thead>
<tr>
<th>UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN</th>
</tr>
</thead>
<tbody>
<tr>
<td>L161—O-1096</td>
</tr>
</tbody>
</table>
This is an authorized facsimile, made from the microfilm master copy of the original dissertation or masters thesis published by UMI.

Prior to publishing, UMI microfilms the original manuscript and returns it to the author or institution granting the degree. When an order is placed, the complete document is reproduced, on paper or in microform, from the master film copy. This is called on-demand publishing.

The bibliographic information for this thesis is contained in UMI's Dissertation Abstracts database, the only central source for accessing almost every doctoral dissertation accepted in North America since 1861.

UMI Dissertation Information Service

University Microfilms International
A Bell & Howell Information Company
300 N. Zeeb Road, Ann Arbor, Michigan 48106
800-521-0600 OR 313/761-4700

Printed in 1987 by xerographic process on acid-free paper
INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in "sectioning" the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.
SMITH, LINDA CHERYL
SELECTED ARTIFICIAL INTELLIGENCE TECHNIQUES
IN INFORMATION RETRIEVAL SYSTEMS RESEARCH.
SYRACUSE UNIVERSITY, PH.D., 1979

COPYR. 1979 SMITH, LINDA CHERYL
University
Microfilms
International 300 N. ZEEB ROAD, ANN ARBOR, MI 48106

© Copyright 1979
LINDA CHERYL SMITH
SELECTED ARTIFICIAL INTELLIGENCE TECHNIQUES IN INFORMATION RETRIEVAL SYSTEMS RESEARCH

by

LINDA CHERYL SMITH

DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Information Transfer in the Graduate School of Syracuse University

MAY 1979

Approved

Date 27 April 1979
PREFACE

Lesen heisst borgen, daraus erfinden abtragen.

To read means to borrow; to create out of one's readings is paying off one's debts.
- Georg Christoph Lichtenberg

The references in the bibliography document my debt to the many authors cited. In addition I want to thank the individuals who have contributed to this work in various ways: F. Wilfrid Lancaster, for introducing me to information retrieval; Michael D. Kelly, for introducing me to artificial intelligence and prompting my early study of artificial intelligence in information retrieval; Jeffrey Katzer, for suggesting that I embark on the study which led to Chapters III-VII; Terry Noreault, for writing the computer programs to gather the data on which the study reported in Chapter VIII is based; Pauline Atherton, for providing access to the data on which the study reported in Chapter IX is based; Edward Storm, for asking questions which led me to think about certain issues in new ways; and Michael McGill, for the advice, encouragement, and criticism which helped me to complete all ten chapters. Finally, I must acknowledge my parents' patience and support throughout my years in graduate school.
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREFACE</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>A. Historical background</td>
<td>1</td>
</tr>
<tr>
<td>B. Definitions</td>
<td>4</td>
</tr>
<tr>
<td>C. New technologies</td>
<td>10</td>
</tr>
<tr>
<td>D. Models of man</td>
<td>15</td>
</tr>
<tr>
<td>E. Limits of discussion</td>
<td>18</td>
</tr>
<tr>
<td>II. RATIONALE AND PROBLEM STATEMENT</td>
<td>20</td>
</tr>
<tr>
<td>A. Information retrieval systems: basic elements</td>
<td>20</td>
</tr>
<tr>
<td>B. Artificial intelligence concepts</td>
<td>24</td>
</tr>
<tr>
<td>1. Pattern recognition</td>
<td>24</td>
</tr>
<tr>
<td>2. Representation</td>
<td>26</td>
</tr>
<tr>
<td>3. Problem solving</td>
<td>27</td>
</tr>
<tr>
<td>4. Learning</td>
<td>28</td>
</tr>
<tr>
<td>C. Method of investigation</td>
<td>29</td>
</tr>
<tr>
<td>1. Model building</td>
<td>31</td>
</tr>
<tr>
<td>2. Testing</td>
<td>33</td>
</tr>
<tr>
<td>D. Problem statement</td>
<td>37</td>
</tr>
<tr>
<td>III. PATTERN RECOGNITION</td>
<td>44</td>
</tr>
<tr>
<td>A. Automatic indexing as a feature selection problem</td>
<td>45</td>
</tr>
<tr>
<td>1. Information retrieval approaches</td>
<td>46</td>
</tr>
<tr>
<td>2. Artificial intelligence approaches</td>
<td>52</td>
</tr>
<tr>
<td>3. Problems for research</td>
<td>56</td>
</tr>
</tbody>
</table>
9. Pattern classification ........................................ 61
1. Information retrieval approaches ............................ 62
2. Artificial intelligence approaches ........................... 63
3. Problems for research ........................................ 64

C. Measures of similarity ....................................... 67
1. Information retrieval approaches ............................ 68
2. Artificial intelligence approaches ........................... 69
3. Problems for research ........................................ 70

IV. REPRESENTATION ............................................. 74
A. Internal representations ....................................... 74
1. Information retrieval approaches ............................ 75
2. Artificial intelligence approaches ........................... 76
3. Problems for research ........................................ 77

3. External representations ...................................... 91
1. Information retrieval approaches ............................ 92
2. Artificial intelligence approaches ........................... 94
3. Problems for research ........................................ 95

V. PROBLEM SOLVING ............................................ 97
A. Question answering as a theorem proving problem ........ 98
1. Information retrieval approaches ............................ 98
2. Artificial intelligence approaches ........................... 100
3. Problems for research ........................................ 105

9. Heuristics in IR ............................................. 110
1. Information retrieval approaches ............................ 111
2. Artificial intelligence approaches ........................... 115
3. Problems for research ........................................ 120
E. Suggestions for further research .............................................................. 183
   1. Representations considered singly ...................................................... 185
   2. Representations in combination ......................................................... 186
   3. Conclusion ......................................................................................... 189

IX. QUERY FORMULATION AS PROBLEM REDUCTION ................................. 190
   A. Process models .................................................................................. 190
      1. Inquirer-intermediary interaction .................................................... 191
      2. Man-machine interaction ................................................................ 194
   B. Problem reduction model .................................................................. 196
      1. Related models ............................................................................ 198
      2. Model development ...................................................................... 200
   C. Data collection ................................................................................ 201
      1. Content analysis .......................................................................... 201
      2. Content analysis category development ........................................ 203
      3. Content analysis category revision .............................................. 205
      4. Reliability analysis ...................................................................... 208
   D. Data analysis .................................................................................. 212
      1. Protocol analysis .......................................................................... 212
      2. Subproblem I ............................................................................. 218
      3. Subproblem II ............................................................................ 221
      4. Subproblem III ........................................................................... 227
      5. Subproblem IV ............................................................................ 233
      6. Subproblem V ............................................................................. 236
      7. Subproblem VI ............................................................................ 242
      8. Tutorials ...................................................................................... 246
      9. Sequence of subproblem solution .................................................. 248
<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Specificity values for 7 representations</td>
<td>173</td>
</tr>
<tr>
<td>2</td>
<td>Exhaustivity values for 7 representations</td>
<td>174</td>
</tr>
<tr>
<td>3</td>
<td>Number of documents retrieved for 7 representations</td>
<td>176</td>
</tr>
<tr>
<td>4</td>
<td>Overlap in sets of items retrieved for pairs of representations</td>
<td>179</td>
</tr>
<tr>
<td>5</td>
<td>Proportions for reliability analysis</td>
<td>209</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Model of an information retrieval system</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>Stages in pattern recognition</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>Artificial intelligence in information retrieval: analysis and studies</td>
<td>39</td>
</tr>
<tr>
<td>4</td>
<td>Schematic of an IR system</td>
<td>41</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

A. Historical background

There have been signs in the past decade that the rather artificial separation of disciplines may be coming to an end. It is no longer a point of honor for each to demonstrate its absolute independence of the others, and new interests have emerged that permit the classical problems to be formulated in novel and occasionally suggestive ways. (Chomsky, 1972, p. 1)

The fields of artificial intelligence and information retrieval share a common interest in developing more capable computer systems. This study reports the results of research which has explored possible contributions of artificial intelligence (AI) to the design of information retrieval (IR) systems. The objective has been to determine how and to what extent concepts and techniques developed in the study of artificial intelligence problems can be applied in information retrieval.

In order to put the discussion of a proposed alliance in proper perspective, it is useful to examine briefly the separate origins of the two fields. Each found an early spokesman to catalyze interest: Vannevar Bush in information retrieval and A. M. Turing in artificial intelligence.

Although the term "information retrieval" was actually coined by Moore in 1950 (Moore, 1959), Bush's paper "As We May Think" published in 1945 indicated ways in which existing photographic and electronic techniques, together with reasonable extrapolations, might be applied to improve recording and retrieval of research results. Bush suggested:

Consider a future device for individual use, which
is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. (pp. 106-107)

In his paper Bush considered what machines could do, recognizing that machines were available which could extend man's mental powers rather than being limited to an extension of man's physical powers as had been the case in the past. Bush was concerned that the growing volume of literature would not be well used, for "the prime action of use is selection" (p. 105), a process inhibited "by the artificiality of systems of indexing" (p. 106). One approach, simple selection, proceeds by examining in turn every one of a large set of items, picking out those which have certain specified characteristics. But this is a tedious process when there are many items. Bush suggested an alternative:

[Memex] affords an immediate step... to associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing. (p. 107)

Bush thus exhibited a belief that basic processes of information retrieval could be reduced to machine processes, at the same time improving the ways in which man stored and retrieved information. This belief shared by others has been a motivation for much subsequent research and development in information retrieval systems.

In his paper "Computing Machinery and Intelligence" published in 1950, Turing addressed the question, "Can machines think?" He replaced an attempt to define meanings of the terms "machine" and "think" with a related question expressed in relatively unambiguous words: "What
will happen when a machine takes the part of A in this game?" The
game referred to is the imitation game in which there are three par-
ticipants: the machine (A), a human (B), and an interrogator. The
object of the game for the interrogator is to determine which of the
other two is human and which is the machine. It is the machine's objec-
tive to cause the interrogator to make the wrong identification. This
experiment is commonly called "Turing's test". Turing (1950) further
specified:

Let us fix our attention on one particular digital
counter C. Is it true that by modifying this com-
puter to have an adequate storage, suitably increasing
its speed of action, and providing it with an appro-
riate programme, C can be made to play satisfactorily
the part of A in the imitation game, the part of B
being taken by a man? (p. 442)

Such a role for a computer would require a library of computer programs
held in store in such a way that the machine could respond to human
interrogation in a coherent and resourceful fashion. Just as we do not
yet have a memex, we have yet to program a machine which could adequately
play the role of A in Turing's test. The two papers do, however, serve
to focus on some of the points of contact between IR and AI. Bush's
interest in selection by association introduces the concepts of repre-
sentation and search. Turing's emphasis on natural language as a medium
of communication introduces the task of developing programs capable of
processing natural language. The initial link between AI and IR is a
common interest in machines as question-answers.

Strictly speaking, there is nothing novel about machine-aided
intelligence. The papyrus, printing, and the library have long fortifed
man's memory (Elithorn & Jones, 1973). Librarians are concerned with
organization of man's "external memory"—the books, journals, reports,
etc., which contain recorded information. The advent of the computer has led to new approaches: librarians now make use of online bibliographic retrieval systems, accessing computers via terminals to retrieve citations from machine-readable data bases. In 1965 Licklider characterized the computer-based "pro cognitive system" as a possible library of the future (p. 6). Criteria for a procognitive system include the ability to (pp. 36-39):

1. Handle both documents and facts.

2. Make available a body of knowledge that is organized both broadly and deeply, and foster the improvement of such organization through use.

3. Handle heuristics (guidelines, strategies, tactics, and rules of thumb intended to expedite solution of problems) coded in such a way as to facilitate their association with situations to which they are germane.

4. Facilitate its own further development by providing tool-building languages and techniques to users and preserving the tools they devise and by recording measures of its own performance and adapting in such a way as to maximize the measures.

These criteria begin to specifically identify possible relations between AI and IR, since they pertain to such basic issues as information representation (item 2), heuristics (item 3), and learning (item 4).

8. Definitions

Just as Turing was faced with the problem of defining "machine" and "think", one must precede a discussion of artificial intelligence in information retrieval systems with definitions of what is meant by "artificial intelligence" and "information retrieval" for the purposes of this study. Usage of the terms varies widely in the literature.

"Information retrieval" may be used in the following senses (Minker, 1977):
1. Document retrieval—retrieval of either a document surrogate, e.g., an abstract, or the document itself which may contain information relevant to a question (also called literature searching, reference retrieval).

2. Question-answering—reply to a question requiring inference from material presented in text.

3. Data retrieval—retrieval of individual or related facts from a file where types of questions to be answered are known in advance (often handled by a data base management system).

Kochen (1974, p. 45) has suggested that the central problem of an information retrieval system is to produce an appropriate response to an inquirer's need for information. A response could in theory take any of the above forms, depending on the expressed need. The retrieval problem emerges only if the information file (whether it be lists of references, actual documents, or tables of data) is so large that it must be organized for partial search or equivalently that selection must first be made by someone or some system other than the inquirer himself.

In all retrieval systems some selection mechanism is required. The system compares the request for information with descriptions of stored items and retrieves all items which correspond in some definite way. One of the main criteria of success of such a system is the extent to which the utilized definition of correspondence between the requests and stored items yields items judged by inquirers to be relevant to their requests. An information retrieval system may be thought of as a receiver-controlled communication system (Paisley & Parker, 1965). In an online retrieval system the inquirer interacts in some way with the file of stored items to progressively modify the set accepted in the search and thus eventually to reach a subset of the file having a high probability of being of some value to the inquirer in relation to the
information need that prompted his search.

Operational information retrieval systems are generally document or data retrieval systems, while question answering systems are largely experimental. What is retrieved in the case of question answering and data retrieval may be immediately useful to the inquirer, whereas the output of a document retrieval system requires further work on the part of the inquirer to identify the desired information. The state of the art thus has systems with responses one step removed from the desired information, producing bibliographies rather than specific answers. The inquirer must locate the documents and read through them to extract the relevant information. The following table summarizes differences between document retrieval and question answering (Salton, 1968, p. 394):

<table>
<thead>
<tr>
<th>Document Retrieval</th>
<th>Question Answering</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data base consisting of items identified by keywords.</td>
<td>1. Data base consisting of items with a variety of relations specified.</td>
</tr>
<tr>
<td>2. Data base considered closed at any given instant since all potential answers are already stored.</td>
<td>2. Open-ended data base with deductive procedures capable of generating new statements (items and relations) from old ones.</td>
</tr>
<tr>
<td>3. Only one type of question permitted.</td>
<td>3. Many question types allowed.</td>
</tr>
<tr>
<td>4. Simple match of concepts assigned to search requests and documents is used to identify retrieved items.</td>
<td>4. Items satisfying relations specified by the query may first have to be deduced from the data base before answers can be retrieved.</td>
</tr>
</tbody>
</table>

Any device that is to perform the function of answering questions must either: (a) obtain the answer by consulting some file where the answer is stored or (b) synthesize the answer from available data according to a set of rules. In the latter case some data are stored in memory and other data are reconstructed as needed. The manner in which data are
organized in most information retrieval systems restricts the class of questions that can be asked to those that can be answered by explicit reference to the data base. Other question types which are answerable in principle given the data contained in the data base are not legitimate questions because the information needed for the answers is represented only implicitly, i.e. it is deducible from the data knowing properties of relations among data items. Questions which can be addressed to document retrieval systems generally must be formulated in a highly formatted language for input. This inhibits the process of establishing a dialogue between inquirer and system for negotiating answers to vaguely formulated initial questions.

Turning to a definition of "artificial intelligence", one can begin with Simon (1969, pp. 5-6) who has provided a useful elaboration of the term "artificial":

(a) Artificial things are synthesized . . . by man.

(b) Artificial things may imitate appearances in natural things while lacking, in one or many respects, the reality of the latter.

(c) Artificial things can be characterized in terms of functions, goals, adaptation.

Research in AI thus seeks to find effective machine processes (a above) for accomplishing complex tasks (c above), imitating human processes only when this proves to be the most efficient way to do the job (b above). McCarthy and Hayes (1969) divide "intelligence" into two parts:

(1) epistemological, dealing with representations of knowledge; and

(2) heuristic, dealing with the problem solving mechanisms that operate on that representation. Armer (1963) offers a broader view of intelligent behavior as an n dimensional continuum. Two dimensions, speed and complexity or sophistication of the information processes available,
provide a contrast between the capabilities of man and machine. While the machine can perform certain operations, such as computations, much more rapidly than a man, the complexity and sophistication of information processes available to a man far exceeds that currently demonstrated by machines. The goal of research in AI is to push machines further out along the dimension of complexity of information processes available. Research is thus directed toward learning how to write programs or build systems that can perform tasks which have never been performed automatically before, usually tasks that have been assumed to require human intelligence. In a recent review Nilsson (1974) has provided a useful enumeration of the current task domains in AI research. They all involve use of portions of the core, i.e. basic mechanisms of intelligence and implementational techniques: (a) modeling and representation of knowledge; (b) common sense reasoning, deduction, and problem solving; (c) techniques for heuristic search; and (d) AI systems and languages, e.g. LISP. AI is thus not about any particular mental task, but about what can be generalized including procedures initially developed on a task-oriented basis but possessing seeds of generality (Michie, 1974, p. 52).

While the goal of AI is to construct computer programs which exhibit behavior that one would call "intelligent" behavior when observed in human beings, there are two different ways to pursue that goal. These have been labeled variously as computer synthesis and computer simulation (Apter, 1970, p. 28) or performance mode and simulation mode (Weizenbaum, 1976, p. 164). Concern with synthesis or performance mode means building machines that behave intelligently, whether or not their performance sheds any light on human intelligence. Concern with simulation means modeling the way humans accomplish certain tasks rather than performing those tasks in the most efficient way the computer possibly could. The first
approach exhibits concern for ends only, while the second shows concern for means and ends. The dividing line between simulation mode and performance mode is not absolute, for often the only way to begin thinking about how to get a computer to do a specific task is to ask how people would do it. The two approaches are complementary and in IR one can learn from both since both the user and the computer are integral parts of an online retrieval system. Choice between the two modes is sometimes dictated by the current state of knowledge in a particular task domain, as suggested by the following passage from Bar-Hillel (1964a):

When, in 1951, I got myself interested in the automation of translation, I tried at first to find out what psychologists knew about human translation, only to discover to my dismay that very little was known that was not purely anecdotal or speculative. Machine simulation of human translating having consequently been discarded as one possible approach, a "let's-see-how-far-we-can-get" attitude was generally adopted. Almost from the start there was a differentiation between those who thought that fully automatic and good quality translation was a reasonable goal to aim at and those who regarded such a goal as utopian, at least for the foreseeable future, and preferred to work towards a man-machine partnership in translation, with the machine doing the routine chores and the man making the "intelligent" decisions. (pp. 180-181)

The passage highlights a number of themes which will characterize the discussion of artificial intelligence in information retrieval systems throughout this study:

(1) an interest in work in simulation mode, Models of Man as a guide to design of IR systems (p. 15);

(2) an interest in work in performance mode, better utilization of the New Technologies for IR (p. 10);

(3) AI in IR as work toward a man-machine partnership.
C. New Technologies

The motivation for the study of artificial intelligence in information retrieval systems stems directly from the development of new technologies. These include:

(1) mass storage devices with random access—capable of storing large amounts of data in machine-readable form online and retrieving selected items rapidly;

(2) machine-readable data bases—as more publications are produced with the aid of computers, large quantities of data in machine-readable form become immediately available;¹

(3) computer-communication networks—coupling computers to other computers and terminals via communication lines means that large data bases are accessible from remote locations and widespread use is promoted by the gradual reduction in communication costs.

(4) intelligent terminals—as terminals become more powerful and less expensive, more individuals will make use of them to access any public computing or information resource (Korfhage, 1978).

A number of online systems using these technologies have been developed and implemented in the past few years. Online technology both creates a need and provides an opportunity for artificial intelligence in information retrieval systems. Batch processed searches of data bases on magnetic tape require a trained intermediary to formulate search strategies to be processed by the system so there is little incentive to make the system transparent to the inquirer. In online systems, on the other hand, it is desirable to work toward more transparent systems, where the inquirer “sees through” the complexity of the system or a series of activities to the end result (Williams, 1977). A degree of transparency is achieved whenever a procedure is instituted which allows the inquirer

¹A directory including only bibliographic data bases lists over 300, a growing number of which are available for searching online (Williams & Rouse, 1976).
to accomplish in a single step what formerly required multiple discrete steps, e.g. searching several data bases simultaneously rather than in sequence. The development of transparent systems involves the delegation of more functions to the machine, a process which may be aided by the application of AI techniques.

Batch tape-based retrieval systems are characterized by: a) little possibility of browsing; b) search strategy cannot be developed heuristically; c) search must be delegated to an intermediary; d) time delay between formulation of a search request and response of the system. In contrast, online systems have the technology to permit: a) random access; b) interactive browsing, heuristic searches; c) person with a need for information can conduct his own search if he desires; d) no significant time delay (Lancaster & Fayen, 1973, pp. 2-3).

Incorporation of AI techniques in online information retrieval systems could have two related consequences: (1) better utilize computer capabilities and (2) facilitate use of systems directly by individuals who have a need for information. The capabilities of computers were not really fully utilized in batch tape-based retrieval systems. As Ziman (1969) suggests, verbal information as found in tape-based systems is a "rather passive material" which leads to inefficient use of computers.

He argues:

There is so very little that one can really do to a classified index of abstracts on tape, except to run through it now and then looking for particular items. One cannot add together, or subtract, or logically collate, or Fourier transform, or determine the convexity of, or what you may call it, such a miscellaneous set. The specific technical advantages of a computer—the power to perform vast numbers of such operations at enormous speed—are seldom exploited. (p. 322)

This observation that computers are not used to full advantage in batch-
processed systems is echoed by Lancaster (1971) who notes that the large computer systems do little more than mechanize searching principles already existing. While such systems do allow processing of larger volumes of documents and queries than would be feasible in manual systems, they provide a level of search sophistication and flexibility little better than that available in certain manual systems. In addition they rely on the human user to completely formulate search strategies to identify desired items. In particular many batch systems are characterized by:

(a) use of Boolean logic in construction of search strategies. This makes it difficult to recognize degree of match between query and document and to rank output by possible relevance.

(b) a strong adherence to human indexing using controlled vocabularies. While such a practice does promote indexing and searching consistency, the resulting surrogate is less specific than the associated document and the controlled vocabulary may be expensive to maintain and apply.

If one accepts the desirability of designing systems that can be used directly by inquirers, one should be prepared to develop techniques different from those used in offline, batch, delegated systems. As information technology develops and online services expand, it becomes increasingly clear that success of information systems is more likely to be achieved through adjusting the system to meet the specific needs of an individual rather than trying to adapt the individual inquirer to match the wholesale output of an information system (Garvey, Tomita, & Woolf, 1974). Because queries vary in specificity and complexity and inquirers differ in retrieval requirements, an essential feature of a retrieval system is flexibility. Technology should allow us to customize
searches, producing a result suited to the needs of a particular inquirer.

While the new technologies clearly create new opportunities, they also generate certain problems. The machine-readable data bases are for the most part production data bases, used in creating printed tools, and it is not clear how they should be modified for searching on-line. In networks there is a lack of uniformity among command languages, retrieval functions, indexing vocabularies, and output formats. Thus the user of an online system must currently bring to the terminal a considerable amount of background knowledge concerning the characteristics of the data base both in terms of content and organization (Artandi, 1976). This must be combined with a thorough understanding of system parameters in order to formulate queries in the most effective way. An individual interested in a topic spanning unconnected data bases must search each one separately, possibly using different search strategies in each. Online systems as currently designed are too rich, in the sense that they require too much intellectual effort to make use of all the available features. Frequently users do not exploit the power of complex search strategies and numerous access points, especially when these vary across systems (Sparck Jones, 1979). One approach to dealing with this diversity is the translating computer interface, which performs the translations necessary to allow the user to access multiple diverse IR systems in a common framework (Reintjes & Marcus, 1974). Its features include:

1. virtual system concept by which users perceive the network as a single homogeneous system;

2. common command language synthesized from a basic language of primitive IR functions;

3. master index and thesaurus which stores vocabularies of separate data bases along with index term interrelationships and counts;
(4) common bibliographic data structure in which data elements for bibliographic information are hierarchically structured and interrelated among different data bases.

Through such an interface, storage technology can be coupled with accessing technology that facilitates retrieval of the data stored.

The translating computer interface represents one experimental effort to shift the burden of searching online systems at least in part to the machine. Remote terminal availability is not enough to guarantee system use. This observation has been made in the form of certain empirical "laws" and "principles":

Mooers' Law: An information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it. (Mooers, 1960)

Principle of Least Action: The design of any future information service should be predicated on the assumption that its customers will exert minimal effort in order to receive its benefits. Furthermore they won't bother at all if the necessary minimum is higher than some fairly low threshold. (Swanson, 1966)

The bottleneck limiting widespread use of online systems is not in hardware speeds and storage capacities, but in programming and the design of the man-machine interface.

In summary, one can compare the development of retrieval system technology with the three stages in the evolution of a technology identified below (O'Connell, Fubini, McKay, Hillier, & Hollomon, 1969):

The first stage in the evolution of technologies is one in which what is being done now can be done cheaper, faster, and better with the help of that technology than without.

The second stage in the evolution of technologies occurs when one can do things that could not have been done without those technologies.

The third stage occurs when there is a change in behavior and ways of doing things to match the new capability that the new technologies give.
Existing online systems are examples of the first stage and to some degree the second. They enable one to do literature searching faster than is possible with printed tools and allow certain types of searches (e.g. on words in title and abstract) which are not feasible in printed tools. The purpose of work in performance mode is to outline in some detail how AI techniques may be used at both the conceptual and operational level to guide the evolution of this new technology of online retrieval systems in the third stage.

D. Models of Man

A scientific generation patterns its models upon its dominant metaphors. (Miller, 1978, p. xiv)

Each psychological theory which is applied to the study of human behavior has inherent in it a model of man. One of the more recent metaphors on which some models are based is the computer metaphor, where a human being is viewed as a processor of information. The computer metaphor has two parts (Mayer, 1977, p. 134):

the human-machine analogy—the human being may be viewed as a complex computer

the thinking-program analogy—the thought processes used by the human being to solve a problem may be viewed as a computer program

To better understand the model of man as an information processor, it is useful to contrast behaviorism and information processing psychology. The behaviorists stress study of overt behaviors, the responses of individuals to stimuli. They are reluctant to postulate the nature of any intervening variables between stimuli and responses, because such variables are not directly observable. In the information processing approach, on the other hand, the mental activities that occur between stimuli and responses are of central importance. The focus is knowledge—how it is acquired, modified, used, stored; in short, how it is processed.
Man as an information processor can be thought of in terms of Images and Plans, structures for knowledge and strategies for using knowledge to solve problems. A human being "builds up an internal representation, a model of the universe, a schema, a simulacrum, a cognitive map, an Image" (Miller, Galanter, & Pribram, 1960, p. 7). An Image is all the accumulated, organized knowledge that an individual has about himself and the world. The meaning of a message is interpreted as the change which it produces in the Image (Boulding, 1956, p. 7). Coupled with Images there must be Plans, for "unless you can use your Image to do something, you are like a man who collects maps but never makes a trip... A Plan is needed in order to exploit the Image" (Miller et al., 1960, p. 2). "A Plan is any hierarchical process in the organism that can control the order in which a sequence of operations is to be performed" (Miller et al., 1960, p. 16) and is essentially the same as a hierarchically organized program for a computer. A Plan need not result in overt action, since it can be used for collecting and transforming information.

The information processing view of man has recently been suggested by Belkin and Robertson (1975) as the basis for information science and hence for our approach to the design of information systems. The two basic concepts of information science which they present are (p. 201):

- A text is a collection of signs purposefully structured by a sender with the intention of changing the image-structure of a recipient.

- Information is the structure of any text which is capable of changing the image-structure of a recipient.

Studies consistent with this view are beginning to address such issues as the development of indexing languages and the design of man-machine dialogs in document retrieval systems. With respect to indexing languages Paisley (1968) observes that such systems as the Dewey Decimal
Classification, faceted classifications, and title-keyword systems "are proposed and defended on logical, rather than psychological, grounds" (p. 8). Gatty (1973) has developed procedures to reveal, analyze, and synthesize user concept systems to develop index languages for groups of workers in new interdisciplinary fields. With respect to retrieval systems Paisley (1968) observes "users are expected to adapt to information systems rather than vice versa, even when a typical adaptation is simply to ignore the system" (p. 8). In trying to reverse this situation Oddy (1977) has developed a system which uses comments from the user to try to form a better image of his region of interest before showing him additional citations. These studies are representative of a growing interest in designing systems which adapt to the needs of individuals or particular groups.

The design of an information retrieval system implicitly contains a model of its prospective users; patterns of possible access and output reflect anticipated queries and acceptable responses. Taylor (1972, p. 222) suggests some of the reasons why retrieval systems have proved unsatisfactory in the past:

(a) the relationship between a researcher and the body of relevant printed knowledge was (and is) not understood;

(b) there are very few cases where complete and exhaustive bibliographical search is necessary, yet most devices are designed on this basis;

(c) the intellectual access to most systems is so artificial and difficult that they invite neglect.

Current approaches to information systems design try to remedy some of these problems by focusing on the inquirer as an information processor and recognizing that the system can assist in the subtasks which make up the information retrieval task (e.g. formulating the query, searching
for items in the data base) in three ways (Paisley & Butler, 1977):

- **acceleration**—change the time frame in which perceptual/cognitive processes occur but not the processes themselves (e.g., by providing an index to accelerate an overview of the information store);

- **augmentation**—assisting inquirer in the substance of his task and not just in the pace of its execution (e.g., by ranking output based on similarity between documents and the query);

- **delegation**—decisions made by the system itself using programmed criteria rather than by the inquirer (e.g., decision as to which data bases to search).

In order to develop computer systems as aids in information retrieval, one needs to analyze the information retrieval process and determine to which parts of that process the computer can contribute. The purpose of work in simulation mode is to provide insight into how the computer system can be used for acceleration and augmentation in the information retrieval task and at what points delegation is appropriate.

**E. Limits of Discussion**

Before beginning a literature review, it is necessary to make explicit the limits of this discussion. Future retrieval systems may be hybrids with files expanded to include such items as pictures, diagrams, numerical data, graphs, and charts for which it will be necessary to find appropriate description languages and retrieval mechanisms. This study is limited to consideration of IR systems where the files contain only textual information, whether document surrogates or full text. The focus is online, interactive retrieval systems which can be searched either directly by the inquirer or by an intermediary. In terms of system design, the emphasis is on logical problems of IR—item representation, matching documents and requests, negotiating queries—rather than the physical problems such as details of file organization and data structures. While solutions of problems at the physical level are important
for system efficiency, attention to logical problems allows one to put more stress on what it takes to help the inquirer rather than on how to optimally utilize the machine.

As the previous section suggests, the view of the inquirer adopted here is one of man as an information processor. One must acknowledge that this single perspective does not include all variables by which one can characterize the user of an IR system. Motivation and personality differences can affect use of and satisfaction with IR systems. Since communication activities are tied to social, professional, and institutional relations and it is out of these relations that information needs often arise, study of the inquirer in a larger context can contribute to an understanding of use of online systems vs. other communication channels. The term "user study" is ambiguous, but the focus here is the inquirer at the man-machine interface of online retrieval systems. Just as computer system design can be viewed at both the logical and physical levels, the approach to the study of human thought processes can be analogistic or reductionistic. Rather than reducing processes to physiological concepts, the analogy to elements and processes of information processing machines like the computer is pursued.
CHAPTER II
RATIONALE AND PROBLEM STATEMENT

A. Information retrieval systems: basic elements

Before considering the possible relationships between AI and IR, it is necessary to review the basic components and operations of an IR system in which AI techniques may be embedded. An IR system can be characterized in a number of ways: how information items are represented within the system; how information items are accessed and manipulated; how queries are formulated and what constitutes a query; what constitutes an answer and how answers are generated. Document retrieval systems differ in certain respects from question answering systems as noted below. The model shown in Figure 1 can be used as a basis for discussing the elements of an IR system (Green, 1969a, p. 2):

![Diagram of an information retrieval system]

Figure 1. Model of an information retrieval system

Providing the interface between the user and the computer subsystem are a number of (possibly different) dialogue languages which allow the user to query the computer and give commands, and which permit the computer to express answers to questions. The translator receives queries and, if necessary, manipulates them into a form which the computer can process. If the computer can solicit information from the user to help answer questions, there may be in addition a query language employed
by the computer and an answer language employed by the user.

In a question answering system the information items to be stored in memory are facts. In responding to questions an answer can be obtained in one of two ways: either an explicit answer is found in memory or it must be inferred from the facts stored there. In document retrieval systems information items are documents and/or document surrogates. Initial processing of documents, done manually or automatically, involves the selection of features to make up document representations (surrogates). Surrogates may include such elements as author, title, abstract, index terms, and classification numbers. An answer in a document retrieval system is usually a collection of document surrogates which best match the query formulation.

Memory in a question answering system is made up of facts stored in the database. It may be indexed to simplify access. For example, if facts are stored as statements in the first order predicate calculus, memory may be indexed by predicate names, function names, and constant symbols. Question answering systems can answer questions encoded in first order logic from facts similarly encoded. Information items in the memory of a document retrieval system are often simply document surrogates. Hayes (1963) suggests that it may also be useful to somehow develop representations for queries asked in the past and for the set of inquirers so that this information may also be stored in memory and used in developing answers to new queries. Just as is the case in question answering systems, memory is often indexed to facilitate access to documents indexed by particular subject terms, written by a certain author during a particular time period, etc.

The executive component must control the process of storing and finding information items in memory and determining answers to queries.
Subcomponents of the executive can include inference mechanisms and search strategies. An inference mechanism is augmented by a search strategy to accomplish more efficient question answering. The search strategy determines where to look and the inference mechanism determines what to retrieve.

This brief discussion of system components has masked the complexity which can exist in sophisticated systems. The query may take different forms. It may be specific, asking for retrieval of a particular document or fact, or more general, asking for documents or information on some subject. Query languages may vary from inflexible, rigid formats to natural language. Translation from query languages to internal representations may range from simple transliteration of one artificial language into another that is isomorphic in structure but different in coding to the much more difficult translation of natural language sentences into the internal representation via syntactic and semantic analysis. The translator may have a separate analytic component to process queries and a generative component to produce answers in a form interpretable by the user.

Inference mechanisms may range from no inference (retrieving data given explicitly) to trivial powers of inference based on some matching or similarity criterion to finite induction or deduction. Combinations of inference mechanisms may be used. For example, statistical correlation can be applied as a first stage filter that selects from a large body of data that portion which is most likely to be related to a query. More refined deductive and inductive approaches can then be used on the resulting small (more manageable) selection of retrieved material.

The examples given on the following page illustrate some of this range of possible information retrieval systems (Colas, 1972, p. 3).
D=data  
Q=question  

<table>
<thead>
<tr>
<th>Depth of Inference Required</th>
<th>Shallow</th>
<th>Deep</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. D: John is a boy. Q: Is John a boy?</td>
<td>2. D: Axioms of group theory Q: Is the order of a finite group divisible by the order of each one of its subgroups?</td>
</tr>
<tr>
<td>Small</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Type 1 is essentially trivial and can be handled by simple matching. Type 2 is a typical theorem proving problem and type 3 is a question answering problem. Type 4 cannot be answered directly by document retrieval systems. Instead, the intent of such systems is to identify specific documents, the contents of which could be used by the inquirer to develop an answer to his query. Future information retrieval systems must address the problem of how best to facilitate answering type 4 questions. Otherwise, the value of large stores of information will not be fully realized because of the difficulty of extracting facts from them on matters for which they were not originally designed and indexed.

The above overview of IR system elements suggests that research is needed on two fronts: (1) query negotiation—obtaining the stated query which best represents the information need of the inquirer; and (2) searching—finding records which satisfy stated requirements. These correspond to what Sikléssy and Simon (1972) have identified as Stage I and Stage II. Stage I gives structure to the query, providing the interface between the query text on the one hand and symbol structures stored in memory on the other. Stage I processing is necessary to translate the query language into an acceptable internal representation. Stage II involves
use of the structured input to perform certain tasks, search and manipulation of the data base possibly using inference mechanisms to produce an answer. Within certain topical limits the system exhibits "understanding" of queries in the precise operational sense that it frequently generates correct answers to them. The division into Stages I and II is a useful one to keep in mind, for it distinguishes two appropriate domains of investigation. The first is the user-computer interface, the activities in Stage I of formulating queries in such a way that they can be processed by the computer. The second is the computer subsystem, the activities in Stage II of query processing once the query is in a structured form that the computer can handle. Investigations of AI in IR may focus on either stage separately or the two together.

8. Artificial intelligence concepts

Four AI concepts which have particular significance for IR are pattern recognition, representation, problem solving, and learning. The discussion which follows juxtaposes explanation of the AI concept with an interpretation of its use in the context of information retrieval.

1. Pattern recognition

Pattern recognition is the identification of an object with a particular set of features as the member of some class. A schematic of a device for character recognition provides a useful example of the stages in pattern recognition (Andrews, 1972, p. 3):

![Figure 2. Stages in pattern recognition](image)

Objects to be recognized must first be represented by data in machine-readable form. A transducer is used to create a digital image, encoding areas of light and dark. Since information in this form is awkward to
manipulate, a feature selector is used to pick out a small set of features (properties, attributes) with which to characterize each object. In the case of numbers and letters, for example, features could be straight and curved line segments. The classifier then assigns the object to one of a number of classes based on the output of the feature selection phase. This division of pattern recognition into stages separates the subject matter dependent aspects of a particular problem (embodied in the feature selector) and the subject matter independent classification mechanisms (mathematical decision algorithms). The latter can be subdivided according to the way classes are defined: (1) sorting, in which class criteria are given explicitly; (2) prototype matching, in which each class is illustrated by a certain number of sample objects (prototypes); or (3) clustering, in which classes are created on the basis of mutual similarity and dissimilarity in a given set of objects. All of these approaches depend on some measure of similarity, determined from the features used to characterize each object.

IR: In information retrieval development of document surrogates and query formulations may be viewed as a feature selection problem; determination of the subset of the memory to be retrieved may be viewed as a classification problem. The term "features" is a useful generic term in the context of IR as well as AI, for it permits one to think of index terms, authors, citations, etc. all as possible features. Feature selection can occur in retrieval systems at two points: when document surrogates are prepared and when queries are formulated for comparison with document surrogates. Classification of items in memory occurs in response to each query, when the portion to be retrieved is separated from that which is not. While sorting as an approach to classification requires that the inquirer specify exactly which features must occur for
an item in memory to be retrieved, the query to find all items in memory "like" one already known to the inquirer is a problem of prototype matching. The system assesses the probable relevance of a document to a query by calculating a measure of similarity between a document surrogate and the query formulation. An item is retrieved if the similarity measure is above some threshold.

2. Representation

AI. The representation is a formalism for the knowledge possessed by a system (Winograd, 1974). It may be thought of as "a set of conventions about how to describe things" (Winston, 1977, p. 15). The same knowledge may be given in alternative representations, but the ability of a system to solve the problems for which it was designed is strongly influenced by the choice of representation. Examples of representations from AI applications include: (1) predicate calculus—used to represent the data available to an automatic theorem prover; (2) semantic networks—used to represent concepts expressed by natural language words and phrases, with nodes representing concepts connected by arcs called semantic relations. These and other representation types are structures. To choose from among them for a particular application, one must also identify what type of knowledge is to be represented.

IR. Just as a representation in AI is a formalism for knowledge possessed by a system, a document representation is "a formalized statement of the nature of a document" (Vickery, 1977a, p. 113). A query formulation may be viewed similarly. When one thinks of computer-based systems rather than manual, one must ask how to take the available information and represent it in a way that the computer can store and manipulate. This includes not only representations for documents and
queries, but also relations between documents (such as citation relations) and between terms (such as those shown by a thesaurus). Design of online information retrieval systems must encompass not only internal representations, i.e., representation of information within the computer subsystem, but also external representations—the displays of information at the user-computer interface. Such displays can include, for example, dictionary terms arranged alphabetically or hierarchically.

3. Problem solving

AI. Problem solving is the art of using knowledge effectively for the attainment of desired goals (Amarel, 1970). A problem exists whenever a problem solver desires some outcome or state of affairs that he does not immediately know how to attain. One strategy of problem solving is problem reduction, partitioning a difficult problem into two or more simpler problems. In so doing the problem solver may reach subproblems of very little difficulty, the solutions of which can be combined to give the solution of the initial problem. Problem solving can be approached using either algorithms or heuristics. An algorithm is a method guaranteed to find a solution, given enough time. Where no algorithm is known or the known algorithm is impractical, one must resort to heuristics which are rules of thumb, strategies, or short cuts aiding discovery of a solution. Heuristics are not guaranteed to find optimal (or any) solutions, but often produce satisfactory solutions in a reasonable amount of time.

One often can identify heuristics applicable to particular problem solving tasks by observing the performance of people confronted with the task. In contrast to well-structured problems such as game playing situations in which initial conditions, permissible actions, and the goal are well defined, one is often confronted with ill-structured problems. Such
problems may lack explicit specification of permissible actions or the goal, for example. Before such problems can be solved, it may first be necessary to convert them into a well-structured situation.

IR. In information retrieval the problem confronting the system is to identify, in response to each query, the portion of the contents of memory which should be retrieved. In this case problem solving includes development of a search strategy and use of some inference mechanism. Application of heuristics may mean use of techniques which allow one to quickly select the subset of memory satisfying the query. In general questions which inquirers bring to retrieval systems can best be regarded as ill-structured problems which become well-structured only in the process of being changed to a form which the computer subsystem can handle. Development of online systems must include consideration of how best to design the user-computer interface to achieve conversion of initially ill-structured queries to a well-structured form.

4. Learning.

AI. System elements such as feature selection routines, representations, and heuristics are of course all initially selected and programmed by a human designer when an AI system is developed for some application. Learning mechanisms by which a system can improve its performance over time are therefore necessary so that the initial design does not circumscribe system capabilities. Basic to learning are the ability to evaluate performance so that improvement can be judged and some way to store and utilize the results of previous experience. Learning mechanisms include, for example, trial and error—keeping a record of the outcome of choosing a particular action in each situation encountered, so that actions with good outcomes can be repeated and those with bad outcomes
can be avoided when the same situation occurs in the future.

IR. The availability of online computer systems makes it reasonable to speak about dynamic systems which change and improve performance over time. Learning in retrieval systems can have either short term or long term effects. Short term learning is the modification of system response during the processing of a particular query in order to retrieve items most likely to meet the needs of the inquirer. This can be accomplished through the use of feedback in query processing, taking account of the relevance status of a sample of retrieved documents as judged by the inquirer. Long term learning could be changing the representation, by modification and/or extension, to improve system response over time. Representation modification can include changes in memory organization and in item representation, e.g. updating the database to reflect new terminology. Representation extension can include techniques for storing previous queries in a form suitable for subsequent use by other inquirers.

C. Method of investigation

Artificial intelligence is . . . an engineering discipline since its primary goal is to build things. (Nilsson, 1971, pp. vii-viii)

In studying any area of knowledge, one must be aware of the basic duality of content and process, conceptual schemes vs. methods of investigation. The description of concepts in the previous section must be accompanied by a discussion of methods. While each particular study will have its own detailed methodology appropriate for the problem to be addressed, it is possible to characterize in general terms the approach used in AI studies. Since, as Nilsson observes, AI is essentially an engineering discipline, the method of investigation follows the engineering design process. The list given on the following page enumerates the stages in this process (Beakley & Leach, 1972, chap. 17):
1. Identification of the problem
2. Collection of information
3. Generation of ideas
4. Preparation and analysis of a model
5. Testing for performance and human acceptance

A description of each of these stages as it relates to AI studies has been given in some detail by Slagle (1971) in his "advice to the potential heuristic programmer" (pp. 178-182):

1. Identification of the problem. This is of course the starting point. As Slagle observes, "opportunities and challenges abound" (p. 178).

2. Collection of information. One needs to find out what has already been done on the problem by asking knowledgeable individuals and doing a literature search. In this way one becomes familiar with the background for the chosen task and can specify the task fairly precisely.

3. Generation of ideas. This involves two parts: specifying at least one procedure to perform the task and finding one or more good representations for the data and the task. Slagle has several suggestions of sources for ideas: consider what other programmers have done on similar tasks, use introspection, observe experts performing the task, analyze the task itself, and analyze correct solutions to see how a program could determine them.

4. Preparation and analysis of a model. A programming language appropriate for the types of data structures and procedures to be used is chosen. The model builder must write and debug the program and familiarize himself with its inner workings.

5. Testing for performance and human acceptance. In order to evaluate the model, the programmer should devise a sequence of experiments
designed to test the most important hypotheses, such as those concerning the effect of using a particular heuristic. These experiments can "take the form of comparing the performance of one version of the program with another, or the performance of the program with that of an expert, 'average' man, or another program" (pp. 181-182). (Slagle's focus is performance testing rather than human acceptance. When AI techniques are included in man-machine systems, however, the latter must be evaluated as well).

As an example, consider the problem of building an automatic theorem prover for some area of mathematics. Collection of information about the task requires identification of axioms and rules of inference. Generation of ideas involves specifying procedures and representations, processes by which a computer can deduce the proof for a theorem and notations for axioms and theorems which the machine can manipulate. Preparation and analysis of a model includes encoding of data in the selected representation and programming of the theorem proving procedure using an appropriate programming language. Finally, testing could include a comparison of the performance of two different versions of the program or a comparison of the program with a human expert, measuring such variables as the time to develop a proof, number of steps in a proof, etc. To further clarify the design process, stages 4 and 5 (model building and testing) are explored in greater detail below.

1. Model building

A model is a simplification, a kind of idealized representation of what it is intended to model. A number of taxonomies for models can be developed, including the following categories (Crosson & Sayre, 1963):

(a) replicates--reproduce at least some of the physical characteristics
of the original object or process which is replicated.

(b) formal models—both components of the system modeled and inter-
connections among these components are represented by symbols which can
be manipulated using known mathematical techniques. Behavior of variables
within the model can be studied under a variety of conditions by strictly
formal manipulations.

(c) simulation models—symbolic models which cannot be solved
analytically in general terms, but behavior can be studied for certain
specified values of variables. Simulation places only weak restrictions
on the complexity of systems that can be studied.

Both (b) and (c) can be implemented on computers and thus are appro-
priate for AI applications. A natural language description also qualifies
as a symbolic model but simulations and formal models are often more pre-
cise. Natural language is flexible and easy to communicate but it is
difficult to rigorously determine predictions from models expressed in
this medium. Formal models can be analyzed in rigorous fashion but
constrain the model builder to consider only that class of models for
which he knows solutions are available. Simulation, especially with the
aid of computers, can be used for a wide range of models without serious
constraints on size and complexity. There is also the ability to make
rigorous determinations of the implications of the model, for the com-
puter can execute the program and determine behavior in particular
situations.

Models can be subdivided within given categories according to their
nature on certain dimensions. Models are dynamic or static depending
upon whether features or symbols change with time. Models are stochastic
or deterministic depending upon whether the model contains intrinsic
probabilistic or random elements which affect the outcome or response of the model.

While the above definitions address characteristics of model structure, one must also be concerned with content. The aim of a model is not to reproduce reality in all its complexity. The model builder selects for inclusion in the model those features of reality that he considers to be essential to his purpose. On the other hand, since a model is a different object from what it models, it has properties not shared by its counterpart. These questions of content illustrate the importance of recognizing that one cannot give an absolute definition of the relation between an object or process and a model of it, because the adequacy of the model depends on the questions one is going to ask it.

The model relation is inherently ternary. Any attempt to suppress the role of the intentions of the investigator B leads to circular definitions or to ambiguities about "essential features" and the like. . . . If A is the world, questions for A are experiments. A* is a good model of A, in B's view, to the extent that A*'s answers agree with those of A's, on the whole, with respect to the questions important to B. (Minsky, 1968b, p. 428)

The model builder thus assesses the sufficiency of his model by comparing the behavior of the model with the behavior of the object or process in the same situation. Model builders investigating applications of AI in IR, for example, hope to identify portions of the information retrieval tasks done by humans which can be fully delegated to the machine.

2. Testing

When applying AI techniques in IR systems, it is useful to identify two levels of evaluation which correspond to the performance and human acceptance testing in the engineering design process: (1) studying characteristics of system behavior, especially particular system components;
and (2) evaluating the utility of the system to system users. Robertson (in press) explains the reason for this approach:

There are many properties that a system might have which might well not be relevant to its utility to the user, but which might, in moving from one situation to another, affect those characteristics which are important.

As an example of the distinction between the two levels, consider the study of similarity measures used in the comparison of document surrogates and query formulations. There is a need to investigate the behavior of various measures initially, comparing their performance as discriminators of relevant documents over several document collections and queries in order to:

(1) determine if they differ significantly in their behavior. If not, then only one similarity measure need be implemented in a particular system.

(2) use knowledge of any identified differences in behavior to improve system performance. Unless one understands the behavior of different measures, one cannot systematically make changes in the similarity measure used by a system to improve the system's utility to the ultimate user. Microevaluations of a particular system component thus can suggest elements to change if macroevaluations of whole systems indicate that performance is not adequate. Microevaluations can include two types of performance tests: (1) internal comparison of one automatic technique to another to better understand their behavior; and (2) external comparison of automatic and manual techniques.

With respect to evaluation of human acceptance of a system, generally one looks at effectiveness and efficiency. Effectiveness considers how well system objectives are met and efficiency considers how system resources are used. In the context of IR these two aspects may be
further described as follows:

**Effectiveness.** While any number of factors can be identified as contributing to system effectiveness (e.g., exhaustiveness, accuracy and appropriateness of the information items in memory; convenience of user access; ease and brevity of user transactions required to obtain a system response), "retrieval system effectiveness has become almost synonymous with retrieval system 'accuracy'" (Katter, 1969). Two measures are associated with accuracy: (1) precision = the percent of items retrieved that are relevant; (2) recall = the percent of relevant items in memory that are retrieved. Effectiveness is thus judged in terms of the percent of relevant material retrieved and the accompanying amount of irrelevant material. Attention only to the recall-precision paradigm for measuring effectiveness may be misleading. There is a need to focus on actual modes of use of information retrieval systems, particularly online systems. The naive concept of the user as always wanting all those and only those documents "relevant" to his query may lead evaluators to evaluate "their way towards better retrieval systems for one-track minds, in consequence putting aside the possibility that retrieval systems to satisfy a multiplicity of needs (in one fell swoop) might be conceivable" (Doyle, 1975, p. 355). An understanding of how systems are used is a prerequisite to assessing their effectiveness, especially when users may generate ideas for uses in addition to those initially intended by the designer.

**Efficiency.** In evaluating efficiency it is important to define system boundaries which include all the system elements which should be considered in the evaluation. One can characterize information retrieval activities in terms of the table given on the following page, where boxes under user and computer would contain the operations performed at each
stage. The table serves as a reminder that system design determines a
division of labor between user and computer, and into three time periods:
pressearch, search, postsearch (Cooper, 1972).

<table>
<thead>
<tr>
<th></th>
<th>User</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presearch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postsearch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table is a helpful device to use as an aid in evaluating system
efficiency—the time spent by the user and computer in each stage and
the cost to each.

There are a number of sources of cost at each stage. Not only does
the user pay for searching the system at a specified price, but also he
places a value on the time and effort spent using the system in addition
to the amount paid to use it. Computer costs depend in part on computer
time spent in each stage. For particular representations one must con-
sider creation costs (e.g. indexing, abstracting, classification, conver-
sion to machine-readable form), processing costs (e.g. file creation and
maintenance, retrieval from the file), and storage costs.

Cost effectiveness is the relationship between level of performance
(effectiveness) and cost of achieving this level. A system is more
efficient if it attains the same effectiveness at lower cost. In retrieval
systems choice of how effort should be divided between man and machine
depends on cost and feasibility. Any cost effective design must take
account of what man and machine each does best. The human can propose
alternatives, establish constraints, and place a utility value on con-
sequences of the search. The computer can handle large amounts of com-
plex data, compare data at high speeds, sort, and maintain large files.
Cooperation is useful in tasks which are either too difficult at present
for either man or machine to do alone or are in some way intrinsically suitable for interactive computing. The issue is one of substitutability of resources—at present it is not clear to what degree human resources can be replaced by machine techniques in the performance of information retrieval. Furthermore, it is not known what level of substitution will prove cost effective.

D. Problem statement

The paradigm consists of a vocabulary, mathematical structures, preferred methodologies, and examples which apply these components to carefully chosen problems. Kuhn indicates that, once the paradigm has been adopted by workers in a field, it provides a channel for experimentalists leading them to tackle tractable problems (T. Martin, 1973, p. 206).

A paradigm in the sense of Kuhn (1970) is thus a general framework in which workers pose and solve problems, since the initial formulation is sufficiently open-ended to leave all sorts of problems for practitioners to solve. A paradigm is a source for meaningful research problems and also provides an organizing framework which permits new research results to make a contribution to some total picture. While ad hoc techniques in information retrieval may work quite well for specific cases, lack of a framework means that it is hard to apply knowledge gained in solving one problem in system design to the solution of other problems.

AI in IR can be viewed as a paradigm. Enumerating the necessary components of a paradigm as given in the above passage from Martin, one has:

(1) a vocabulary—the ability to develop new concepts depends to some extent on the available vocabulary. In IR the vocabulary available to describe systems comes largely from work with manual systems and to a lesser extent with batch systems. Looking at online systems, one is prompted to ask new questions when the conceptual schema developed in AI is applied to IR—ideas from pattern recognition, representation,
problem solving, and learning.

(2) mathematical structures—while AI does not really have areas of mathematics which are unique to it, there are a number of theoretical subjects which have been identified as particularly relevant to AI studies: mathematical logic, computational linguistics, theory of computation, information structures, control theory, statistical classification theory, graph theory, theory of heuristic search (Nilsson, 1971, p. viii).

(3) preferred methodologies—the general method follows the engineering design process described in the previous section, with its stages of problem identification, information gathering, idea generation, model building, and testing.

(4) examples which apply these components to carefully chosen problems—this is the subject of the remainder of this study.

In order to determine what AI can actually contribute to the design and development of IR systems, research must be pursued at two levels:

(1) analysis of possible areas of application of AI in IR; and

(2) completion of investigative studies addressing research problems identified in the analysis phase. Figure 3 on the following page clarifies the relation between these two levels and identifies the chapters where they are discussed.

Analysis. Possible areas of application of AI in IR follow directly from consideration of AI concepts. Areas of application can be briefly characterized as follows:

Pattern Recognition

1. Automatic indexing as a feature selection problem

Examine AI feature selection criteria as possible criteria for selection of index terms in IR.

2. Pattern classification
Figure 3. Artificial Intelligence in Information Retrieval: Analysis and Studies
Consider alternative ways of formulating IR as a pattern classification task (e.g., prototype matching vs. clustering).

3. Measures of similarity

Examine alternative measures of similarity used in AI as a basis for retrieval rules matching query formulations to document surrogates in IR.

Representation

4. Internal representations

Consider the use of AI representations for representing information items in the memory of an IR system.

5. External representations

Consider alternative forms of display at the user-computer interface as aids in developing query formulations.

Problem Solving

6. Question answering as a theorem proving problem

Identify questions which can be answered when posed as theorem proving problems.

7. Heuristics in IR

Examine what form heuristics may take to aid searching in IR.

8. Ill-structured problems

Consider how to design the user-computer interface to facilitate conversion of initially ill-structured queries to a well-structured form which the system can process.

Learning

9. Short term learning

Examine which AI learning techniques may utilize feedback to modify system response to individual queries.

10. Long term learning

Examine which AI learning techniques can be used for representation modification and/or extension to change system response to new queries.

Detailed analysis of each application area given above includes survey of:

(1) IR approaches--since some techniques used in the past may be viewed as AI in IR, IR work relevant to the application area is
(2) AI approaches—AI techniques relevant to the application area are described.

(3) Problems for research—specific research problems in need of investigation are enumerated, identifying questions to be answered in order to determine the applicability of AI techniques to this area in IR. Studies. The analysis phase yields an enumeration of research problems which should become the subject of investigative studies. In order to suggest what form these studies may take, two sample studies, shown at the lowest (i.e. most specific) level in Figure 3, have been completed. Figure 4 below can be used to put the two studies in perspective.

![Diagram of IR System]

**Figure 4.** Schematic of an IR system

Stage I includes the activities of formulating a query in such a way that it can be processed by the computer subsystem. Stage II involves the activities of query processing by the computer subsystem, once the query is in a form which it can handle. The study "Query Formulation as Problem Reduction" looks at stage I: How is the inquirer's initially ill-structured problem converted to a form which the system can process? The study "Processing Documents as Queries" looks at stage II: What happens when different internal representations are used as a basis for retrieval when searching for documents like one already known to the inquirer? The studies are self-contained, but at the same time can be thought of as small pieces fitting within a larger framework. Each study falls
primarily within one category at the next level as shown in Figure 3.

The figure is simplified, however. The following brief descriptions
of the studies include mention of multiple relations omitted from the
diagram.

1. Documents as Queries

The situation is one in which the task is to find documents like one
already known to the inquirer. It is the study of a particular system
component, ways of operationalizing the system relevance judgment.
This research problem relates to several areas:

Representation—internal representations.
The focus of this study is the relationship between the internal
representation of document surrogates in the computer and the set
of items retrieved in response to a query document. The repre-
sentation of the document surrogate in the data base will determine
what the basis of comparison between query document and document
surrogates in the data base can be, e.g., they share common index
terms, they are written by the same author, etc. Existing document
representations have redundancy in the sense that there are different
approaches to the task of locating documents relevant to a given
query.

Pattern recognition—pattern classification and measures of similarity.
The pattern classification task under investigation can be viewed
as prototype matching, where the document submitted as a query is
the prototype. Use of a particular similarity measure will deter-
mine which items in memory are retrieved when they are sufficiently
similar to the query document.

Problem solving—ill-structured problems.
The query to find other documents "like" one already known to the
inquirer is ill-structured because the basis of judging likeness
has not been specified. Elements available in document representa-
tions determine what elements can be compared for similarity.

Learning—short term learning.
Results of this study have implications for ways to implement rele-
vance feedback, where a retrieved document identified by the inquirer
as relevant can be used in the search for similar ones in memory.

2. Query Formulation as Problem Reduction

The task is to develop a model of the problem solving process used
by an intermediary in converting the user interest statement into an
SDI profile. The model is used as a basis for determining what parts
of this process could be accomplished directly by the computer or
through user-computer interaction, as well as for identifying inter-
mediary expertise which cannot be duplicated by machine. This research
problem relates to two areas:
Problem solving—ill-structured problems.

In most SDI systems the machine cannot directly process the data given in the user interest statement. The intermediary’s task of query formulation (SDI profile development) beginning with the user interest statement can be thought of as a problem which can be divided up into identifiable subproblems (e.g., selection of databases to search, selection of terms under which to search).

Representation—external representations.

Results of this study have implications for design of displays at the user-computer interface as aids in developing query formulations.

The research described above under the categories of analysis and studies represents definition and initial exploration of possible AI applications in IR. The intent of the analysis phase has been to identify problems for research. Chapters III–VI consider each concept area separately.

Chapter VII functions like a "topographical map", providing an overview of the research problems identified in Chapters III–VI as well as an assessment of those areas of investigation likely to be most fruitful given the state of the art in other AI applications. In addition, there has been an effort to provide examples formulated and conducted within this conceptual framework by completing the investigative studies reported in Chapters VIII and IX.
CHAPTER III

PATTERN RECOGNITION

In Chapter II (p. 24) a diagram of a pattern recognition device is used to illustrate the three stages in pattern recognition by machine: (1) representation of objects in machine-readable form; (2) feature selection; (3) classification. In information retrieval development of document surrogates and query formulations may be viewed as a feature selection problem; determination of which document surrogates to retrieve in response to the query formulation may be viewed as a pattern classification problem. Just as the classification decision in other pattern recognition tasks depends on some measure of similarity, an IR system bases retrieval decisions on the calculation of some measure of similarity between each document surrogate and the query formulation. While the various components of a pattern recognition system are interrelated, one can distinguish three possible areas of application of AI in IR:

1. Automatic indexing as a feature selection problem

Examine AI feature selection criteria as possible criteria for selection of index terms in IR.

2. Pattern classification

Consider alternative ways of formulating IR as a pattern classification task.

3. Measures of similarity

Examine alternative measures of similarity used in AI as a basis for retrieval rules matching query formulations to document surrogates in IR.

Before considering each of these areas separately, it is helpful to identify two alternative feature selection-classification models. The
first model (forming the basis for statistical or geometric approaches to
pattern recognition) is one in which the features are treated as components
of a vector. Each pattern is considered to be a point in the resulting
n-dimensional feature space. Classification then becomes a problem of
dividing n-dimensional space into regions, each representing a particular
class. While widely used, this model has been criticized for focusing
primarily on statistical relationships among features rather than on the
structural properties that characterize certain types of patterns. The
second model (forming the basis for syntactic or structural approaches to
pattern recognition) defines a class of patterns as satisfying a certain
set of relations among suitably defined features. Relations might be boolean
expressions in a relatively simple case. The applicability of both these
models to IR is explored throughout the remainder of this chapter.
A. Automatic indexing as a feature selection problem

A feature is a property, attribute, mark, quality, trait, or character-
istic. In character recognition, for example, features might be the
occurrence of straight and curved line segments. In pattern classification
each feature plays a dual role, for on the one hand it must be a common tie
binding members of a class together and on the other it must be a distinc-
tive mark separating one class from another. Choice of features to
characterize objects is completely dependent upon the types of objects and
the classes into which they are to be categorized (Gose, 1959). An attempt
is made to identify features which are characteristic of classes or of
differences between classes. If a feature can be found which has one range
of values for members of one particular class and a nonoverlapping range of
values for members of all other classes, then that feature will be sufficient
for determining whether or not an unknown sample is a member of a particular
class. Similarly necessary features can determine non-membership but not membership in a particular class. Most single features used in pattern recognition are neither necessary nor sufficient, so decisions must be based on combinations of features.

Because few features are either necessary or sufficient, one must suggest other characteristics which may serve as a basis for selection. Feature selection may be viewed in two contexts:

(1) with respect to the dimensionality reduction task alone—to reduce the number of features to a manageable number before classification is attempted;

(2) with respect to the pattern recognition system as a whole—to determine which features are most important to the classification task.

The first sort of feature selection uses only information contained in the machine-readable representations of objects, e.g., statistical parameters calculated from this data. The second sort actually uses knowledge of class membership as a basis for selection of important features. Somehow one must develop feature selection algorithms to determine if there are features which can aid in (Brick, 1969):

(1) reducing complexity (dimensionality) of class description;

(2) clustering class members (intraclass features);

(3) separating different classes (interclass discriminators).

Likewise, one must identify extraneous, redundant, or useless features.

The quality of classification decisions one makes depends on the features one chooses to select. In problems where it is not evident how to formalize human feature selection methods, experimentation to determine suitable bases for selection of features is the alternative route.

1. Information retrieval approaches

The problem in information retrieval is: given a data base, to
determine the set of features which, when used as a basis for document surro-
gates and query formulations, leads to efficient retrieval of the desired
subset of items in the data base in response to each query. If one con-
siders first the situation in which the feature selection process is accom-
plished by humans, two parts become evident. Descriptive features (author,
title, date, citations, etc.) may be transferred verbatim from the document
to the surrogate and specific elements will be included or excluded depen-
ding on system policy. Subject characterization, i.e. indexing and/or
abstracting, is more complex and is not necessarily limited to the selection
of words or sentences from document text. Subject characterization by
humans may involve three operations:

(1) text of document is scanned to determine the information content;

(2) a decision is made as to which parts of the document content are
to be recorded (may depend on user interests and queries likely to be made);

(3) document content selected is expressed in the language of the
system. Where no control over index terms exists, terms may be selected
from text. Otherwise, content considered important must be translated into
controlled vocabulary terms.

With document texts available in machine-readable form, a number of
methods for mechanizing the feature selection process have been imple-
mented. The task is one of devising procedures by which a machine can
index, since the selection of other types of features (e.g. author) by
machine is straightforward.

There are at least two different kinds of relationships that can
exist between terms, i.e. semantic and syntactic. Semantic closeness
includes such factors as synonymity among index terms and generic rela-
tionships. Such relationships between terms are based strictly on the
meaning of the terms in question. Syntactic closeness is based on term co-occurrences in documents. One thus finds two approaches to automatic indexing; both syntax and semantics have been used as the basis of criteria for index term selection. Each criterion reflects certain assumptions as to how occurrences of words relate to the content of documents in which they occur. It should be noted that any criterion for suitable index terms based only on occurrence of words in text does not permit direct inference as to retrieval effectiveness. Just as is the case in pattern recognition in AI, the feature selection process does not exist independent of a pragmatic context. In IR access of inquirers to relevant documents may be hindered if feature selection is done without regard to the type of queries likely to be made. Justification for any automatic indexing technique must ultimately be based on successful retrieval.

While discussion of automatic indexing generally emphasizes feature selection from documents, it may be possible to treat queries in the same manner by identifying features to be included in the query formulation used in search of the file (Swanson, 1962). In manual systems this relationship is evident when one views the task of indexing as some mapping of statements of document content into controlled vocabulary terms and similarly the task of searching as one of mapping the query into the controlled vocabulary. When the operations are mechanized, they may be less constrained and alternative criteria for feature selection can be investigated. The discussion below follows the analytic framework proposed by Sparck Jones (1974). A detailed review of specific automatic indexing techniques investigated in the past can be found in Stevens (1970).

Syntax.

In studying syntactic approaches to automatic indexing, the question
of interest is whether syntactic information should be represented in
document surrogates and whether automatic procedures can be used to obtain
it. Implicit syntax involves the identification of word strings (e.g.
phrases) to function as units, whereas explicit syntax spells out the rela-
tional structure of text. Parsing procedures can satisfy the demands of
either implicit or explicit syntax: partial parsing identifies phrases of
a sentence without specifying syntactic relations between them, while
full parsing gives an overall grammatical characterization of a sentence.

Presently more use is made of syntactic criteria than syntactic
information; feature selection procedures make use of syntax to identify
semantically important features for indexing rather than to preserve syn-
tactic information for its own sake. Limited syntactic analysis to iden-
tify noun phrases yields terms which are candidates for selection. This
process is used, for example, in the LEADER system (Hillman & Kasarda,
1969). In evaluating the usefulness of syntactic criteria, one must
determine whether selected features are more useful in retrieval than ones
selected by some other criteria, such as statistical. Syntactic-based
indexing has not been fully tested. Questions remain as to the degree
to which implicit syntax of compound terms is useful, as to whether ex-
plicit syntax has any value, and as to how easily either type of syntactic
information can be supplied automatically.

Semantics.

Selection of features not involving syntactic processing is either
dictionary-based or statistical. A negative dictionary contains words
which have such a high frequency in the data base (e.g. prepositions,
articles, conjunctions) that they fail to assist in the task of distin-
guishing one document from another and thus are omitted. When a negative
dictionary is used, feature selection gives a list of all nontrivial words in the original text. In the absence of other applicable techniques for semantic analysis one has the adoption of statistical approaches to document content characterization. They provide a posteriori rather than a priori interpretation of documents by selecting terms from text which may be particularly appropriate for retrieval purposes. Computers are of course well suited to such processing.

The purpose of statistical analysis of text is to identify content-bearing words, but statistical word criteria are really applicable only to reasonably long input texts where sufficient information about word frequency is available. The general idea that word frequency within texts is broadly correlated with semantic importance was first publicized by Luhn (1958). The justification of frequency of occurrence as a measure of word significance is based on the fact that a writer normally repeats certain words as he advances his arguments and as he elaborates on an aspect of his subject. The possibility of higher weighting for words in the title, first paragraph, and summary was also suggested.

The fundamental thesis of this approach is that statistics on kind, frequency, location, order, etc. of selected words are adequate to make reasonably good predictions about the subject matter of documents containing those words. This is a very simplified approach and one which ignores the ways in which certain types of words combine to convey information. It is interesting that Maron (1951) cast this procedure in a pattern recognition model early in the development of automatic indexing:

Consider the following way of talking about clue words and prediction methods for automatic indexing. The basic objects in this universe of inquiry are the class of documents under consideration. These objects (documents) have properties 'features', viz., the clue words that they contain. The properties are the observables
In our universe and we take measurements on them. Thus a measurement is a list of the kinds of properties that an object has. (More sophisticated measurements would provide information about the frequency, distribution, order, etc., of the properties of our objects.) The information supplied by the measurement when properly formulated constitutes the evidence. . . . Statistical data relating clue words and subject categories constitute hypotheses. (p. 406)

In selecting terms based on frequency counts, it is seldom adequate simply to count occurrences of terms in the raw text. Keyword normalization through word stem routines seeks to equate terms with variable endings, where the endings can be ignored for the purposes of retrieval. There is a danger, however, of either overstemming or understemming; overstemming gives erroneous amalgamation of distinct words and understemming represents a failure to associate different word forms which share a common stem.

Simple frequencies may also be modified by various weighting schemes, where information about within document frequencies of terms, lengths of documents, and frequencies of terms in the collection as a whole may be used to generate weights by a variety of functions.

Weighting can also play an important role when one considers development of a retrieval system over time. Selection of a fixed vocabulary may be unsatisfactory for a changing collection. An alternative strategy is to control the indexing vocabulary by weighting. Instead of choosing certain terms at a particular time and then retaining them, terms can be given different (possibly equal to zero) weights at different times in the life of the collection.

The discussion of syntax and semantics highlights the fact that there are two distinct languages with which one is concerned: the natural language of input texts and the artificial language of document surrogates. Linguistically the latter may be more or less closely related to the former.
The objective in automatic indexing is to select those features of natural language text which make possible good retrieval performance.

This brief review of automatic indexing has identified feature selection criteria found useful in a particular domain—automatic indexing for information retrieval systems. The use of syntactic and/or semantic criteria can be summarized as follows:

1. Terms are chosen from document or query texts. Terms may be words, word stems, noun phrases, prepositional phrases, or other content units which exhibit certain specified properties.

2. Weights may be assigned to each term on the basis of its type, its frequency of occurrence, or its position in the document.

3. Relational indicators may be put between terms to express syntactic or logical relationships.

Criteria used for automatic indexing have been chosen primarily as a matter of convenience rather than being derived from any firm theoretical basis. Syntactic criteria have been little used and semantic criteria have in most cases been limited to certain frequency characteristics of features. A brief review of approaches to feature selection in pattern recognition is given below since it is possible that criteria shown to be effective in other AI applications will prove to be useful for feature selection in IR as well.

2. Artificial intelligence approaches

Just as techniques for automatic indexing can be categorized as semantic or syntactic, feature selection techniques in pattern recognition can be divided into two categories: mathematical and structural. These reflect the two alternative feature selection-classification models described on p. 45, statistical vs. structural approaches to pattern recognition.
Mathematical techniques.

All mathematical techniques for feature selection can be classified into one of two major categories: (1) feature selection in measurement space and (2) feature selection in the transformed space (Kittler, 1975). The first category includes feature selection techniques to accomplish dimensionality reduction by reducing substantially the number of required features, selecting a subset of the original set. This is achieved by eliminating features which contain irrelevant and redundant information. The second category of techniques utilizes all information in the pattern vector to yield a feature vector of lower dimensionality. Examples of each approach are given below. Detailed discussion of the mathematics may be found in Kittler (1975).

In feature selection in measurement space, the task is to select the best subset Y of n features where \( Y = \{ y_i \mid i = 1, 2, \ldots, n \} \) from the set \( X = \{ x_j \mid j = 1, 2, \ldots, N \} \) of N measurements representing the pattern. The "best subset" is that set of n features which minimizes classification error, or some other criterion related to classification error, with respect to any other combination of n features taken from X. Therefore to determine the best subset, one must check all possible combinations of n features from N measurements. Two approaches have been used:

1. Methods based on minimization of entropy. Prior to looking at any of the features of an object, the uncertainty that exists as to the category to which the object in question belongs is the entropy, as defined in information theory. Given that a particular feature \( y_i \) does occur in the object, one can calculate how much the initial uncertainty will change. Given two features \( y_1 \) and \( y_2 \), \( y_1 \) is a better feature if its occurrence in an object removes a greater amount of the initial uncertainty than would
the occurrence of $y_2$. Selection of the $n$ best features maximizes the information available about probable class membership.

(2) Euclidean distance measures. Use of Euclidean distance originates from the intuitive argument that the greater the distance between elements of different classes, the better the class separability. A typical example of such a criterion is the average distance between pattern vectors, which is calculated using all possible subsets of $n$ features to choose the subset of the $N$ original features which is to be used.

In feature selection in the transformed space, each pattern vector $\mathbf{x}$ is mapped into a lower dimensional vector $\mathbf{y}$ by some transformation. The following are examples of transformations which have been utilized:

(1) Weighting. Allow the weight of a particular feature to be inversely proportional to the variance of that feature in pattern space. This approach gives greater weight to features with occurrences concentrated in a few patterns.

(2) Factor analysis finds a lower dimensional representation that accounts for correlations among features. It generates new features which are linear combinations of old ones. The problem with factor analysis as a dimensionality reduction technique is that it is overly concerned with faithful representation of the data. But for classification, one is interested in discrimination, not representation (Duda & Hart, 1973). Roughly speaking, the most interesting features are the ones for which the difference in class means is large relative to the standard deviation.

(3) Discriminant analysis involves reducing dimensionality from $N$ dimensions to one dimension by projecting $N$ dimensional data onto a line. The objective is to make the difference between class means large relative to some measure of standard deviation for each class. This procedure can
be generalized to multiple discriminant analysis for more than two classes. As a feature selection technique, in discriminant analysis one uses a step-wise selection procedure to eliminate features which either of themselves or in combination with other variables lack discriminating power.

In contrast to methods of feature selection in measurement space where all subsets of features have to be considered, implementation of feature selection techniques in the transformed space only requires determination of the optimal transformation and is thus less computationally demanding. Choice of the appropriate mathematical feature selection technique for a particular problem must be a compromise between computational complexity and the reliability of the method. Computational complexity includes both the computer time and memory space required (Chen, 1975). For real time recognition problems in particular, one needs computationally efficient feature selection criteria. An exhaustive search for the optimum feature subset may not be practical. The significance of mathematical techniques for feature selection is that through the vehicle of mathematics, results of research in one area of pattern recognition can be used in other areas, which is not necessarily the case for nonmathematical techniques.

**Structural techniques.**

The second category of feature selection techniques tries to overcome the limitations of mathematical techniques which simply produce lists of features. Because of its fixed size, the list of features is limited (for any given set of features) in the detail of distinctions it can make. What is required is a list of features together with some statement of relations among them. This introduces the distinction between monadic and structural levels of representation (Michie, 1974): at the monadic level, input is decomposed into features which form a description without regard to struc-
tural interrelations among the features themselves, while the structural level exploits organization and interrelation of primitive features.

Grammatical analysis of a sentence is a structural description in which parts of speech are the components and the order in which they are related is a structural relation. This also suggests that hierarchies of structural description are possible—in natural language one may look at phonemes, syllables, words, phrases, clauses, and finally sentences. The description level chosen reflects those aspects of the class of patterns in question important to the task at hand. Just as syntactic criteria in automatic indexing can be used to supplement semantic criteria, structural techniques in pattern recognition are now beginning to be explored as a source of richer descriptions than that possible with the isolated features selected by mathematical techniques.

3. Problems for research

The reviews of automatic indexing using semantic and syntactic criteria and of feature selection in AI based on mathematical and structural techniques suggest a number of as yet unexplored areas.

Comparative studies.

While automatic indexing has been the subject of considerable experimentation since data bases first became available in machine-readable form, there is a dearth of comparative studies. From experiments performed to date, something is known about how particular techniques perform on particular data bases, but relatively little is known about how to choose the best technique for a particular application on a particular data base. Although Maron and Kuhns (1960) assert that one can index by machine with increasing accuracy if one is willing to collect enough statistical data relating words and classes and if one is prepared to consider more of the
relationships that exist between individual words, word combinations, word types, etc. and classes, there is a question as to when such a process is no longer cost effective in terms of the improvement in retrieval performance to be expected. More research is needed systematically comparing alternative techniques for automatic indexing using such evaluation factors as cost of computation and retrieval performance (ability to distinguish relevant from nonrelevant documents in response to a query). One can propose three categories of comparative studies: automatic vs. manual; automatic vs. automatic; and monadic vs. structural.

Automatic vs. manual. The real question in automatic indexing is not whether the same terms are selected by human indexers and the machine system for a collection of documents, but rather how effective the index terms selected in either case are for retrieval purposes. The speed and storage capacity of computers allow methods of indexing which would be utterly impractical for humans. At the same time automatic techniques must surmount difficulties not faced by humans (Gar-Hillel, 1964b):

1. identification of what is to be regarded as instances of the same word;

2. synonym identification;

3. automatic procedures yield a set of terms taken exclusively from the document itself.

With the availability of online systems it may prove desirable to establish a compromise between fully automatic and fully manual feature selection by making processes computer-aided. In indexing, for example, the terms used to index references cited by a new article entering the system could be retrieved as candidate index terms for that article (Gray & Harley, 1971). Human indexers would then make the final decision on terms to be included.
Studies are needed to determine the cost effective combinations of manual and automatic techniques for both initial document indexing and later query formulation. A test comparing MEDLARS (manual indexing using a controlled vocabulary) and SMART (automatically generated document surrogates) led Salton (1973) to suggest that the intellectual input conventionally provided by expert indexers could be replaced by automatic methods. Whether Salton's conclusions hold true for processing of large data bases remains to be tested.

Automatic vs. automatic. As discussed in the section on testing (p. 34), an important type of study is internal comparison—comparing performance of different automatic techniques to better understand their behavior. An example of such an internal comparison is a study done by Carroll (1975) comparing several statistically based keyword selection criteria. Before choices can be made among techniques suggested either by previous automatic indexing studies in information retrieval or feature selection studies in pattern recognition, there is a need to determine if such techniques differ significantly in the indexing terms selected and how they differ.

Nomadic vs. structural. Just as syntactic criteria have had limited use compared with semantic (statistical) criteria for automatic indexing, structural techniques are only beginning to be explored as an alternative to mathematical techniques for feature selection in AI. There must be an attempt to relate the feature selection technique chosen to the classification task to be performed. In AI structural techniques have been used in scene analysis, for example. In IR isolated index terms may be a sufficient basis for retrieving document surrogates, but success in performing different tasks such as question answering, fact or passage retrieval may
depend on the availability of other types of information. Since feature selection determines to a large extent what elements are available for the internal representation in a system, the possibility of retaining structural (syntactic) as well as statistical (semantic) information should be explored.

Heuristics for feature selection.

In addition to mathematical and syntactic techniques for feature selection, one must also recognize the possibility of incorporating heuristic techniques in a system (Tou & Gonzalez, 1974). These rely on human intuition and experience to select features. Thus, unlike mathematical and to a lesser extent structural techniques, heuristic techniques are problem dependent. Application of mathematical feature selection techniques to automatic indexing may not be successful without the use of heuristics to aid in selection of terms. A study by Borbash (1970) is an illustration of this. He compared retrieval performance resulting from index terms selected automatically and by human analysts. In the machine method a training set of documents previously identified as relevant or nonrelevant was used as a basis for selection of the most "informative" index terms. An information theoretic measure was used to rank terms according to the number of bits of information which each index term individually provided about the category of documents in the training set. Terms selected by human analysts led to better performance, however, because analysts supplemented the information in training sets with information not made available to the machine system, such as data on frequency of terms in the collection. Human analysts avoided using index terms which had a high frequency of occurrence in the collection as a whole even though they were excellent discriminators over the training set. Mathematical techniques for feature selection may thus have to be augmented with heuristics
used by human indexers in term selection. The possible types and applications of such heuristics, whether used alone or in combination with mathematical and structural techniques, require further study. The use of heuristics in selecting terms during query formulation should also be considered.

**Feature selection in short term learning.**

A review of the mathematical feature selection techniques described above suggests that for the most part they are based on knowledge of class memberships of sample objects; in the case of information retrieval, this means that the system begins with examples of relevant and not relevant documents. While such feature selection techniques are therefore inapplicable for indexing at input, there is still the possibility that they can be employed in feedback routines in an online retrieval environment. With an increase in available computer storage space, it may be possible to eliminate feature selection from document text at input (except for processing using a negative dictionary and word stem routines, perhaps). The problem of deciding which are the "important" features of a document has no simple solution where there is a time interval between initial feature selection and actual search of the collection of document surrogates. Term significance is not inherent—it is a relative measure which depends upon system users, their particular interests and the kinds of queries they are likely to have. Subject indexing thus involves not so much indexing of intrinsic subject matter as indexing of documents in relation to the types of queries which are anticipated. According to the probabilistic interpretation of indexing, if an index term is selected for a document, it represents a prediction about how those inquirers who will be satisfied with that document will probably ask for it. The system is created in
anticipated of needs not fully known, but the measure of adequacy of a
system is the ability to satisfy the inquirers' needs as they arise.

Since indexing necessarily involves some selection of terms from a
larger set, performance of this operation at input means that the infor-
mation discarded could be that which is most vital to subsequent identifi-
cation of documents relevant to a particular query. If documents relevant
to a query do not resemble one another in initial descriptions, retrieval
performance (in terms of recall and precision) will not be good unless there
is some mechanism to enable the system to learn to discriminate between
relevant and nonrelevant items based on the inquirer's identification of
relevant documents among those retrieved. Storing full text requires no
initial analysis. But given such a data base, one is in a position to
determine the set of features which leads to retrieval of the desired items
in response to each query. When an inquirer submits a query and retrieves
a sample set of documents, identifications of each as relevant or nonrelevant
can possibly be used by mathematical feature selection techniques to select
those features which characterize relevant documents best for this parti-
cular query. Such an approach is feasible, of course, only if computationally
efficient feature selection techniques can be found to operate in an online
environment based on small training sets of documents judged relevant and
nonrelevant by the inquirer. Experimentation is needed to determine if
mathematical feature selection techniques satisfy this requirement.

3. Pattern classification

In a pattern recognition system, the output of the feature selection
stage forms the input to the classification stage. Classifications can be
categorized according to certain general properties, as determined by
relations among features, objects, and classes (Sparck Jones, 1970).
Relations between features and classes determine whether a classification is monothetic or polythetic. If a class is monothetic, all its members possess the same common feature(s), which is not true of polythetic classes. Relations between objects and classes determine whether one can have classes which are exclusive or overlapping. If objects are assigned to one class only then one has exclusive classes, while the assignment of objects to more than one class leads to overlapping classes. Relations between classes determine whether a classification is ordered or unordered. In an ordered classification, the classes are systematically related to one another, while they are not so in an unordered classification. When considering how best to handle a pattern recognition problem, one must not lose sight of the variety of types of classifications possible.

Computers play a central role in modern classification for a number of reasons (Sokal, 1974):

(1) Computers can find solutions to problems that are otherwise analytically intractable.

(2) Computers can handle classification of many objects characterized by many features.

(3) Development of algorithms for classification by computer forces one to characterize the process in detail, thus making it more objective. The influence of the computer on classification in IR as well as in AI is illustrated in the discussion which follows.

1. Information retrieval approaches

Before enumerating the alternative approaches to classification as found in AI applications, it is useful to identify the processes within an IR system which involve classification. There are three:

(1) Identification of the class of documents to be retrieved in
response to a query formulation. In most IR systems an exact match (e.g. document representation contains terms satisfying the boolean expression given in the query formulation) or a partial match (e.g. document representation contains n of N requested terms) is the basis for determining class membership. The retrieval set may be further divided into subclasses using a ranking algorithm or some other device to form groups of documents within the set retrieved.

(2) Grouping document surrogates in the data base. This may be done by manual assignment of a classification number (e.g. from the Dewey Decimal Classification) to each document or by automatic clustering of documents based on patterns of occurrence of terms in document representations (as described in Salton, 1975, pp. 333-357).

(3) Grouping of terms. This may be done by manual creation of categorized lists in thesauri or by automatic clustering of terms based on the document representations in which they occur (as described in Salton, 1975, pp. 364-375).

It is helpful to note the properties of some of these classifications. Automatically created classifications, such as term and document clustering, tend to be polythetic, overlapping, and unordered. This is in contrast to manual classification of documents using Dewey, for example, where the classification as it is usually applied in libraries is monothetic, exclusive, and ordered. Computer-based IR systems can accommodate more flexible classifications.

2. Artificial intelligence approaches

In AI applications one can identify three approaches to class definition:

(1) Sorting—unambiguous criteria for class belonging are externally given from the beginning so the actual process of placing objects into
classes according to features is mechanical.

(2) Prototype matching—the number of classes is given and each class is illustrated by a certain number of sample objects (prototypes). The task of the classifier is to extract features characterizing classes from prototypes and to place newcomers (objects with known features and unknown class belonging) into classes like the prototypes.

(3) Clustering—no information is available as to what classes are to be and what class samples are. Classes are created on the basis of mutual similarity and dissimilarity in a given set of objects. Various algorithms for generating clusters impose a structure on the objects to be clustered which one represents as vectors of features. The attempt at clustering implies a belief that the objects exhibit some structure and are not randomly or uniformly distributed throughout the space defined by the features. In clustering the computer groups together objects which share the largest number of features in common, or which have the highest mutual correlation. Assignment of higher orders of meaning is a human activity (Giuliano, 1967). The effectiveness of clustering is judged by whether the resulting groups are meaningful for the application at hand.

3. Problems for research

Given the approaches to class definition defined above, it is necessary to assess their applicability to IR. Sorting is a mechanical operation once the features to be matched have been identified. IR systems in which query formulations are made up of boolean expressions can be thought of as using sorting as a basis for classification, the determination of which document surrogates to retrieve in response to a query. Similarly, grouping documents in the data base using manually assigned
classification numbers is a sorting operation. Prototype matching and clustering have received less attention in IR, but the increasing availability of data bases in machine-readable form means that pattern classification no longer need be limited to a sorting operation. A number of questions require further study to determine the feasibility and utility of prototype matching and clustering in IR.

Prototype matching.

Most IR systems have rather limited options for query formulation. The inquirer must express his question as terms linked by boolean operators or sometimes simply as a list of terms. This is then compared with document surrogates in the file. An alternative approach would be to begin with a known relevant document. The task of the IR system would then be to find "similar" documents, an example of a prototype matching situation. In general, document surrogates are made up of a number of different types of features, any of which could be used as a basis for judging similarity—authors, free text words from title and abstract, assigned subject headings, citations. While definition of a prototype using free text terms or subject headings is most common, retrieval based on common sets of citations, called bibliographic coupling, has also been explored (Weinberg, 1974). Bibliographic coupling assumes that when two documents have common citations, they have subject matter in common. The strength of the subject relation is determined by the number of common citations.

While a number of alternative bases for judging similarity are possible, there have been few systematic comparisons characterizing the classes which are defined when alternative types of features are used. Recognizing the task of locating "similar" documents as one of prototype matching, it
is evident from AI that the task of the classifier is to extract features characterizing a class from the available prototype(s). Before choices can be made among alternative sets of features, there is a need to determine through comparative studies whether the use of alternative types such as subject headings and citations leads to overlapping sets of items retrieved.

Clustering.

- Three areas of application of clustering in IR can be identified:

(1) Document clustering to automatically classify documents in the database. Each cluster is represented by its centroid, a weighted set of terms derived from document vectors in the cluster. In processing queries in a clustered file, a query formulation is first compared with each cluster centroid and subsequently matched only to those document vectors contained in clusters with centroids of high similarity to the query formulation. This technique reflects the cluster hypothesis that documents relevant to one request are on the whole more like each other than they are like the nonrelevant documents. While index terms may be used as a basis for identifying sets of similar documents, clustering may also be done using citations. Cluster analysis using either subject terms or bibliographic coupling is based on the assumption that a classification scheme which is to be used for retrieval should depend upon the properties of the document collection itself rather than upon some a priori scheme (Price & Schiminovich, 1963).

(2) Term clustering for automatic thesaurus generation. Term clusters can be used to broaden a query automatically (Gammon, 1963). The query formulation need not contain only those terms mentioned explicitly by the inquirer, for related terms may be added from term groups given in an
automatically generated thesaurus. Terms known to be strongly related to query terms thus are used to augment the query formulation in an effort to retrieve additional relevant documents.

(3) Clustering of items retrieved in response to a particular query.
In this case clustering is used to identify useful subgroups in large retrieval sets (Precey, 1975). Existence of a number of different applications of clustering in IR suggests that one should investigate whether certain clustering algorithms yield better results for particular applications (i.e. document, term, or retrieval set clustering). The classification produced can vary in a number of ways: size of clusters, number of clusters, criteria for membership in clusters, overlap in cluster membership vs. nonoverlap (Bailey, 1974). The IR system designer has little guidance in the choice of algorithms, but some research in AI is beginning to provide a rational basis for comparing clustering methods. Dubes and Jain (1976), for example, compare eight clustering programs representing three clustering techniques, recording variables such as time required, memory used, number of clusters formed, and number of objects misclassified. More comparisons like this are needed if the IR system designer is to make the best use of automatic classification techniques. As Swanson (1973) observes:

Although it seems premature to try to predict the extent to which meaningful, operational applications of association techniques can be realized for information handling purposes, it seems that this can be a rich field for theoretical and pragmatic developments if researchers do not close off possibilities before they have been adequately investigated. (p. 73)

C. Measures of similarity

The various pattern classification techniques depend on having a way to determine similarity and dissimilarity among objects. Since success
in classification is likely to be a function of the similarity measure used, it is necessary to look at this component of pattern recognition systems in greater detail.

1. Information retrieval approaches

In the field of IR, similarity measures are used principally for comparing the following types of items: (1) pairs of document surrogates in automatic classification; (2) document surrogates and query formulations during retrieval; (3) pairs of features, such as index terms in automatic thesaurus generation. The range of similarity measures possible in cases 1 and 2 is circumscribed by the set of features obtained at the feature selection stage. Consider first the case where the features are a set of index terms. Each document surrogate may be viewed as an n-dimensional vector where components of the vector stand for n possible index terms. If the jth term is applied to a given document $D_i$, then the jth component of the corresponding vector would have 1 as its value; and conversely, a 0 in the kth component would mean that the kth index term is not assigned to $D_i$. The query formulation may likewise be interpreted as a vector. In two-valued matching one can require that all the terms present in the query formulation be present in the document surrogate for the latter to be retrieved. Given the vector interpretation, an alternative to two-valued matching is the determination of the degree of similarity between two vectors. A number of different measures of closeness (and distance) between vectors can be used; e.g. the cosine of the angle between two vectors in n-space. Weighting of index terms can be introduced by allowing terms to have values other than 1 or 0. If the inquirer does not consider each term in the query vector to be equally important, he may likewise differentially weight terms. Here, also, different measures of closeness
(and distance) between weighted vectors could be used in order to compute the degree of similarity between query formulation and document surrogate. Paice (1977, pp. 107-118) gives a number of examples of similarity measures for both weighted and unweighted vectors which have been used in IR applications.

In any of the above approaches, the probable relevance of a document to a query is determined by calculating the measure of similarity between document surrogates and query formulations. Document surrogates are retrieved if the similarity measure between the two is above some threshold. Calculation of similarity rather than a binary classification decision permits ranking of the output so that the inquirer obtains access first to what is presumed to be most relevant to his query.

2. Artificial intelligence approaches

In pattern classification it is necessary to determine intraset similarities (similarity measure) or interset discrimination (dissimilarity measure). The measure chosen depends on the available features. Any of the following measures is possible (Brick, 1969):

- Euclidean distance either weighted or unweighted
- Correlation functions or functions of correlation functions
- Normalized correlation
- Boolean 'AND' and Boolean correlation
- Frequency of co-occurrences
- Intuitive and observed similarities

The last category of "intuitive and observed similarities" suggests the subjectiveness which often characterizes human pattern recognition. Establishing likeness of two patterns depends both on the observer—be it man or machine—and on the context, because different observers do use
different sublanguages (sets of features) of patterns. The concept of pattern is essentially relative. What one is concerned with is the notion of pattern-for-an-agent, and the problem of pattern recognition is ill-defined until the class of agents has been specified (MacKay, 1969).

'Pattern recognition' implies a relationship between a pattern and an individual; and the individual, with his own peculiar experiences, should not be too lightly dismissed, as is sometimes the case. Recognition is not explicable in terms of properties of the patterns, or signs, alone. It is conceivable, for instance, that you and I would recognize the face of a third person, from quite different feature characteristics. This is conceivable, though perhaps unlikely. Even if we use the same set of features, we may weight them differently. (Cherry, 1956, p. 292)

In determining similarity, one must select the set of features to be used as the basis for measurement. Classifications based on many features will be general; they are unlikely to be optimal for any single purpose, but might be useful for a great variety of purposes. By contrast, a classification based on a few features might be optimal with respect to these features, but would be unlikely to be of general use (Sokal, 1974).

3. Problems for research

Recognition that there are different possible measures of similarity does not help one to decide which is the best for a particular application. Retrieval performance using different measures must therefore be studied. Several categories of research questions can be identified, based on the above discussion of similarity measures in AI.

**Similarity based on index terms.**

Determination of similarity between document surrogates and query formulations using index terms can include four phases: (1) weight document surrogate terms according to their relevance to each document;
(2) weight query formulation terms according to their relevance to the query; (3) determine correlation between each document surrogate and the query formulation; (4) arrange document surrogates in order of decreasing correlation value. These four phases make up a ranking algorithm. Since a number of possible weighting schemes and correlation measures are possible, there is a need for studies systematically comparing alternatives to determine their behavior. Sager and Lockemann (1976) completed a study comparing the results of using several weighting functions for document terms, weighting functions for query terms, and correlation measures. They found that certain ranking algorithms did differ substantially in the order in which retrieved documents were ranked. Observing that a single ranking algorithm may not satisfy every user or subject area equally well, they suggested that a generally applicable IR system should offer several ranking algorithms that differ substantially in their behavior, i.e. result in different orderings of the same set of document surrogates. Additional studies are needed to further investigate the behavior of different ranking algorithms, as well as the appropriateness of specific similarity measures for certain subjects and/or users.

**Similarity based on several types of features.**

Thus far the discussion of similarity measures in IR has emphasized comparison of index terms. Recognizing that several other types of features (e.g. authors, citations) make up document surrogates, it is reasonable to suggest an extension of the definition of similarity measures to include comparison based on these other types of features as well. Cleveland (1975) has illustrated how one might measure similarity between document surrogates using a Euclidean distance function based on author(s), journal, citations, and index terms. With the many types of features available in machine-readable document representations, one need no
longer be limited to comparisons based solely on index terms. The retrieval performance resulting from use of alternative types of features both singly and in combination as a basis for similarity measures deserves further study.

Alternatives to correlation and distance measures.

The measures described so far, whether based on index terms or other types of features, treat document surrogates and query formulations as feature lists or vectors. These measures ignore the possibility that individual terms may be supplemented by some syntactic or structural information. Work in AI dealing with structured descriptions of objects has already provided examples of what form similarity measures might take in this more complex domain. Winston (1977) observes that it is necessary to translate the abstract notion of difference into a concrete scheme.

For pairs of structured descriptions, he suggests that a simple way of doing this is to count up the number of differences. Should this be too coarse a measure, each type of difference can have an adjusted influence on the result through the use of weights reflecting relative importance. When comparing two directed graphs, for example, one might find differences in the number of nodes, the number of arcs, the direction of arcs, and the labeling of arcs and nodes. When syntactic information is retained as part of document representations, different types of similarity measures between document surrogates and query formulations will be required. This is likely to be needed in systems for question answering and fact retrieval from the full text of documents. The IR systems which retain information as to the sequence of words in text provide a simple example of how one could define a similarity measure using structural information. If one is interested in finding documents on "information retrieval", one could
say that the document surrogates most similar to the query are those in which the words "information" and "retrieval" are adjacent in the text of the document; those document surrogates where the two words appear in the same sentence but not adjacent are less similar; and so on. As IR and AI applications make more use of structured representations, there will be an increasing need to explore in detail possible similarity measures in this domain.

**Interactive pattern recognition.**

The last area of concern stems from the recognition that measures may be based on intuitive and observed similarities, as noted on p. 59. In IR systems it may therefore be necessary to involve the inquirer in an interactive dialogue with the system to identify similarities between queries and documents which are meaningful to him. There is increasing use of interactive pattern recognition systems in other AI applications (Chien, 1973). Experience with these systems should prove useful to IR system designers as they investigate ways to involve the human inquirer in the determination of the most informative features and the similarity measure appropriate for processing a particular query.
CHAPTER IV

REPRESENTATION

Just as a representation in AI is a formalism for knowledge possessed by a system, a document representation is a formalized statement of the nature of a document. A query formulation may be viewed similarly. Design of online IR systems must encompass not only internal representations (of documents, queries, relations between documents and between terms) within the computer subsystem, but also external representations—the displays of information at the user-computer interface. Both types of representation are possible areas of application of AI in IR:

1. Internal representations

   Consider the use of AI representations for representing information items in the memory of an IR system.

2. External representations

   Consider alternative forms of display at the user-computer interface as aids in developing query formulations.

A. Internal representations

   A representation may be thought of as a set of conventions about how to describe things. Many of the representations used in computer-based IR systems have simply been adopted from manual tools such as the card catalog without an analysis of their suitability for machine-readable files. Representations in AI, on the other hand, have in many cases been developed for knowledge-based systems utilizing computers. Following a summary of representations currently used in IR, a review of alternative representations used in AI applications is presented. While many of these
representations are apt to have only limited applicability to document retrieval systems, they are of considerable interest as one begins to consider the design of question answering and fact retrieval systems.

1. Information retrieval approaches

To build a representation one needs a language. Sharp (1975) has suggested that one can describe a language in terms of its vocabulary and structure. He contrasts natural language with artificial languages. The use of natural language vocabulary in IR systems means that there is no control over word form or synonyms; artificial languages have some degree of control over one or both. The structure of natural language is its syntax; the structure of artificial languages allows one to combine words from the vocabulary to represent more complex concepts. In addition there may be a thesaurus structure exhibiting relationships among terms and a classificatory structure exhibiting relationships among documents. Each of these is described in greater detail below.

Documents.

In spite of the name, document retrieval systems do not generally store the full text of documents, but rather document representations. The elements available for the representation are highly dependent on the output of the feature selection process. In its simplest form a document representation may be thought of as a property list made up of attributes with their associated values. One usually finds such attributes as author, title, source (e.g., journal), date, index terms, and classification number. The values which these attributes can have may be determined solely from the documents to be represented or may be subject to some external controls during the process of developing the representation. For example, author names may be listed as found in the document (John Doe) or may be reduced
to a standard form of entry using only first initial and last name (J. Doe). Similarly indexing terms may be words selected from the text of the document and thus relatively unrestricted in form or they may be limited to one or more terms selected from a predefined controlled list.

These document features or attributes may be categorized in various ways. Two useful categorizations are author-defined vs. indexer(system)-defined and content vs. context. Author-defined attributes are aspects of the representation determined by the author. In addition to author name(s), these can include title, journal where the article appears, and citations to other works. Indexer-defined features include classification number and assigned indexing terms. If an abstract is included as part of the representation, it may be author-defined or indexer-defined depending on its source.

Turning to content vs. context, content elements include title, abstract, classification codes, and indexing terms. Indexing terms may be selected from the document text, in which case they are referred to as free text or natural language terms. A degree of control can be effected using a series of normalization techniques: elimination of unwanted terms using a stop list, elimination of plural forms, creation of word stems by removal of prefixes and suffixes, and use of a synonym dictionary to replace terms equivalent for purposes of retrieval by a common identifier. If indexing terms are chosen from a controlled vocabulary, such as Library of Congress subject headings, rather than simply being selected from document text, greater standardization is possible. Indexing terms may be differentiated by assigning weights, associating with each term a numerical value reflecting its relative importance. Topics treated peripherally would be given a lower weight.
than the central topics of a document. Weights are sometimes assigned automatically, reflecting the frequency of occurrence of a particular term in a document.

To supplement content features, it is possible to include context features: those elements that describe various objective properties and relationships that hold for individual documents, authors, professional societies, journals, etc. (Maron & Shoffner, 1969). They are external attributes such as the journal in which a document is published, the other papers it cites, and sources of reviews about it. The context of an author may include such factors as education, place of employment, membership in professional societies, people co-authored with, etc.

While the feature selection process for context attributes is much simpler than for content attributes, their utility in retrieval is less well established. If there is a significant connection between context and content, then context information can be used to make inferences about content.

Retrieval using a document representation made up of attribute-value pairs usually requires a match on some attribute-value pair (e.g. retrieve document representations where author=John Doe) or some combination (e.g. retrieve document representations where author=John Doe and subject heading=linguistics). When retrieval is to be somewhat more complex, allowing matches of individual words in title and abstract, it has been found useful to retain some additional information in the representation regarding the relative position of words in text. This allows one to specify occurrence of two words in the same paragraph, the same sentence, or even adjacent in text as a basis for retrieval. The same principle has been found useful as a basis for searching the few IR systems which
currently retain the full text of documents, such as legal retrieval
systems.

Another type of document representation often derived from the text
of title and abstract is the document vector. The vector representation
may be binary (1 or 0 denoting occurrence or nonoccurrence of a particular
term in the document) or may include weights to indicate frequency of
occurrence. When the document representation takes this form, it is
possible to use similarity between document vectors and similarly repre-
sented query vectors as a basis for retrieval.

Document representations as found in retrieval systems serve pri-
marily a clue-providing function rather than an information-preserving
function. Portions of the document representation are used to gain access
to the full representation which in turn includes the bibliographic data
(author, title, source, date) needed to locate the full document in the
library. Even if an abstract summarizing the content of the document is
included as part of the representation, it is used in retrieval simply as
a basis for word matching. If one considers design of a document repre-
sentation solely to serve a clue-providing function, it is not clear that
our understanding of the most appropriate form for that representation has
advanced very far beyond the situation described by Bar-Hillel (1957):

An index set is a tool whereby a document is to be
cought whenever it is pertinent to a certain topic
and should be judged accordingly. The hook whereby
a fish is caught is not, in general, supposed to be
a miniature or condensed fish. . . . There is no
intrinsic reason why an index should be a minia-
ture, or condensed, document. That an index . . .
is often, but by no means always, a word or phrase
occurring in the document or in its title is prob-
ably only a transitory phenomenon, the result more
of convenience and conservatism than of any inherent
reason. (p. 106)
Queries.

The internal representation of a query must be in a form that the computer can use and manipulate. Thus query representations usually depend on the form which document representations have in a particular system. Where the document representations are property lists of attribute-value pairs, queries can be boolean combinations of attribute-value pairs. Document representations satisfying the boolean expression will be retrieved in response to a query. If queries simply include values without identification of the attributes with which they are to be associated, then the entire document representation must be searched for occurrence(s) of the stated value(s). For example, if the query is simply "linguistics", then documents having that as a title or as a subject heading would be retrieved. Where document representations take the form of document vectors, terms in the query are likewise used as the basis for developing a query vector to be compared with document vectors. If the document representation contains information regarding the relative position of words in text, then this can also be included in the query formulation, such as the query to find all document representations in which "information" is adjacent to "retrieval" in title or abstract.

It is evident from the above examples that the document representation—its attributes, their permissible values, and any structural information—should be developed with regard to the queries which inquirers are likely to bring to the system. The possible internal representations of a query are governed by the content and organization of the document representations. An IR system therefore restricts inquirers to asking the sorts of questions that can be answered using available document representations.
Relationships among terms.

Relationships among terms are most frequently shown by a thesaurus, which may embody two types of meaning for each term. One is its dictionary meaning, or words which can be treated as synonyms for the purposes of retrieval. These are indicated by SEE or USE references, e.g. "nonprint materials" USE "audiovisual aids". The other meaning could be labeled its "state of the world" meaning, i.e. how the term relates to other terms (Simmons, 1966). In a thesaurus these would be given by relational pointers labeled broader term (BT) or narrower term (NT) for hierarchical relations and related term (RT) for other useful relations. The thesaurus representation internally in the computer becomes a network in which nodes are terms and each arc is labeled with the name of a particular relation. One possible modification to this representation would give arcs labeled by strength as well as by type of association. Without such a well developed network, identification of the best terms for searching would be more difficult since "the network of relations of one descriptor to other descriptors thus provides a kind of definition by placing the descriptor into semantic space" (Unesco, 1971a).

The thesaurus is generally confined to exhibiting relationships among terms in a single data base. As online IR systems now provide access to a large number of different data bases with different controlled vocabularies, there is a need for some kind of "intermediate lexicon" to relate terms used in one controlled vocabulary to the corresponding terms in one or more other vocabularies (Gardin, 1969). While the semantic relationships displayed in a thesaurus or intermediate lexicon are generally identified by their human builders, it is also possible to explicitly store relationships among terms based on statis-
tical associations derived by computer. In this case words which tend to co-occur frequently in text are linked together. A purpose for storing relationships among terms, whether in a thesaurus, intermediate lexicon, or groupings of co-occurring terms, is to provide a tool useful in the query development process. The inquirer's initial choice of terms can be expanded to include additional terms which have previously been identified as related in some way.

Relationships among documents.

Garfield (1968) has pointed out that the traditional approach to IR system design has been to treat individual documents as independent entities. Using citations, however, one has available linkages between documents provided by authors. This pattern of document-document coupling may be represented as a network. Classifications, whether derived from manually assigned classification numbers or from automatically generated clusters, can likewise be used as a basis for explicit representation of relationships among documents. A purpose for storing relationships among documents is to support browsing. Once a document of interest to the inquirer has been identified, other related documents can be displayed for his inspection or added to the retrieved set automatically.

2. Artificial intelligence approaches

Within the last few years, as many AI researchers have focused their attention on the task of designing expert problem solvers, they have identified a number of key questions. Addressing these questions in system design creates a kind of knowledge engineering (Winston, 1977):

1. What kind of knowledge is involved?

2. How should the knowledge be represented?
3. How much knowledge is required?

4. What exactly is the knowledge needed?

Choice of representation (the answer to question 2) cannot be made in isolation from consideration of answers to the other questions. Some knowledge is procedural, while other knowledge is factual. Some representations will work adequately for small sets of facts, but break down for large sets. Perhaps a combination of representations is required for a particular application. The features of several forms of representation are described below under categories suggested by Winograd (1977):

a. Predicate calculus (Green, 1969a)—a small number of possible structure types are used in a general way to describe knowledge. Facts are expressed as atomic statements and quantifiers are used to express more complex facts. The representation is frequently chosen for theorem proving tasks. Its advantage lies in modularity and the main problem is efficiency because in a large data base there may be no mechanism for easily identifying the particular facts relevant to solution of a given problem.

b. Simple programs—involve a separation between program and data as opposed to the more uniform representation of the predicate calculus. They combine special procedures and specific data to give explicit control and high efficiency. They can, however, only do things explicitly planned in the organization of the knowledge base. Thus programs are efficient at the cost of low generality while representations like the predicate calculus are general at the cost of low efficiency.

c. PLANNER-like languages (Hewitt, 1971)—research in AI has led to the development of programming languages which give some of the benefits of a more flexible representation in which knowledge can be represented
and strategies or procedures for solving problems specified. They keep
the efficiency of programs by avoiding general search of a data base while
breaking loose from some of the rigidities of program control required
in conventional programming languages. These languages have a data base
of facts and a set of theorems (programs) embedding more complex knowledge.
Rather than simply stating a fact, a PLANNER theorem states a particular
sequence of actions to be taken if there is a goal of establishing that
fact. The fundamental mechanism that makes PLANNER work is pattern
matching: a PLANNER theorem may use pattern matching to search a data
base for certain expressions and, if it finds them, add a new expression
to the data base or a PLANNER theorem may use pattern matching to search
the data base for other theorems that are designed to add certain facts
to the data base. Theorems are stored with a special index which can
determine which ones match the goal pattern. When a particular goal is
to be attained, the system automatically tries the various theorems
which are indexed as being useful for this goal. PLANNER gives the ability
to be explicit in controlling what theorems are tried by adding a recom-
mendation list to the goal.

d. Production systems (Newell & Simon, 1972, p. 32-33)—the body
of knowledge is represented by a linearly ordered set of rules called
productions. Productions are usually specified by a set of patterns to
the left of an arrow and a set of actions on the right. Thus the left
side is a list of things to watch for and the right side is a list of
things to do. If patterns in memory match the indexing pattern of a
production, then the actions specified in that production are performed.
Depending on its indexing pattern, a production can be triggered by a
combination of facts or patterns in memory. An ordered list of produc-
tions is called a production system. Advantages of this form of representation are its homogeneous representation of knowledge, the incremental growth of the knowledge base possible through addition of individual productions, and the absence of rigid program control.

e. Semantic networks (Bell & Quillian, 1971; Simmons, 1973; Woods, 1975)—a form of representation designed to encode only descriptive information, omitting any explicit reference to plans for action. Networks are structures composed of sets of nodes connected by directed arcs. A semantic network represents concepts expressed by natural language words and phrases as nodes connected to other such concepts by a set of arcs called semantic relations. It thus is a move away from attempts to represent knowledge as a collection of separate, simple fragments. Semantic networks combine in a single mechanism the ability not only to store factual knowledge but also to model the associated connections exhibited by humans which make certain items of information accessible from certain others. The strength of the network form of representation lies more in finding connections than in making use of them.

f. Frames (Minsky, 1975)—the basic object in this system is a frame, a collection of facts and procedures associated with a concept or situation. A frame is a network of nodes and relations in which the top levels are fixed, representing aspects of the situation which always hold. Lower levels or terminal nodes can be thought of as slots to be filled in by facts specific to the situation at hand. Terminals may already be filled in with default assignments which can be replaced by other values in situations where defaults do not hold. Attached to a frame are procedures for using the facts stored.

The representations described above are alternatives available to a
system designer. At present there is no hard theory comparing these representational schemes and capable of indicating which will be the most useful in any particular application (Doran, 1977). To choose from among the representations, one should consider certain evaluation criteria. Jackson (1974, p. 253) has suggested two: (1) epistemological adequacy (sufficient knowledge should be present in the representation for problems presented to the system to be in principle solvable); and (2) heuristic adequacy (a representation should offer ways of avoiding or greatly reducing search, expressing information helpful in solving problems). The possible contribution of a representation to problem solving is explored in greater detail in Chapter V. Another basis for evaluation, important for some applications, is extendability—the ease with which new information can be linked onto previously constructed structures so that the representation can be expanded whenever new information is received. This is necessary if the representation is to support learning as described in Chapter VI.

3. Problems for research

In the discussion of IR approaches four types of internal representation were identified: document representations, query representations, relationships among terms, and relationships among documents. But this view of the IR representation problem is somewhat limited, as suggested by the following additional representation types which are possible:

(1) Microlevel representations—derived from a single document, representing in some detail portions of the text and illustrations;

(2) Macrolevel representations—representations to characterize a data base or file of document representations as a whole, so that choice can be made among data bases;
(3) Representations of "information packages" other than documents which may be the desired objects of retrieval—maps, engineering drawings, audiovisual materials, etc.

In a paper comparing in detail properties of printed and machine-readable files, Pollock (1977) observes that many machine-readable files are essentially electronic copies of printed products. He suggests that it should be possible to enhance the searchability of machine-readable files by modifying their structure or adding extra data. Alternative internal representations already used in various AI applications need to be examined as possible bases for machine-readable files in IR. In particular techniques used in natural language understanding systems can be considered for microlevel representations in IR, as one begins to design systems for question answering and fact retrieval in addition to document retrieval.

Alternative internal representations.

The structure of knowledge no longer lends itself to a linear representation . . . nor even to representation as a "tree". To give a formal representation of this structure now requires the employment of the notions of networks, grids and systems. (Pingenol, 1971a, p. 30)

The review of IR representations indicated that a relatively limited number of structures have been used for internal representations to date; property lists of attribute-value pairs, vectors, and networks are the principal forms. The discussion of AI representations suggests two issues which should be considered as one compares alternative representations for any application: procedural vs. declarative knowledge and structure vs. function.

a. Procedural vs. declarative knowledge—the discussion of alternative formalisms for representation in AI suggested that in many cases one could either represent knowledge explicitly or implicitly, leaving it to be
generated by a procedure. The choice between declarative and procedural representations can be a tradeoff between storage space and time—how frequently must facts or relationships be used before it is worthwhile to store them directly rather than use a procedure for reconstructing them? Simon's (1962) distinction between state and process description is interesting in this regard. A state description need only specify those criteria required to identify an object, e.g. a circle is the locus of all points equidistant from a given point. The elements of a document surrogate would be a state description since they provide sufficient information to permit matching of the surrogate and the query formulation. A process description must provide a means for producing or generating objects having the desired characteristics, e.g. to construct a circle rotate the compass with one arm fixed until the other arm has returned to its starting point. In an IR system, a clustering algorithm is an example of a process description, since it can be used to generate classes of associated documents, replacing assignment of a predefined classification number.

The distinction between declarative and procedural representations includes the contrast between stored relations and algorithmically defined relations. Swanson has observed that "it is in fact far more reasonable to design library representations on the basis of the way in which the users tend to organize the subject matter rather than the way in which indexers imagine that it ought to be organized, yet it seems that this procedure is seldom followed" (Swanson, 1961). The availability of algorithms or procedures by which the inquirer can define relations in online systems can supplement the stored relations already given in the retrieval file. A similar situation exists in question answering systems. Existing
systems tend to adopt a uniform representation, e.g. all statements in
the first order predicate calculus or as relations. Question answering
may be more efficient if declarative knowledge is combined with procedural
knowledge indicating how to use statements to answer questions. Procedures
can be thought of as second order knowledge, specifying ways in which
facts will be accessed or identifying strategies to try in problem solving.
IR system designers seem to have focused on document and query represen-
tations to the exclusion of procedures. This is reflected in the man-
machine dialogue in which the inquirer simply issues a sequence of com-
mands rather than procedure calls. The possibilities of using procedural
representations, as supplements to existing internal representations in
IR systems which are primarily declarative, should be one aspect of the
study of AI representations in IR.

b. Structure vs. function--in making a choice among representations
for an IR system one needs to determine which structures are best suited
to the functions the system is being designed to support. One must con-
sider the types of knowledge to be represented and made accessible, as
well as how to coordinate use of various types of knowledge in the task
of generating responses to queries. The review of representations already
used in IR indicates that they are heterogeneous--relationships among
terms can be displayed in a network while document representations may
simply be property lists. These particular representations illustrate
structures serving two different functions, i.e. generic survey and
specific reference (Vickery, 1971b). Term relationships displayed in a
thesaurus allow one easily to expand a search, while document represen-
tations allow one to locate documents having specific combinations of
features.
Considering first document retrieval systems, the function of the document representation is primarily clue-providing rather than information-preserving. Research is needed exploring the possible uses of representational elements which are author-defined vs. indexer-defined, as well as the role of content vs. context elements. While all of these elements presumably are "clue-providing", it is not clear what combinations should be used in generating a response to a particular query. If research succeeds in identifying relationships between query characteristics and document representation elements leading to good retrieval performance, knowledge of such relationships could be embedded in the pattern-action formalism of productions, for example. The pattern of characteristics of a particular query (perhaps such factors as subject area and desired size of retrieved set) could then lead to selection of certain portions of the document representation as a basis for retrieval. When attention shifts from document retrieval to the information-preserving function of representations in question answering and fact retrieval, it is clear that several AI representations are of potential interest since they have already proved useful for knowledge-based systems in other domains. AI representations also may be fruitful sources of ideas as one begins to investigate representations for which there is no body of established practice, particularly the tasks of developing macrolevel representations and representations of various kinds of "information packages" other than texts (as described on pp. 85-86). In addition AI provides some insight into the problem of representation evaluation. The notions of epistemological adequacy, heuristic adequacy, and extendability (as defined on p. 85) are potential criteria for evaluating IR representations as they relate to functions associated with problem solving and learning.
Natural language understanding.

It has never been assumed that a retrieval system should attempt to "understand" the content of a document. Most IR systems at the moment merely aim at a bibliographic search. Documents are deemed to be relevant on the basis of a superficial description. I do not suggest that it is going to be a simple matter to program a computer to understand documents. What is suggested is that some attempt should be made to construct something like a naive model, using more than just keywords, of the content of each document in the system. (Van Rijsbergen, 1975, p. 137)

The passage from van Rijsbergen indicates a dissatisfaction with the "superficial description" of documents currently found in IR systems and suggests the need "to construct something like a naive model" of the content of each document in the system. This is microlevel representation of content units smaller than the document as a whole. Research to date in AI on representations to support natural language understanding is clearly relevant to this task. Representations for natural language understanding systems can have two parts (Wilks, 1977):

1. complex system structures for representation of text that are significantly different from the "surface structure" of the text;
2. cognate structures representing conceptual and real world knowledge that is not explicitly present in the text.

Each of these is readily interpreted in the context of IR. The "surface structure" of text may contain extraneous words not useful in representing the content of the document and relationships between parts of the text may be masked by the linear character of running text. One therefore is likely to generate structures different from the "surface structure" as document representations. Cognate structures may be required as supplements to text because interpretation of certain words may be subject-dependent, a fact not evident from the text itself. In a multidisciplinary
IR system there may be a need for such cognate structures, because the same word may have widely varying meanings in different disciplines. While AI research on natural language understanding systems can provide guides to the development of representations much richer in structure than the simple indexing languages presently available, the feasibility of "scaling up" must be the subject of future research. Methods to date for building natural language data bases deal with natural language in small quantity; managing large data bases may entail a new order of complexity and may require development of entirely new techniques.

8. External representations

The need to consider external representations at the user-computer interface as distinct from the internal representations within the computer subsystem arises for two reasons:

(1) Representations best suited to machine processing are not necessarily easily interpreted by humans and vice versa.

(2) There is a distinction between tasks performed by the human which external representations are designed to accelerate or augment and tasks delegated to the machine which are performed using internal representations. This section is necessarily more suggestive of areas for future research than descriptive of past work. Most online IR systems have to date had only limited display capabilities; attention in AI has only recently begun to include machine-aided intelligence as well as machine intelligence. Design of the user-computer interface includes decisions regarding both the dialogue language and display formats. While the man-machine dialogue is considered in Chapter V, display formats are considered here as examples of external representations.
1. Information retrieval approaches

While the IR literature shows that IR system designers are aware of some of the functions which online displays might serve in interactive systems, operational systems generally limit the inquirer to printing sample documents and viewing portions of alphabetically and/or hierarchically arranged dictionary files. These displays have been adopted from printed tools, where thesaurus displays can include alphabetic, permuted, hierarchical, and category groupings (Surace, 1970). There is a growing interest in exploring alternative displays, however, as designers work to make systems accessible to a wider group of users. One finds references to "user-oriented" or "user-cordial" interfaces, which emphasize efforts to make the system "friendly" to the inquirer rather than having the inquirer conform to rigid protocols (Goldstein & Ford, 1978). External representations in such systems are used to give the inquirer some guidance in the query formulation process.

Robertson (in press) notes that the inquirer has an image of the problem he wants to solve and of what sort of documents he might find useful for solving that problem. Given the initial linguistic expression of his problem and the image of what documents might be suitable, he then chooses search terms. His choice may be modified if provided with some aid by the system, such as a display of a thesaurus showing term relationships. This view is corrective to an over-mechanistic view of the information retrieval process: the thesaurus serves as an image-modifying device rather than simply a mechanical translation device for converting natural language to the controlled language of the internal document representations. Doyle (1961) has expressed a similar point of view in his discussion of "semantic road maps", displays showing relationships
among terms derived from their co-occurrence in documents in the data base. Each inquirer would find this empirical map of relations somewhat different from his expectations (his own pattern of associations), and to the extent that it is different it conveys information about how his initial selection of terms must be modified to achieve satisfactory retrieval within this particular system.

Displays of terms thus can serve to work against two possible sources of search failure: (1) lack of coincidence between the language of the inquirer and language of the system and (2) a failure to cover all possible approaches to retrieval. Displays can bring the vocabulary of the inquirer into coincidence with the vocabulary used by the system and provide the means whereby the inquirer can vary his search strategy to achieve better results. Forms for external displays can include:

(1) collections of words and phrases which tend to mean about the same for the purposes of retrieval and which may be used in the process of constructing "alternative ways of expressing the same concept" (Swanson, 1960);

(2) hierarchical displays to suggest terms to be used in making the original query either more specific or more general;

(3) mappings linking terms used to describe the same concept in different databases so that in searching multiple databases one knows which elements are common to the two and which change;

(4) a directory of available databases to help orient inquirers by identifying available resources;

(5) search reviews, summarizing the progress of a search up to a certain point;

(6) sample document surrogates matching the initial query formulation.
A sample of items retrieved in response to the initial formulation should reveal how relevant items have been represented in the system and can suggest additional elements to be added to the query formulation.

One caution is that displays may pose too many alternatives without indications of what each can do. Postings data (for each term, the number of documents in the data base in which it appears) are one source of information which can be used in selecting among alternatives.

2. Artificial intelligence approaches

The AI work dealing with external representations can be found in displays used in decision making and problem solving systems. Such systems reflect a shift in focus from delegation of tasks requiring intelligence (e.g. automatic theorem proving systems) to acceleration and/or augmentation (e.g. man-machine theorem-proving). Two studies can be cited as examples of this trend. In exploring various approaches to interactive pattern recognition, Chian (1978) notes that the central consideration of his book is how to present numerical and nonnumerical information in a form most suitable for human observation using computer-generated displays. Bledsoe and Bruell (1974) describe a man-machine theorem proving system in which the theorem being proved is presented on the display screen in a form that the user can easily comprehend and a series of interactive commands allows him to help with the proof when he desires. Just as IR system designers are interested in "user-cordial" interfaces, one finds statements in the AI literature regarding the need to refine the interface to an intelligent agent to make the system appear "comfortable" to the human user (Feigenbaum, 1978). In particular even in delegated tasks, it may be helpful to make the system's reasoning and inference processes understandable to the user. This capacity for explanation has been
provided in MYCIN, a program designed to assist physicians with the selection of antibiotic therapy for patients with bacterial infections (Shortliffe et al., 1975).

3. Problems for research

Interactive systems provide a technology which allows one to link man and machine in problem solving tasks. External representations, the displays at the man-machine interface, are integral components of this link. In IR a number of proposals have been made as to the form displays might take, but little systematic study has been done exploring the options and assessing their utility as aids in query formulation. In AI the relevant work deals with machine-aided intelligence, but it is thus far limited in scope as the interest in such systems is relatively recent. Nevertheless, AI research does suggest two problem areas which can serve as a focal point for future research on external representations:

(1) investigation of the relationship between internal representations and external displays;

(2) mechanisms to represent processes as well as objects and relationships as part of displays.

Internal vs. external representations.

Representations readily processed by machine, such as the predicate calculus, may have limited utility to the human user, whereas representations meaningful to the human user, such as passages of text, may not be easily manipulated by machine. Yet there may be a need to translate from one to the other. By recognizing the existence of two types of representations, external and internal, and their interrelationships, one can seek to refine each separately while also developing mechanisms to translate from one to the other as required for the query formulation process.
Process description.

The ability to represent and interpret processes will become increasingly important as IR systems delegate more tasks to the machine. To enable the inquirer to exercise some control over the query formulation, it should be possible to provide him with a description of processes, the way in which the system makes certain decisions such as which database to search. Such explanations are useful both in providing the inquirer with a better understanding of what is going on and in giving him the information needed to intervene in the process should he wish to.

AI researchers developing knowledge-based expert systems have found that the ability of a system to explain its actions can be a key to its acceptance by users.
CHAPTER V

PROBLEM SOLVING

Problem solving is the art of using knowledge effectively for the attainment of desired goals. Problem solving can be approached using either algorithms or heuristics. An algorithm is a method guaranteed to find a solution, given enough time. Where no algorithm is known or the known algorithm is impractical, one must resort to heuristics which are rules of thumb, strategies, or short cuts aiding discovery of a solution. In contrast to well-structured problems such as game playing situations, one is often confronted with ill-structured problems. In IR the problem confronting the system is to identify, in response to each query, the appropriate answer. In document retrieval systems the answer is the portion of the data base retrieved, while in question answering systems the answer is a specific fact or statement. In IR problem solving includes development of a search strategy. Application of heuristics in IR can include use of techniques which allow one to quickly select the portion of the data base satisfying the query. In general questions which inquirers bring to retrieval systems can best be regarded as ill-structured problems which become well-structured only in the process of being changed to a form which the computer subsystem can handle.

This discussion of problem solving suggests three possible areas of application of AI in IR:

1. Question answering as a theorem proving problem

   Identify questions which can be answered when posed as theorem proving problems.
2. Heuristics in IR

Examine what form heuristics may take to aid searching in IR.

3. Ill-structured problems

Consider how to design the user-computer interface to facilitate conversion of initially ill-structured queries to a well-structured form which the system can process.

A. Question answering as a theorem proving problem

Within IR one type of problem solving activity is question answering. Raphael (1976, p. 194) has suggested that an ideal question answering system should be able to:

1. accept facts and questions, and make appropriate responses, all in the form of natural English;

2. store, remember, and make efficient use of a large amount of data—at least thousands of elementary facts;

3. answer questions that require it to figure out the logical consequences of the facts stored explicitly in its memory;

4. operate conversationally.

At present a given question answering system exhibits only some of these capabilities. In particular, systems which view question answering as a theorem proving problem tend to excel at determining logical consequences of facts stored in memory. These systems are described in some detail under AI approaches, and possible comparisons between this approach to question answering systems and others being pursued by AI researchers are suggested as problems for research.

1. Information retrieval approaches

Fact retrieval or question answering systems in information retrieval have usually been one of two forms: special format or text based. Examples of both types are given in the reviews by Simmons (1965, 1970).
Special format systems have a specific, often narrowly defined format used to represent information in the data base as well as questions addressed to the system. The only information used is that which fits their particular format. They are thus designed to handle only a given set of facts and a given restricted query set. Text based systems are much like document retrieval systems in that text excerpts are retrieved based on similarity to search requests. Simple word matching may be used, or more sophisticated syntactic and semantic processing may be applied to develop representations for document and query text which can be compared.

Text based systems are more general than special format systems since the same procedures can be used for different subject areas, whereas databases and queries in special format systems are usually restricted to very narrow subjects. Text based systems are still somewhat limited as question answerers, however, because they employ a rather weak form of inference. Passages are retrieved based on similarity of term occurrence in query and document text, under the assumption that this indicates that answers to questions are likely to be found in retrieved passages. Before turning to a discussion of question answering as a theorem proving problem in AI, it is useful to summarize advantages and disadvantages of general and special purpose systems.

Special purpose or problem dependent question answering systems can be characterized as follows:

1. class of questions asked is small, completely specified in advance, and concerned with a particular subject;

2. special question answering subroutines may be optimized for the data base and question class.
Advantage: can achieve good performance in terms of memory utilization and running speed.

Limitation: loss of flexibility and easy extendability because questions dealing with more than one subject area cannot be handled without development of additional subprograms.

General purpose or problem independent question answering systems have the following characteristics (Green, 1969b):

1. Representation language is well defined, unambiguous and rather general so one can hope to describe many subjects, questions and answers;

2. Inference mechanism is subject independent, so to describe a new subject or modify a previous description only some portion of the data base needs to be changed and it is not necessary to make changes in the program;

3. Range of questions allowed is broad.

Advantage: automatic theorem provers are becoming more efficient so one can more readily derive answers not stored explicitly in the file. Loveland (1978) provides a comprehensive survey of automated theorem proving.

Limitation: generality is achieved at some expense. Much of the data base is irrelevant to answering a given query. For the general approach ultimately to be successful it will be necessary to provide a means for representing semantic or problem dependent information in the system to use in limiting search for an answer.

2. Artificial intelligence approaches

In general purpose question answering systems both representations and inference mechanisms are subject-independent, but there is still a choice available to the system designer. Examples of two types of general
purpose question answering systems are given below as illustrations of
question answering as a theorem proving problem.

Predicate logic and resolution.

The general representational language is first order predicate
calculus which may be used to express statements about widely varied
subjects. Axioms can take two forms. Data axioms are instantiated
predicate expressions, e.g. FATHER(Sam, Sara). General axioms are used
to define predicates and their interrelations, e.g. FATHER(x,y)∧FATHER(y,z)→
GRANDFATHER(x,z). The associated general deductive mechanism is a theorem
prover based upon the resolution inference principle (Robinson, 1965).
The resolution rule of inference embodies in a single rule instantiation
mechanisms and simplification rules. Using resolution in question answer-
ing, the deduction problem is posed as the problem of deducing a contra-
diction from a finite set \( S = A \cup \neg \neg Q \) of clauses where \( A \) is the set of
clauses in the data base and \( \neg \neg Q \) is the negation of the query. Individual
steps of deduction are all of the form \( A \vdash \neg \neg C \vdash B \) where \( A \) and \( B \) are clauses
either in \( S \) or already deduced from \( S \). \( C \) is a resolvent of \( A \) and \( B \).

Contradiction takes the form of the empty clause \( \Box \). Deduction of clause
\( \Box \) from a set \( S \) of clauses has the form of a tree with a clause at each
of its nodes; at each of the initial nodes is a clause in \( S \); at the term-
inal node is the empty clause and the clause at each noninitial node is
the resolvent of clauses at the two immediately preceding nodes. The
resolution principle combines the following ideas (Slagle, 1971):

1. syllogism principle of propositional calculus
\[ \neg a \lor b \land a \lor c \rightarrow b \lor c \]

2. instantiation principle of predicate calculus
from formula \( F(v_1, \ldots, v_n) \) one may infer formula \( F(t_1, \ldots, t_n) \)
obtained by substituting terms \( t_1, \ldots, t_n \) for variables \( v_1, \ldots, v_n \) respectively. One can denote a substitution of a set of terms for a set of variables by \( \sigma = \{ t_1/v_1, \ldots, t_n/v_n \} \).

As an example consider the following:

**Axioms**

A1. Mr. Jay's shipment is on car 69 or on car 78.

\[ \text{ON}(J, \text{C69}) \lor \text{ON}(J, \text{C78}) \]

A2. Car 78 is empty.

\[ \neg \text{ON}(x, \text{C78}) \]

A3. Car 69 is at Atlanta.

\[ \text{AT}(\text{C69}, \text{Atlanta}) \]

A4. \( \forall x \forall y \forall z \left[ \text{ON}(y, z) \land \text{AT}(z, x) \right] \rightarrow \text{AT}(y, x) \)

\[ \neg \text{ON}(y, z) \lor \neg \text{AT}(z, x) \lor \text{AT}(y, x) \]

Q: Where is Mr. Jay's shipment?

Prove \( (\exists x) \text{AT}(J, x) \) 

Negation of query: \( \neg \text{AT}(J, x) \)

The proof answers two questions: Does an \( x \) exist? If so, what is it? The deduction is represented below as a proof tree. Clauses forming the nodes of the tree are from one of three sources: (1) axioms retrieved from memory; (2) negation of the query; (3) resolvent of two clauses.

\[ \text{ON}(J, \text{C69}) \lor \text{ON}(J, \text{C78}) \quad A1 \]

\[ \neg \text{ON}(x, \text{C78}) \quad A2 \]

\[ \text{ON}(J, \text{C69}) \]

\[ \neg \text{ON}(y, z) \lor \neg \text{AT}(z, x) \lor \text{AT}(y, x) \quad A4 \]

\[ \sigma = \{ J/y \} \]

\[ \neg \text{AT}(\text{C69}, x) \lor \text{AT}(J, x) \]

\[ \neg \text{AT}(J, x) \quad \neg Q \]

\[ \sigma = \{ \text{Atlanta}/x \} \]

\[ \text{AT}(\text{C69}, \text{Atlanta}) \quad A3 \]

A contradiction (\( \square \)) has been found, proving the existence of an answer.

In order to get an explicit answer, it is necessary to extract information from the proof tree, as described by Nilsson (1971). In this case the answer is that Mr. Jay's shipment is at Atlanta, \( \text{AT}(J, \text{Atlanta}) \). The resolution principle is complete in answering questions, i.e. if a
question is answerable, then an answering clause will be generated by repeatedly applying the resolution principle (given enough time and computer memory).

Relational data bases.

In semantic networks or relational data bases concepts are represented by nodes and relationships by labeled connections between nodes. Consider the example below from a data base describing family relations among individuals (Minker, 1975):

```
   1   2   3   4   5
F    M    F    M    
|    |    |    |    |
1---2---3---4---5
W    M    W    M    
|    |    |    |    |
6 ---7 ---8 ---9 ---10
F    F    H    
|    |    |    |
S = sibling
```

The data axioms or facts depicted in the above graph can be expressed as $F(1,2), F(1,3), \text{etc.}$. If a relation is listed explicitly in the data base, questions asking whether the relation exists can be answered readily. Where information is implicit, the answer must be inferred from the given relations. Given a relational data base system, it is possible to define new relations in terms of specified relations. Rules for relational composition can be given such as $P_1(x, y) \land P_2(y, z) \rightarrow P_3(x, z)$. The process of defining new relations in terms of other relations makes it possible to derive facts implicit within the relational data base and make them explicit by storing new relations in the data base. The process of defining new relations in terms of given relations can be thought of as inferring new facts.

The binary relations, e.g. $F(1,2)$, represent facts known to the system (extensional data), where arguments of the relation become nodes in the graph and binary relations become labeled directed arcs. To infer
relationships other than those listed explicitly it is necessary to have some general rules (axioms) which serve to relate facts to one another (intensional data). The problem of deriving an inference becomes one of making new facts explicit by using both extensional and intensional data. Let us supplement the graph given on the previous page by some general axioms: (1) \( w(x, y) \land m(x, z) \rightarrow f(y, z) \); (2) \( m(x, y) \land m(x, z) \rightarrow s(y, z) \); (3) \( f(x, y) \land f(x, z) \rightarrow s(y, z) \); (4) \( h(y, x) \land m(x, z) \rightarrow f(y, z) \); where (1), for example, can be interpreted: "If \( x \) is the wife of \( y \) and \( x \) is the mother of \( z \), then \( y \) is the father of \( z \)." Axiom (1) can be used with data axioms \( W(2, 6) \) and \( M(2, 11) \) to infer \( F(6, 11) \) even though this is not shown explicitly in the original data base. One could therefore draw an arc from 6 to 11, labeling it \( F \). By applying axioms 1–4 at all nodes successively until one can no longer derive additional inferences, the entire graph may be completed. Once complete, all facts are explicit and no additional inferences can be made using the given general axioms. Unless new axioms are provided or new data are added, no new facts can be inferred.

In relational data bases one must handle the problem of solving for the answer to a question given in the form \( R(x, a) \), \( R(a, x) \) or \( R(x, y) \) where \( x, y \) are variables for which values are to be determined and \( a \) is a constant. A single relational expression is solved if there is an entry in the extensional file matching the relation. The question answering algorithm first gives preference to extensional data, i.e. the facts in the system, and then intensional axioms are applied to derive an answer when there is no explicit answer in the data base. Hence whenever a problem is to be solved for which an explicit answer exists in the data base, the answer is found immediately.

Types of questions.

It is possible to identify four types of questions to which question
answering systems might be designed to respond:

(1) Class A: requires a yes or no answer (closed questions);
(2) Class B: requires a "where is", "who is", etc. answer, e.g. John is in Paris, John is Mary's husband (open questions requiring answer construction);
(3) Class C: requires an answer in the form of a sequence of actions: Go to New York by train and then go to Paris by airplane.
(4) Class D: requires an answer involving testing of conditions (branching): If a bus is available go to New York by bus; otherwise go to New York by train.

Theorem proving techniques may be used in answering all of the above classes of questions as illustrated by Chang and Lee (1973). While question answering systems have generally been restricted to classes A and B, with class C considered to be problem solving and class D program synthesizing, all four cases could be capabilities made available in a more general question answering system.

To summarize this section, it is useful to review the formulation of question answering as a theorem proving problem. The correspondence is the following:

(1) knowledge of the world is expressed as axioms, facts necessary for question answering;
(2) questions asked are presented as theorems to be proved;
(3) the process of proving a theorem becomes the process of deducing the answer to a question.

3. Problems for research

Although it is clear from examples given in the previous section that question answering can be posed as a theorem proving problem, further
research is necessary to determine if question answering systems in
information retrieval are a practical application area for theorem proving
techniques. Certain differences between question answering systems and
more common theorem proving applications should be noted. Facts in the
question answering system's file have the status of nonlogical axioms—
propositions true not for purely logical reasons but because they are
observations made in the world or simply because they are assumed to be
true (Sparck Jones & Kay, 1973). Finding answers to questions about
ordinary objects and events usually requires a relatively short chain of
inference to be constructed on the basis of facts chosen from a large
data base of nonlogical axioms. Proof of a nontrivial theorem, on the
other hand, typically requires much longer chains of inference constructed
on the basis of a much smaller inventory of already known facts.

Three problem areas for research can be identified, two within the
question answering as a theorem proving problem paradigm and one comparing
this paradigm with others currently being explored. If theorem proving-
based question answering systems are to satisfy Raphael's criteria for
an ideal question answering system (given on p. 98), it is necessary to
look at the extent to which natural language statements can be expressed
in the predicate calculus and the form heuristics might take to guide
search for answers in large data bases. It is also of interest to compare
theorem proving systems with other question answering systems in order
to identify potential strengths and weaknesses of each.

Adequacy of predicate calculus.

Of any representation language one must ask if it permits expression
of a useful subset of natural language. Question answering systems to
date have for the most part investigated simple relations readily expressed
in the predicate calculus, e.g. set membership, set inclusion, part-whole, family relations. Axioms for limited topic areas can usually be expressed with predicates, variables, and constants. While axiomatization of mathematical topic areas in predicate calculus is common, there is relatively little experience with how to handle natural language. Representation of natural language is a two step process (Sandewell, 1971):

(1) Formulate pieces of the linguistic conceptual framework in predicate calculus, e.g. adjective-noun composition, adjective comparison.

(2) Formulate axioms which express properties of various concepts.

Development of general purpose question answering systems in IR based on the theorem proving paradigm will depend in part on how easily natural language can be translated into predicate calculus and what types of knowledge are readily expressible in this formalism. The possibilities and limitations of this use of predicate calculus must be the subject of further study.

Heuristics to guide search for answers.

For at each stage when one is using a logical system there is a very large number of alternative steps, any of which one is permitted to apply, so far as obedience to the rules of the logical system is concerned. These choices make the difference between a brilliant and a footling reasoner, not the difference between a sound and a fallacious one (Turing, 1950).

Although the discussion of question answering as a theorem proving problem suggests that such a formulation is a natural one, there has been no consideration as yet of the problems which may arise in the search for a proof tree using large data bases. A need for heuristics to guide search for answers to questions occurs in two areas:

(1) data base clause selection—to determine the sequence in which to select clauses from the data base;
(2) Inference mechanism—to determine the order in which to perform deductions by selection of clauses for resolution. The resolution rule of inference tells one how to derive new clauses from a specified pair, but it does not tell one how to choose which clauses to resolve. Heuristics are needed to guide and control selection of clauses for resolution, avoiding time lost producing clauses not needed in the proof. While it may be theoretically possible to locate ∅ in the implicit search space defined by the inference system, axioms in the data base, and query clauses, a practical question answering system has time and space constraints. In a practical system there may be thousands of clauses stored in the data base. The search strategy must restrict the search to the most relevant axioms in the data base.

It is possible to categorize heuristics or search strategies used in question answering. They direct search by restricting the search space (e.g., forbidding some resolution steps) and/or by ordering the search (giving preference to some resolution steps). The former is called a refinement strategy and the latter is an ordering strategy (Nilsson, 1971). Choice of clauses may be based strictly on structural characteristics (e.g., clause length) or it may be based on semantic characteristics (e.g., interpretation of the meaning of clauses).

To compare the effectiveness of various strategies, some performance measure is needed. One measure for evaluation of search strategies is penetrance (Fishman, 1973). Penetrance of a search \( P = \frac{L}{T} \) where \( L \) is the length of a path from start to the goal node and \( T \) is the total number of nodes generated by a search. A strategy is perfect if \( L = T \) and \( P = 1 \); otherwise \( P < 1 \). Other performance measures include amount of computer memory and time required. Such measures allow comparisons of heuristics
when analyzing their performance in different question answering situations. Comparative studies.

The general purpose question answering systems utilizing automatic theorem proving represent one approach to applying AI techniques to the task of question answering. While automatic theorem proving techniques are highly developed, other AI approaches designed to avoid some of the potential weaknesses of systems based on theorem proving have also been explored. These alternative approaches are described briefly to suggest the motivation for comparative studies.

One approach is to embed general representations in a larger problem solving or question answering system. It is not always natural to describe the whole of a problem by writing axioms in the predicate calculus and it is unnatural to have as the only inference mechanism a uniform and subject-independent rule. Using resolution on subproblems may prove to be more efficient than a resolution-only approach. Consider, for example, STRIPS (Stanford Research Institute Problem Solver) (Fikes & Nilsson, 1971). In STRIPS the current state is represented by a set of clauses in first order predicate calculus. Operators correspond to actions transforming one state into another. The problem solver is given the starting state and the goal state as inputs. It then attempts to find a sequence of operators transforming the initial state into the goal state. In the process of a search for solution the problem solver can call the theorem prover to answer questions concerning which operators are applicable to a state and whether or not goals are satisfied by the current state. The problem solver must transform the given initial state into a state in which the given goal can be proven true. Thus STRIPS represents an approach in which the resolution theorem prover is a module in the
program combined with other problem solving methods.

A second approach to question answering focuses on natural language processing. A representative system, QUALM (Lehnert, 1978), is distinguished from question answering systems that are motivated by information retrieval or automatic theorem proving. If the latter attempt to answer questions phrased in natural language, they are designed in two pieces: (1) a memory retrieval system and (2) a natural language interface. Rather than treating natural language processing as a "front end", QUALM seeks to model the process by which humans understand and respond to natural language questions, developing a theory of question answering.

The contrast between question answering systems focusing on automatic theorem proving and natural language processing reflects the distinction between work in performance mode and work in simulation mode (p. 8). STRIPS is one of possibly many hybrid approaches to question answering and problem solving. Lehnert (1978, p. 266) suggests that it is useful to differentiate "technical information processing" (e.g. retrieving responses from a large data base of facts) from "conceptual information processing" (e.g. story understanding). She states that it is too early to predict how technical and conceptual processing strategies relate to each other and where the boundaries of their knowledge domains lie. Clearly the term "question answering" can encompass a wide range of systems serving many different purposes. Comparative studies are needed to begin to define the areas within which the various approaches to question answering being explored in AI prove to be most fruitful.

8. Heuristics in IR

Heuristics are things that aid discovery of solutions to problems. Heuristic reasoning must not be regarded as final and strict, but as
provisional and plausible only whose purpose is to discover the solution to the present problem (Polya, 1957, p. 113). The following definitions suggest the differences between algorithms and heuristics:

Algorithm— an ordered, finite set of rules for finding the solution of a mathematical or logical problem in a finite number of steps; decision processes which are guaranteed to produce the solution being sought given enough time.

Heuristic— rule of thumb, strategy, method or short cut which limits search for solution in large problem spaces. While it may facilitate arriving at a solution, it does not guarantee that a solution, even if it exists, will be found.

The need for heuristics arises because systems (whether human or machine) have limited processing resources relative to the complexity of situations with which they are confronted. In IR one uses heuristics to locate the answer to queries addressed to the system. In AI heuristics have been applied in the solution of a variety of problems. Some heuristics are special purpose, developed for use in solving a particular problem. Others are general purpose with wide applicability. Study of heuristics in AI—their form, acquisition, and use—can be of particular interest to designers of interactive IR systems which can support heuristic searching.

1. Information retrieval approaches

The situation in machine systems is no different from that in manual systems where success in responding to a query is highly dependent on the indexing, classification, and organization of available material, i.e. the representation. The form of a query which the machine can process depends on the internal representation of its data base, both the elements
of a document surrogate and their organization.

Since use of such terms as "search strategy", "query formulation", "retrieval rule", etc. varies in the literature, it is helpful to characterize each in some detail for discussion of information retrieval as a problem solving process. A search strategy is made up of two parts: (1) the query formulation, specification of query elements and a method of combination; (2) identification of the portions of the data base(s) to be searched. Once the portion of the data base to be searched is identified, each item in this subset is compared with the query formulation and retrieved or not depending on the retrieval rule applied. The retrieval rule may also be thought of as an inference mechanism, since if the decision is made to retrieve an item from the file, the system has in effect inferred that this item is a response to the query. Some examples should clarify these distinctions and illustrate the close relationships among choices of internal representation, search strategy, and retrieval rule. In particular, if multiple data bases are searched, a different query formulation may be required in each.

Search strategy

Query formulation:

Query elements, like document representation elements, can characterize content (controlled vocabulary or free text terms) and/or context (author, journal title, etc.)

Methods of combination can include:
(1) linking terms with Boolean operators (AND, OR, NOT)
(2) arranging terms in a list or vector

Portion of the data base(s) to be searched can include:
(1) all of a single data base
(2) multiple data bases
(3) the cluster with the centroid most similar to the query formulation (when documents in a clustered data base and queries are represented as vectors)
Retrieval rule
Those used to date in IR systems utilize measures of similarity such as those discussed in Chapter III. An item is retrieved if the measure of similarity calculated between the query formulation and document representation exceeds some threshold. In the case of queries formed using Boolean operators, document representations satisfying the specified logic are retrieved.

The search strategy can be thought of both as a process, i.e. a series of decisions as to query elements and methods of combination, and as a product recording the outcome of these decisions. In batch processing systems the machine handles only the product, since all decisions as to the form of the search strategy are made prior to running it on the machine. In online systems, however, where the searcher takes advantage of the interactive capabilities of the system, a process description is necessary because decisions can be made in sequence and can be based in part on intermediate search results. This process point of view is evident in recent attempts to model search strategy development and to identify heuristics used in IR. Because there are no algorithms providing rules for search strategy development, the inquirer must employ heuristics.

Markey and Atherton (1978) have identified five strategies used in searching operational online systems which have query formulations made up of terms linked by Boolean operators:

(1) building blocks (subassembly, modular)—develop each facet of the search separately as if it were a subsearch all on its own and then make a final logical assembly of all of these subprograms (which are in effect solutions to subproblems);

(2) successive fractions (divide and conquer, file partitioning)—establish successively smaller sets or subfiles by developing and ANDing facets in succession;

(3) citation pearl growing—use interactive capabilities of the
system: locate a relevant citation, review it and add appropriate terms to the query formulation, iterating until no additional terms are found for inclusion;

(4) most specific facet first—in a multifacet search, begin with the most specific concept and add additional facets only if the retrieved set is too large;

(5) lowest postings first—if the frequency with which terms have been used in indexing is known, the facet having the lowest postings may be entered first. If the resulting postings are low enough, the search may stop at this point.

Choice from among these strategies may depend on the goal of the search, e.g. high recall, high precision, or a brief search to locate sample citations. While these alternative approaches have been identified, it is not known at present whether they simply reflect differences in searching “style” or whether there are specific types of questions for which each is most appropriate. In addition, there may well be other approaches used by searchers but not included among the five models described above.

The models of search strategy described by Markey and Atherton are macrolevel plans for the whole search. In order to analyze strategies in greater detail one must look at the microlevel, “information search tactics” which are moves made to further the search (Gates, 1973). Examples are “specify” (search on terms as specific as the desired topic) and “cleave” (apply the binary search principle). Gates suggests that these general tactics may have wide applicability in many specific search situations. Analysis of search strategies and tactics can provide some insights into heuristics employed in IR. To study this area more
systematically, however, one can benefit from an understanding of AI approaches to heuristics.

2. Artificial intelligence approaches

Problem representation.

A problem exists whenever a problem solver desires some outcome or state of affairs that he does not immediately know how to attain. Most work on automatic problem solving in AI centers on the notion of search. The basic idea is that solving a problem involves selecting a good sequence of actions from the set of possible actions. Determination of the best sequence is accomplished by tracing through consequences of alternatives, a process referred to as search.

In the previous chapter the concept of representation was introduced as a formalism for knowledge possessed by the system. Another aspect of representation emerges when one turns to the task of problem solving, for problem definition may also be viewed as a form of representation. There are two search-oriented approaches:

(a) state space approach--the state space representation for a problem consists of a specification of the structure of a state, a description of the start state(s) and the goal state(s), and a set of operators which map sets of states into states (Vanderbrug & Minker, 1975). To solve a problem using this approach, one successively applies operators to currently generated states until the goal state is reached. This may also be viewed as the problem of finding a path through a graph from the initial state to the goal state, where the states are nodes and applications of operators are denoted by arcs.

(b) problem reduction approach--the problem reduction representation for a problem consists of a specification of the structure of a problem
description, a description of the original problem, a characterization of the problems whose solutions are immediately known (primitive problems), and a set of operators which map problems into sets of problems (VanderBrug & Minker, 1975). It is a strategy of partitioning a difficult problem into two or more simpler problems (subproblems) each of which may be further partitioned. The hope is that ultimately the initial problem can be partitioned into subproblems of very little difficulty. The problem reduction space can be illustrated conveniently by an AND/OR graph.

The task then becomes one of identifying a subgraph which represents the solution to the problem graph. The AND/OR graph gives a hierarchical goal structure. The initial problem node is the goal and to nodes at each level belongs either a disjunction of subgoals, any one of which would attain the supergoal (OR node), or a conjunction of subgoals, all of which must be attained (AND node).

The following table offers a convenient summary of the relationships between these approaches:

<table>
<thead>
<tr>
<th>Problem reduction</th>
<th>State space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problems</td>
<td>States</td>
</tr>
<tr>
<td>Primitive problems</td>
<td>Initial states</td>
</tr>
<tr>
<td>Initial problems</td>
<td>Goal state</td>
</tr>
<tr>
<td>Problem reduction</td>
<td>State space operators</td>
</tr>
</tbody>
</table>

State space and problem reduction can be contrasted as follows: the state space representations search from fragmented initially known pieces of information (initial states) toward the goal (goal state) while problem reduction representations search from the goal (statement of the original problem) toward fragmented pieces of information (primitive problems).
whose solutions are initially known. Since state space and problem reduction representations can be viewed as being related by the direction in which the problem space is searched, one may ask in which direction (initial to goal, goal to initial or both) it is better to search. This depends on the nature of the problem but certain criteria can be suggested (VanderBrug & Minker, 1975): (1) the direction in which operators are most easily applied; (2) the direction in which heuristic information is most easily used; (3) location in the problem space which contains the most information; (4) number of goal nodes.

It should be noted that it is not necessary that the main problem and all subproblems be attacked using the same representation. The problem reduction method is a useful device for representing very global aspects of a problem, while the state space approach may be of more use when more specific problems are attacked (Hunt, 1975, p. 217).

**Heuristics.**

The development of problem solving programs in AI has led to the recognition of several heuristics, some of which are problem specific and some of which have wider applicability. Special purpose heuristics are often obtained from intimate knowledge of a particular problem area. One approach to discovering heuristics may therefore be to try to learn from experts the heuristics used in solving problems in a particular task domain. AI has also led to the discovery of certain general purpose heuristics, applicable to the solution of several different problems. Two examples can be given from CPS, the General Problem Solver (Ernst, 1970):

1. **Means-ends analysis:** proceeds by dividing the problem into easier subproblems. The problem solver compares what is given with what
is wanted, producing from this comparison a set of descriptions of how the two situations differ. Operators are then selected that eliminate these differences.

(2) planning: if the goal is to transform a into b, abstract a and b, eliminating most of their details and creating new objects (abstractions or images) a' and b'. Now formulate the goal of transforming a' into b'. This new problem will generally be easier than the original. Once it has been solved, its solution is used to guide the solution of the original unabstracted problem. A plan thus acts as a rough outline of a possible solution. Planning gives the problem solver a technique to analyze the problem structure in the large using a simplified model of the problem situation.

The following techniques are among those used by heuristic programs to limit search for a solution:

a. Evaluation functions—mechanisms used to select among alternatives in the search for a solution. Any search heuristic other than random or the systematic exhaustive one involves a presumption that certain areas of the problem space are more likely to contain the solution than others (Cooper & Elithorn, 1973). Evaluation functions reflect this presumption in two ways; they can answer either the question, where shall I search next?, or the question, when shall I stop search and accept the solution as satisfactory? They can thus serve either as a steering mechanism for search or as a "satisficing" criterion for terminating search by estimating the gain to be expected from further search. Satisficing involves looking for good or satisfactory solutions instead of optimal (Simon, 1969, pp. 64-65). As has already been noted, one usually cannot within practicable computational limits generate all possible alternatives and compare their
respective merits. Nor can the best alternative be recognized before all have been seen. One satisfies by looking for alternatives in such a way that an acceptable alternative can generally be found after only moderate search. The search proves sufficient, not necessary, for attaining the goal and the result is not necessarily unique. By drastic search limitations the best solution may be overlooked, but the only alternative is exhaustive search. The use of an evaluation function to guide search often introduces human ad hoc knowledge on characteristics of promising search strategies and the relative weights to be assigned. In AI, for example, game positions are evaluated to determine the next move to be made. Existence of evaluation functions allows a search to proceed by trial and error; the success of each attempt at a solution is assessed and allowed to improve subsequent attempts until a solution acceptable within defined limits is reached.

b. Control—the issue of control of a search is closely related to evaluation functions since outputs from the evaluation process yield the flexibility in control characteristic of heuristic searches. Control specifies the order of operations to be performed in the search for a solution. Evaluation functions determine control of the search at particular nodes, but more global control is evident in choices between breadth vs. depth first searching, order of application of operators, etc. Bruner et al. (1956) in their studies of concept attainment noted that some orders of inquiry are better than others for human problem solvers. A number of things can be gained by varying order: (1) opportunity to obtain information appropriate to the objectives of the inquiry; (2) increase or decrease in cognitive strain involved in assimilating information and keeping track; (3) control degree of risk involved.
Given a number of different types of knowledge available in the representation, control determines when to use each.

6. Constraint satisfaction—heuristic methods use not only knowledge embedded in the representation, but also details included as part of a particular problem statement to limit the search space. Each added condition in the problem statement is one more item that can be exploited in finding a solution.

In addition to characterizing the form of heuristics, it is helpful to identify some principles for their effective use (Pohl, 1977):

(1) principle of robustness—heuristics are approximate uses of knowledge. They should be employed in schemas which can recover from error should the application of a particular heuristic fail to lead to a solution.

(2) principle of adaptability—heuristics tend to be locally good. Their most effective use is adaptive (dynamic), applying them when conditions are appropriate.

These are reflected in the use of online systems. A user interacts with the computer in general accordance with some plan of action. At each step the user evaluates the computer's output, reformulates his ideas concerning possible approaches to problem solution, and decides among alternatives for subsequent courses of action using heuristics appropriate at that point.

3. Problems for research

With the growing availability of online IR systems, frequent reference has been made to their ability to support heuristic search. It is necessary to clarify the notions of heuristic and heuristic search in this context. The discussion of heuristics in AI suggests three areas in need
of further research: their form, their acquisition, and their effective use in problem solving in IR.

Form of heuristics.

The description of heuristics in AI gives one a vocabulary to use when analyzing the form of heuristics. The distinction between general and special purpose heuristics, as well as the notions of evaluation function, control, and constraint satisfaction, must be interpreted in the context of IR to determine their applicability to that domain.

While special purpose heuristics developed to aid problem solution in other AI application areas are unlikely to be of interest in IR, it may be possible to use general purpose heuristics such as means-ends analysis and planning. As one example of means-ends analysis, consider the use of index language devices and search language devices to increase flexibility in subject searching. Certain devices group terms together into classes and increase recall while others allow one to increase possible shades of meaning and increase precision. Index language recall devices include: (1) control of synonyms; (2) control of quasi-synonyms; (3) control of word forms; (4) hierarchical grouping; (5) grouping by statistical association. Index language precision devices include: (1) coordination of terms; (2) weighting; (3) links to indicate syntactic relationships; (4) roles to indicate functions of terms (Lancaster, 1972, p. 121). These can be supplemented by search language devices. In searching recall devices include: (1) use of OR logic; (2) truncation; (3) "exploding" a search term by automatically searching on subordinate terms in the hierarchy; (4) expanding a search to include words statistically associated with the initial search terms. Search language precision devices include: (1) use of AND logic; (2) use of information
regarding the relative position of terms in text; (3) use of weighted term searching with a threshold to limit retrieval. Use of recall and precision devices can be viewed as an example of means-ends analysis. Consider, for example, a situation where a searcher is not satisfied with the number of citations retrieved and his goal is to retrieve more items from the file which still have a high probability of being relevant. Any of the recall devices listed above could be selected to reduce the difference between the set of items already retrieved and the set which the searcher would like to retrieve. Use of this heuristic could be mechanized by developing a table identifying the effects of various devices available for means-ends analysis in retrieval. Although actual selection of the device to be used may still be left under the control of the human searcher, the machine could consult its table to determine what devices are available to reduce the difference identified by the searcher. While the various devices noted above contribute to recall or precision, further research is needed to characterize the effect of each on retrieval more precisely. When known, this information can be embedded in the machine to aid in choice of devices through means-ends analysis. Other general purpose heuristics already identified in AI, such as planning, should be examined to determine their relevance to IR as well.

Use of heuristics incorporating evaluation functions, control, and constraint satisfaction was described above under AI approaches. These can likewise be interpreted in the context of IR:

a. Evaluation functions—features to be used in evaluation functions are defined by the task area. Within a data base, evaluation functions can be used to select features for inclusion in the query formulation. One might choose to include terms based on the number of postings, for
example. Evaluation functions can also be used to aid in making the initial choice among data bases as part of a data base selector. In this application the evaluation function ranks data bases according to their applicability to the query as stated by the inquirer (Williams & Preace, 1977). Evaluation functions can assist the inquirer in determining where to search next (within a data base or in switching from one data base to another) and in deciding when to stop searching and accept the result as satisfactory.

b. Control—in searching there are two aspects of control. At the global level control is choice of the appropriate search strategy for the given query; at the local level control is choice of the particular search tactic to be used at each point. The discussion of search strategies and tactics (pp. 113-114) shows that some effort has been made to model ways in which human searchers control a search. Building blocks and successive fractions are two alternatives at the global level; selection of “specify” as a tactic for choosing terms is an example of control at the local level.

c. Constraint satisfaction—the inquirer’s specification of certain parameters as part of his query can reduce the set of alternatives to be explored. Examples of parameters are: (1) weights to reflect the relative importance to him of various elements in the query; and (2) the context that surrounds each inquirer (education, place of work, etc.). Such constraints must be expressed in a form that the machine can handle. Cuadra and Katter (1968), for example, have cited the need for a descriptive language for use orientation to reflect the inquirer’s intent in asking a question with respect to how he expects to use the information received.
It is evident from the above discussion that evaluation functions, control, and constraint satisfaction all are aspects of the form which heuristics can take in IR. Research is needed to more completely define the role of each in interactive IR systems.

**Acquisition of heuristics.**

If machines are to actively assist inquirers in the search for answers to questions, a way must be found to incorporate heuristics into machine systems. There are two possible approaches, work in simulation mode vs. work in performance mode. In simulation mode an attempt is made to model the heuristics used by "expert" problem solvers. In the case of IR one would like to identify techniques used by intermediaries. Since much of the planning of search strategies such as choice of which data base(s) to search may go on before the intermediary actually begins searching online, it is necessary to study the whole process of search strategy development beginning with the statement of the question by the inquirer. One can ask the problem solvers (in this case the intermediaries) to "think aloud" as they develop a search strategy, thus providing an indication of the basis for various decisions. The study reported in Chapter IX provides examples of protocol analysis based on dialogues between intermediaries and inquirers which have revealed some heuristics used in selecting data bases and terms for particular queries. Once heuristics used by human problem solvers have been identified, one must find ways to build them into the machine. The alternative source of heuristics, derived from work in performance mode, focuses on the computer subsystem. Working with the internal representation, one explores ways of using it to solve the problem at hand without regard to heuristics used by human problem solvers. In the case of document retrieval systems, for example,
calculation of a similarity measure between the query vector and centroids of clusters of document vectors as a basis for limiting search to a portion of the data base represents a heuristic based on exploitation of available internal representations rather than any attempt to model human processes. Both types of study, work in simulation and in performance mode, must be pursued to generate ideas for heuristics which can be incorporated in machines.

Use of heuristics.

The process of problem solving in IR includes query formulation (query elements and methods of combination), identification of portions of the data base(s) to be searched, and application of a retrieval rule. Since there are no algorithms for search strategy development, heuristics are used in each part of this process. As noted on p. 120, when working with heuristics it is important to consider the need for robustness and adaptability in system design. This has been acknowledged in existing online systems which allow the searcher to break the search down into search statements which represent a sequence of steps in the development of a search strategy. If certain tasks prove to be unfruitful (e.g. no postings are found), they can be abandoned. Since development of the search is stepwise, the strategy can be adapted as one proceeds with the search. At present application of heuristics in this process depends on the initiative of the human searcher. Once a repertoire of heuristics has been acquired by the machine, how they can best be used in search strategy development must be investigated.

The appropriateness of problem reduction and state space representations as frameworks within which to view the use of heuristics must also be explored. Strategies identified by Markey and Atherton (1973)
already provide some examples of their usefulness. Building blocks is a problem reduction approach, dividing search strategy development into a number of subproblems. Successive fractions is a state space approach, isolating successively smaller portions of the data base until the goal is reached. It is evident that operational IR systems do not constrain all queries to fit within a single framework. Choice from among alternative search strategies is itself heuristic. To do this automatically it is necessary to further explore the suitability of particular search strategies for particular types of queries and/or document representations.

C. Ill-structured problems

The ill-structured nature of most requests for information may go unrecognized because current systems require that a query be well-structured in the format acceptable to the system before processing can begin. The query formulation as processed by the machine is often simply an explicit goal statement: find the set of references indexed under a set of terms satisfying conditions specified in the query formulation. The "method" of solving this problem is then trivial for the machine; it simply compares the query formulation to each document representation in the file and retrieves those meeting the specifications. The discussion of problem solving and heuristics in this chapter has thus far dealt with well-structured rather than ill-structured problems. In so doing many of the difficulties which arise when inquirers bring queries to retrieval systems have been ignored. If one hopes to implement machine techniques to assist in the process of search strategy formulation, it is first necessary to recognize these difficulties and to explore to what degree machine systems can be expected to handle them.

1. Information retrieval approaches

The first distinction to be made is that between questions and
commands. In existing batch processing retrieval systems, for example, the search strategy takes the form of a command to the machine, specifying precisely what elements are to be searched and their logical relationships. A command assumes either (or both) of two things on the part of the inquirer: (1) he knows exactly what he wants and can describe its form (book, paper, etc.) and its label (author and title or specific subject); (2) he knows the organization of the system, the rules for formulating queries so that they may be obeyed as commands (Taylor, 1968). In response to a command the system locates a specific set of items satisfying the command and the process ends. The use of commands in communicating with the system ignores the whole process of question negotiation which precedes statement of the command. Question negotiation begins with an initial statement of the problem which may be highly unstructured and loose. In existing systems the inquirer, perhaps with the aid of a human intermediary, must transform this initial statement into a command which the system can follow. The main distinction between the way one programs a computer and the way one instructs a human is this: the machine is instructed mainly in the form of a sequence of imperative sentences, while the human is instructed mainly in declarative sentences describing the situation in which action is required and a few imperatives that say what is wanted. Each has certain advantages (McCarthy, 1959):

Advantages of imperative sentences: (1) Procedural description in imperatives is already laid out and is carried out faster. (2) One starts with the machine in a basic state and does not assume previous knowledge on the part of the machine.

Advantages of declarative sentences: (1) Advantage can be taken of previous knowledge. (2) Declarative sentences have logical consequences
and it can be arranged that the machine will have available sufficiently simple logical consequences of what it is told and what it previously knew. (3) Meaning of declaratives is much less dependent on order than is the case with imperatives. This makes it easier to have afterthoughts. (4) The effect of a declarative is less dependent on the previous state of the system so less knowledge of this state is required.

Existing systems exploit the advantages of imperative sentences. New techniques must be developed to facilitate machine handling of declarative sentences. Such a capability is necessary if the machine is to participate in the question negotiation process now carried out by intermediaries, in which the inquirer is prompted to qualify and specify his search.

The distinction between command and question is closely related to that between well- and ill-defined (or structured) problems. In general problems which inquirers bring to retrieval systems can best be regarded as ill-structured problems which become well-structured only in the process of being prepared in a form which the system can handle. Much problem solving effort is therefore directed at structuring problems and only a portion of it at solving problems once they are structured.

2. Artificial intelligence approaches

Many tasks studied in AI have been well-defined or well-structured, as suggested by the formulation of problems in state space or problem reduction representations. A problem may be regarded as well-structured to the extent that it has some or all of the following characteristics (Simon, 1973):

(1) criterion for testing a proposed solution and a mechanizable process for applying this criterion;
(2) at least one problem space in which can be represented the initial problem state, the goal state, and all others that may be reached or considered in the course of attempting the solution of the problem;

(3) results of applying operators can be represented;

(4) knowledge acquired by the problem solver can be represented in one or more problem spaces.

While there is probably a continuum of well- to ill-definedness, several hypotheses have been put forward on the general characteristics of ill-structured problems (Newell, 1973):

a. Open constraints (Reitman, 1965)—aspects of the problem such as givens, goals, operators which are simply not specified in the problem statement. These are "open" and have to be "closed" by the problem solver before he can work on the problem. The problem statement does not determine how such constraints are to be closed. Instead, the problem solver must reach out into his general knowledge to decide how to close these constraints. The idea of open constraints emphasized that one cannot artificially restrict the context in which problems get solved. All problems depend, not only on the specific problem statement, but also on the range of background information that the problem solver must be presumed to have in his head. Open constraints are taken as the locus and source of ill-definedness in problem statements since values assigned by one problem solver to close open constraints may be unacceptable to another. When one closes open constraints in a problem description given to him by someone else, one is in effect behaving with respect to a hypothesis of what is wanted.

b. Weak methods—there is no difference between well- and ill-structured
problems except that a problem solver will call a problem ill-structured if he has only weak methods that apply to it. The designation depends on the individual problem solver, as well as on the collection of methods available. A general problem solver has a collection of weak but widely applicable methods such as heuristic search and match. The generality of a method is judged by the size of the set of problems to which it can be applied. The power of a method, on the other hand, is judged by such factors as the probability of solution, the quality of solution, and the amount of resources used. There is usually an inverse relation between the generality of a method and its power. One way to move toward well-structured problems, therefore, is to make available to the problem solver more powerful methods. Newell and Simon (1976) point out that it also may be possible to select an appropriate representation which makes problems easier to solve. This suggests two levels of problem solving. The higher level is concerned with finding improved representations of the problem to be solved. The lower level is finding a solution within the chosen representation. This approach suggests a shift in interest from search techniques to the question of how to represent large amounts of information in a fashion that permits effective use.

c. Problem specification—problems are ill-structured to the extent that the activity in solving the problem goes on concurrently with the activity in formulating the problem. The problem solver does not engage in the act of specification all at once at the beginning of the problem session followed by problem solving proper. Rather both problem solving and problem definition go along together. Thus the problem retains its character of being ill-structured until the final solution is reached.
3. Problems for research

If the inquirer is to successfully use online IR systems without the assistance of intermediaries, it is necessary to consider how to design the user-computer interface to facilitate conversion of initially ill-structured questions to a well-structured form which the system can process. One aspect of interface design has already been addressed in the discussion of external representations in Chapter IV. Remaining problems for research include identification of the ways in which questions addressed to an IR system can be ill-structured and development of a man-machine dialogue to convert ill-structured to well-structured queries.

Sources of ill structure.

The three characterizations of ill-structured problems as suggested by studies in AI also have significance for queries in IR.

a. Open constraints. Some examples of open constraints which must be handled in query negotiation include:

(1) ambiguity—Two or more meanings may be given to a word or string of words. Clarification may be possible by obtaining context information, such as the inquirer's area of expertise (e.g. physics vs. medicine).

(2) analogical attributes (Reitman, 1365)—These specify an object and a relationship between that object and some component of the problem. In retrieval an example would be to find all documents "like" one already known to the inquirer. The question remains: what is a reasonable measure of likeness? One could compare documents using subject headings or citation sets, for example. If one is to find an x similar to y in some respect, it is first necessary to acquire or assume additional information about the nature and degree of similarity referred to in the problem
description. In the case of IR the system can either ask the inquirer to specify the basis for comparison or the choice can be made automatically.

The ill-structured nature of queries brought to a retrieval system means that clarification of concepts may become an important part of the problem solving process. Recognizing that queries are apt to have open constraints, systems can be designed to prompt the inquirer for clarification and to lay out the possibilities when there are multiple interpretations.

b. Weak methods. The inquirer using an IR system without the aid of an intermediary may have to manage with weak but general methods for locating the subset of the file to be retrieved. As suggested in AI, one approach to making the problem well-structured is to make available more powerful methods to the problem solver. In the context of IR such methods could be, for example, search strategies already developed by previous inquirers to find answers to similar queries.

c. Problem specification. In an online system the problem can be considered ill-structured because it is often the case that the activity in solving the problem goes on concurrently with the activity in formulating the problem. The original query formulation may be modified on the basis of information found during the course of the search. The heuristic search process can include the following steps: (1) select terms using vocabulary displays provided by the system; (2) combine terms into various logical statements; (3) see postings; (4) view selection of retrieved citations at the terminal, looking at indexing terms assigned; (5) modify the strategy (where necessary) after examining selected citations, including new terms. In particular, the use of request sets (storage of query fragments) allows query formulation to go on concur-
rently with problem solving. By changing query formulation from a single process to a set of processes, request sets both simplify the task of query formulation and lessen the need for preplanning (Martin, 1974). The searcher knows that he can switch to browsing through document representations and can pursue interesting tangents suggested by these representations. Query fragments that do not prove helpful do not need to be included in the final query formulation. Often the searcher is unable to make his information need determinate without first seeing some items which he identifies as relevant.

The three sources of ill-definedness identified by researchers in AI are seen to have significance in IR as well, and various approaches to aiding the inquirer in the task of formulating his query have been suggested above. An interactive system is important in this regard, for the computer is to aid in formulation of the inquirer's problem. It needs information only the inquirer possesses and it must check back that it is working on the inquirer's problem rather than substituting a different problem of its own.

**Man-machine dialogues.**

The tools we are trying to use and the language or notation we are using to express or record our thoughts are the major factors determining what we can think or express at all. (Dijkstra, 1972, p. 364)

Discussion of ill-structured problems and the need for query negotiation leads naturally to the question of designing the language for the user-computer interface in retrieval systems. A major issue is whether the inquirer can express his query without resorting to utterances which are not included in the subset of natural language interpretable by the system; as Watt inquires, "Will people find that sublanguage 'habitable'?" (Watt, 1968, p. 339).
The man-machine dialogue can take many forms, e.g. English (natural) language, limited English, form-filling, menu-selection (J. Martin, 1973). Any form except natural language may not be "habitable" in that inquirers may find it difficult to use as a medium for expressing their queries. AI research dealing with natural language processing is therefore relevant to the design of the user-computer interface, for it can provide a means with which to make the dialogue language more "English-like". A program to process natural language as part of a man-machine dialogue confronts four highly interrelated problems (Jackson, 1974, p. 294): (1) a syntax problem; (2) a semantics problem; (3) an inference problem; (4) a response generation problem. Overarching these four is the integration problem—determining how these problems are interrelated and what use a language processing program can make of this interrelation. In natural language processing, programs must use syntactic, semantic, and inferential components in an interlaced manner. Rather than consider a sentence, parse by syntactic processes, attempt to assign meaning using semantic routines, etc., the program must use available semantics to disambiguate at the syntactic level, use information based on partial parsing to disambiguate meaning, and, if necessary, use the inferential component to enable what has gone on before in the discourse to select possible meanings and parsings (Winograd, 1972).

The generality of a natural language processing program can be judged on at least two dimensions: the range of (1) syntactic structures and (2) subject areas that it can successfully handle. While AI natural language processing systems presently cover a limited range on both of these dimensions, this has not precluded their successful application as a natural language "front end" to a data base management system.
(Harris, 1977) is a software preprocessor that translates English language requests in the form of questions or commands into the formal query language of a database management system (DBMS) and translates the DBMS response back into English for the user. ROBOT has demonstrated that AI techniques are sufficient for the "microworld" of database query. Because questions addressed to a document retrieval system may require more complex processing than queries of a DBMS with a well defined content and record format, additional research will be needed to develop natural language processing programs capable of handling queries for more diverse IR systems. But natural language processing systems like ROBOT demonstrate the utility of a natural language interface which eliminates the inquirer's need to know the internal structure of a system, so that he can begin with an ill-structured query. Either the computer automatically converts the initial query to a well-structured form for subsequent processing (e.g. Harris, 1977) or a man-machine dialogue can be used to elicit from the inquirer the information needed to eliminate the sources of the query's ill-definedness (e.g. Brown, 1977).
CHAPTER VI
LEARNING

System elements such as feature selection routines, representations, and heuristics are of course all initially selected and programmed by a human designer when an AI system is developed for some application. Learning mechanisms by which a system can improve its performance over time as a result of experience are therefore necessary so that the initial design does not circumscribe system capabilities. Basic to learning are the ability to evaluate performance so that improvement can be judged and some way to store and utilize the results of previous experience. The availability of online computer systems makes it reasonable to speak about dynamic systems which change and improve performance over time. Learning in retrieval systems can have either short term or long term effects. Short term learning is the modification of system responses during the processing of a particular query in order to retrieve items most likely to meet the needs of the inquirer. Long term learning includes changing the representation by modification and/or extension, to improve system performance over time. There are thus two possible areas of application of AI in IR:

1. Short term learning

Examine which AI learning techniques may utilize feedback to modify system response to individual queries.

2. Long term learning

Examine which AI learning techniques can be used for representation modification and/or extension to change system response to new queries.
This chapter varies slightly in format from that found in Chapters III-V.
A general discussion of IR approaches and AI approaches to learning is
followed by the section on problems for research which is subdivided
into two parts: short term learning and long term learning.
A. Information retrieval approaches

For information retrieval the important distinction between
computer-based retrieval and other modes based to some degree on the
printed page is that a computer-based representation can be easily
updated. In contrast to dynamic update, the printed text and bound
volume have a much slower cycle of revision. As a result, most manual
retrieval systems tend to be object-oriented (static) rather than
inquiry-oriented (dynamic). This creates a conflict since a system is
created in anticipation of needs that are not fully known. Yet the
measure of adequacy of a system is the ability to satisfy inquirers'
needs as they arise (Lipetz, 1966). Whether any automated system can
achieve high performance depends to some extent on the ability of the
system to adapt to the needs and expectations of the particular user
population being served. Learning can be used to accomplish this since
it is (Andreee, 1969):

(1) directional—a learning system seeks goals or at least recog-
nizes a goal when it is reached;

(2) transitional—introduces changes with time.

To test for directionality one needs a measure of performance of the
system with respect to a particular task in a particular environment.
On the average for a learning system this performance measure should
increase with time while the system adapts to the task environment.
It is not immediately clear what factors this performance measure should
include in an online IR environment, but definition of such a measure (or measures) should be part of attempts to introduce learning capabilities into retrieval systems. Recall (the percent of relevant items in the data base that are in fact retrieved by the system) and precision (the percent of items retrieved by the system which are relevant) are two examples of performance measures that have been used in information retrieval.

Learning in retrieval systems can have either short term or long term effects. Short term learning is illustrated by the use of feedback in query processing. Feedback is a method of controlling a system by reinserting into it results of its past performance. If the information which results from performance is able to change the general method and pattern of performance, one has learning. In IR this is evident in the view of document retrieval as a trial-and-error process (Swanson, 1977). A query is no more than a guess about the attributes a desired document is expected to have. In general, the response of the system to the query is used to correct the initial guess for another try.

In online systems each inquirer develops a search strategy to locate items relevant to his query. Once done, no attempt is made to store results in some form that would be useful to later inquirers. There may be a good deal of learning on the part of the human inquirer as to ways to better form strategies, but none of this learning is incorporated in the machine. Implementation of learning with longer term effects would mean that each search could have two results: (1) a response to the given query and (2) a record of the operations used to produce this response (Dodd, 1971). The rationale for saving earlier actions which have led to a favorable outcome (or for discarding earlier
actions which have led to an unfavorable outcome) is that one takes the
outcome value as evidence that the given earlier action was good (or bad)
in some more general sense than that it happened to lead to this parti-
cular result in this particular case.

8. Artificial intelligence approaches

In AI various approaches have been taken to give systems learning
capabilities. A system which learns can increase its ability to produce
correct responses in classification tasks, refine heuristics, and/or
modify and extend its representation as experience is accumulated. The
following are examples of available learning techniques:

a. Incremental adaptation—if an evaluation function is made up of
features with coefficients denoting weights to be attached, programs
learn by successive modifications of weights in the evaluation functions
based on favorable or unfavorable outcomes. The basic problem with this
kind of learning program is that once the program has been run, one has
only numerical values of some parameters. The information in such an
array of numbers is unstructured, for the weight associated with each
feature may depend so much on what other features are also included in
the function that each number in itself has no separate meaning. It may
be useful as a terminal learning scheme once a good set of features has
been identified.

b. Learning by building descriptions (Winston, 1975)—from various
examples the program determines the minimal description of a concept
(e.g. an "arch") in terms of objects and their interrelationships.
Counterexamples and near misses are used as sources of information about
what parts of the description are most important. Learning is structural
rather than an update of numerical parameters.
c. Trial and error—an ad hoc technique that involves a series of questions at each decision point: Have I been in this situation before? If so, what did I do? What were the consequences of my action? If satisfactory, choose the same action again. Otherwise, choose something else. The technique is limited in scope since there is no generalization from experience. Situations are assessed separately without attempting to link them together into meaningful categories.

d. Procedural (method) learning—learn procedure or technique which allows one to deduce facts. It can replace storage of data in a table in memory by a function which allows determination of an outcome for particular values of variables, e.g. learning the procedure for addition rather than table lookup.

e. Learning with a teacher (with supervision, with performance feedback)—the most elementary type of learning with a teacher occurs in the development of pattern classification algorithms in which each pattern in the training set has been assigned a class label by the teacher. A more sophisticated version would be an advice taker. In this case human experts could gradually transfer their skills to a machine system through online interaction. Another type of learning with a teacher is book learning. This was used, for example, by Samuel's (1969) checkers program which stored a record of actions taken by experts in particular situations. These could be checked in determining the next move to make.

f. Learning without a teacher (without supervision, without performance feedback, cluster seeking techniques, mode seeking techniques, clumping techniques)—in this pattern classification task the training set of patterns does not have associated class labels. Techniques sort the set of unlabeled heterogeneous patterns into classes such that patterns
in each class are "similar". The training patterns must be typical of
classes from which they come in order to provide information for creating
useful subsets and a sufficiently accurate description of each class to
be able to correctly classify patterns not yet seen.

g. Aggregation and generalization—aggregation is accomplished by
chunking pieces of information into larger units and generalization is
a form of learning where results of one successful problem solution are
saved in a form applicable to solution of subsequent problems or a number
of particular solutions are used to develop a single solution schema.
Aggregation can yield new operators of greater power to be used in the
solution of new problems. Sequences of basic operations become new
chunks useful in the construction of even more complex combinations of
operators, since the sequence can be labeled with an identifier and
manipulated as a unit. Storage and use of aggregates is a form of learning
that can reduce solution time for similar tasks as well as allowing for-
mation of more complex solutions. Such a form of learning is possible,
however, only if the system can represent aggregate operators as part
of its knowledge base. The compact representation of results of previous
experience requires an adequate descriptive language to permit general
statements about both problem domain matters and problem solving methods
(Minsky, 1958a). The system must be able both to package a successful
method for efficient later use and be able to recognize situations in
which such methods are applicable.

C. Problems for research

Learning techniques have thus far been used primarily in experimental
IR systems, examples of which are given below. Online IR systems provide
an appropriate environment within which to investigate both short and
documents from among those retrieved. An alternative mode casts the computer in the role of advice taker. In this case the inquirer instructs the computer as to which terms to include in a query formulation. This can occur at each of two stages: presearch and postsearch (Lesk & Salton, 1971). Presearch interaction consists of investigating the indexing vocabulary in order to decide which terms should be included in the query. Postsearch interaction involves examination of a sample of retrieved documents as a basis for decisions on how to modify the query. Oddy (1977) describes a program THOMAS which has a well developed postsearch interaction stage, during which the computer can accept two types of advice from the inquirer. In addition to making relevance judgments, the inquirer can: (1) suggest new names or terms not previously occurring in the query formulation or appearing in sample document representations; (2) select from authors or terms in sample document representations those of primary interest. Terms of no interest can also be identified as such.

The two activities of making relevance judgments vs. giving advice through selection and suggestion of terms illustrate a range of inquirer involvement as teacher. It is possible to build a graded set of feedback methods ranging from automatic procedures which make only minimal demands on the inquirer and are suitable for novices to systems permitting sophisticated interaction for advice giving. The possible roles of the inquirer as teacher should be investigated further. In systems designed to accommodate a variety of users, it may prove best to leave open a range of possibilities; the machine can perform more tasks automatically when handling requests of a novice, while allowing the expert to develop his own search strategy making use of his own experience and available external representations. Another type of learning with a teacher is "book learning".
To determine the forms which this can take in IR, it is necessary to study approaches to long term learning which can include storing the strategies of expert searchers for subsequent use by other inquirers.

2. Long term learning

There are two approaches to making changes in the internal and/or external representations of an IR system which can lead to different answers in response to new queries. Representation modification can include changes in the content and structure of representations; representation extension can include a variety of augmentations to serve as new aids to problem solving. Each of these was anticipated by Bush (1967) in his revised description of the memex, as noted below.

**Representation modification.**

The memex notices (we have to use such terms; there are no others) that nearly every time [its master] pursues the trail there are a series of items on which he hardly pauses. It takes them out of the main trail and appends them as a side trail. It also notices that when he comes to a certain item he usually goes off on a side trail, so it proceeds to incorporate this in the main trail. (Bush, 1967, p. 96)

In Chapter IV, it was noted that internal representations in IR systems can include representations for documents, relationships among documents, and relationships among terms. Each of these can be subject to modification in a learning system.

Indications of needed modifications in document representations are provided by data obtained from monitoring query formulations developed by inquirers. Data can be accumulated on: (1) frequency of use of different feature types (e.g. author, subject heading); (2) terms appearing in document representations which are used in searching; (3) terms used by inquirers which are not included in the document representations. This data can be used as a basis for evaluating the adequacy of existing
document representations, since good retrieval performance cannot be achieved where there is a large discrepancy between terms used by inquirers in searching and terms appearing in the document representations. Research is needed to determine how data accumulated in this way can be used, either automatically or with human intervention, to modify document representations to achieve better retrieval performance.

Relationships among documents can be established by clustering, also referred to as learning without a teacher. Before any queries are processed against the database, the only items available to be clustered are document representations. Once many queries have been processed against the database and saved as part of the internal representation, additional information is available to guide organization of the database. Queries can be clustered and the resulting query clusters can be used to cluster the document collection in one of three ways (Yorona, 1971): (1) all documents correlated highly with the centroid of a query cluster form a cluster; (2) all documents correlated highly with one or more queries of a query cluster form a cluster; (3) all documents judged relevant to one or more queries of a query cluster form a cluster. The third way has the underlying assumption that documents jointly relevant to one query are likely to be jointly relevant in the future. Using document clusters based on query clusters, groups of documents are readily available to other inquirers who make similar queries at a later time. The retrieval performance resulting from the use of document representations vs. query representations as a basis for clustering should be investigated.

Development of relationships among terms to represent a concept can be considered an example of learning by building descriptions. In IR
these include: (1) common search facets, such as the set of terms relevant to the concept of "higher education" in searching an education database (Markey & Atherton, 1973), and (2) "hedges" as defined in MEDLARS, agglomerations of terms that cut across conventional genus-species hierarchies (Lancaster, 1972). These could either be built by a single expert or developed from the collective experience of several searchers. In addition to including as part of the internal representation search fragments developed by human searchers, ways to derive them automatically (e.g., by identifying statistical associations among terms) should be investigated.

Representation extension.

(The memex) can build trails for its master. Say he suddenly becomes interested in the diffusion of hydrogen through steel at high temperatures, and he has no trail on it. Memex can work when he is not there. So he gives it instructions to search, furnishing the trail codes most likely to have pertinent material. . . . In the morning its master reviews the new trail, discarding most of the items, and joining the new trail to a pertinent position. (Bush, 1967, p. 96)

One would like to permit query formulation not to be simply a product of the given query and operations on it, but instead to provide a mechanism so that query formulations could be supplemeted with relevant material selected from past experience. One must be able to "package" successful searches for efficient later use. The AI learning techniques of aggregation and generalization are of interest in this regard. Recalling that query formulation using an interactive system is often a trial-and-error process, it is unlikely that simply storing the trace of the man-machine dialogue is likely to prove very useful to future inquirers. Learning using aggregation includes extraction of terms linked by "OR" from these traces to be stored as a chunk. This can be displayed for
the inquirer who enters one of the component terms, seeking possible additional terms for inclusion in his query formulation. Saving all query formulations indiscriminately is not likely to be particularly helpful, because some may be in error or too specialized to be of any interest to most inquirers. Learning using generalization includes comparison of chunks containing common terms to develop aggregates likely to be of general usefulness and eliminating groups of terms unlikely to be of help to other inquirers. This type of learning has been little studied to date in IR. SUPARS (Atherton et al., 1972) did have a file of stored searches. In response to entry of any keyword, the search strategy file could be accessed to obtain a listing of all previous search strategies that incorporated that keyword or simply all unique words found in the search strategies associated with this word. As an experimental system, SUPARS could afford to save all searches entered by inquirers. But it is probable that designers of large scale systems will find the AI learning techniques for aggregation and generalization of interest as approaches to maintaining a manageable and useful file of stored searches.

Procedural (method) learning, noted in the discussion of AI learning techniques, also can have a role in representation extension in IR systems. The sequence of commands issued by the inquirer is the procedure used in developing the answer to his query. By gathering statistics one can develop a model of "typical" procedures used in searching an IR system. These models (i.e., sequences of commands) can provide the basis for error detection and adaptive prompting to assist inexperienced searchers of an IR system (Penniman & Perry, 1975).

The above discussion of approaches to short term and long term learning in IR has provided an indication of the relevance of a variety
of AI learning techniques. Most of these have received little attention
to date in IR system design. If initial design is not to circumscribe
system capabilities, research on learning must be conducted concurrently
with research on pattern recognition, representation, and problem solving.
CHAPTER VII

RESEARCH PROBLEMS: AN OVERVIEW

The best part of our knowledge is that which teaches us where knowledge leaves off and ignorance begins. (Holmes, 1896, p. 211)

In the sense of unaware and as yet unlearned, our ignorance and our recognition of that ignorance may be the best motivation both for problem-solving and for creative activity. (Hackerman, 1974, p. 907)

The discussion of AI in IR has focused on four concepts: pattern recognition, representation, problem solving, and learning. Chapters III-VI have considered each conceptual area separately, surveying IR approaches and AI approaches as well as identifying problems for research. Because the conceptual areas are interrelated, it is appropriate to conclude this portion of the dissertation with an overview of the research problems which have been identified. The first part of this chapter describes the "terrain" in general terms, identifying trends in IR and AI which are creating increasingly overlapping areas of concern. The second part of the chapter functions like a "topographical map", providing an overview of the research problems identified in Chapters III-VI as well as an assessment of those areas of investigation likely to be most fruitful, given the state of the art in other AI applications.

A. The terrain

The part of science for the renewal of which the scientist assumes responsibility is surrounded by a sea of information on which he must rely for his enterprise. The scientist may regard his selected field as his "calling," which necessarily includes his submission to the vast area of information and belief surrounding his selected field of inquiry. Each scientist's calling has a different geography.
Each must try to choose a problem that is not larger or more difficult than he can master. His faculties would not be fully utilized if he applied them to a lesser task, and would be altogether wasted on a larger one. The degree of originality any particular scientist trusts himself to possess should thus determine the range which he will venture to tackle and hence also the range of information which he will unquestioningly accept. (Polanyi, 1966, p. 79)

Chapter I concludes with a section on the "limits of discussion", defining the areas of emphasis for this study. To reiterate, this study of possible applications of AI in the design of IR systems has focused on document retrieval and question answering systems accessible online. The emphasis is on the logical or intellectual problems of IR—developing representations, comparing documents and requests, negotiating queries—rather than the physical problems such as details of file organization and data structures. The view of the inquirer has been one of man as an information processor, working at the man-machine interface of online IR systems.

The review of work in AI for the purposes of this study has been limited to those areas which seem to this researcher to be most relevant to the design of online IR systems. It should be recognized that there is a good deal of other work being done in AI in such areas as robotics, computer vision, speech understanding, and game playing. In the future work in speech understanding, for example, may be of interest to interface design in IR systems. But the application areas described under the categories of pattern recognition, representation, problem solving, and learning encompass the research problems of immediate interest.

1. Science vs. engineering

In addition to emphasizing certain content areas of AI, this study has adopted an engineering rather than a scientific point of view of AI.
Two recent statements of goals of AI research illustrate this difference:

The central goals of Artificial Intelligence are to make computers more useful and to understand the principles which make intelligence possible. (Winston, 1977, p. 1)

The ultimate goals of this kind of research are (i) to de-mystify the process by which new science and art are created, and (ii) to build tools (computer programs) which enhance man's mental capabilities. (Lenat, 1978, p. 257)

This dualism in goals is reflected in the two modes of work described on p. 8: performance mode and simulation mode. Findler (1976) has called these the "engineering approach" and the "modeling approach". In the former, the researcher wants to create a system that can deal with intellectual tasks, regardless of whether the methods used are similar to those used by humans. In contrast, the latter has the basic research objective of trying to gain an understanding of the inside mechanisms of a real life system in order to explain and predict its behavior.

Hayes (1973) provides some examples to illustrate the distinction between the engineering and scientific approaches:

Much AI work can be classed as "engineering" or applied AI, as opposed to basic or "scientific" AI. Thus for example one might contrast Waltz's "engineering" approach to natural language comprehension with Schank's "scientific" approach. Waltz's system is designed for use; it is supposed to be able to answer questions put to it in natural English by untrained users. . . . Its linguistic abilities are limited, but usable. In contrast, Schank and his students are concerned to investigate people's general ability to understand the full range of sentences which occur in English. Their work is consciously scientific: an approach to psycholinguistic theory.

One can make similar contrasts throughout AI. Industrial robotics vs. integrated problem-solving robotics research; visual pattern recognition for controlling machine tools vs. computational theories of the human visual system. And yet—and this is my point—both "applied" and "pure" AI involve constructing working programs. For applied AI, the program is the end product; for scientific AI, it is an experiment. (pp. 297-298)
While it is possible to distinguish the engineering and scientific orientations, they are of course not mutually exclusive. Knowledge of how humans perform certain tasks may be incorporated into those programs whose main objective is to accomplish these intellectual tasks as well as possible. Likewise building systems to perform intellectual tasks may provide insights into how humans perform these same tasks.

This study has emphasized "applied AI" using the engineering design process (p. 29) as the method of investigation. In so doing it reflects Garvin's (1957) philosophy with respect to research in machine translation: "Our research philosophy is that an approximate model will do in the beginning and that we will enhance our theoretical understanding by attempting engineering applications and learning from them (pp. 72-73)."

This philosophy is echoed by Sridharan (1978) who notes that "attempting applications is a sound way of extracting and explaining more core AI problems (p. 1)."

2. Information management and knowledge engineering

Variations in usage of the terms "information" and "knowledge" can obscure the fact that there are areas of common concern in AI and IR. In IR there is a growing emphasis on systems for information management, rather than simply information storage and retrieval:

The information research community should, as soon as possible, shift from a preoccupation with document housekeeping and delivery mechanisms to the related, but much broader, problem domain that is based on the need to discover the principles of (and to develop the means for) the optimal husbandry of one of man's key resources, knowledge. (Slamecka, 1975, p. 320)

At the same time one finds an increasing number of references in the AI literature to "knowledge engineering" and "knowledge-based systems". AI applications have penetrated medical practice and research and a variety
of scientific domains. These include: analytical chemistry, synthetic
organic chemistry, protein x-ray crystallography, biochemistry, molecular
genetics, cognitive psychology, and geological prospecting. Design and
development of knowledge-based systems require (Sridharan, 1978):

(1) formulating the application problem—recent efforts view the
problem as being one of utilizing expert knowledge when appropriate.
This represents a shift in emphasis away from the earlier concentration
on search management to the current focus on the construction and use
of knowledge bases;

(2) designing, constructing and refining a knowledge base of expertise—this involves designing the representation appropriate to the
processing that will be carried out and acquiring this knowledge through
formalizing, structuring and making explicit the private knowledge of a
group of experts;

(3) developing schemes of inference, search or problem solving;

(4) winning the confidence of experts;

(5) evaluating and testing the programs—measuring speed, correctness,
competence; tolerance to error in input; robustness against minor defects
in the knowledge base;

(6) developing production versions of the programs—the essential
characteristic at this stage is responsiveness to the user.

Through the design and implementation of such knowledge-based systems,
AI is developing a rich set of concepts about how information can be
represented, processed, and used. By investigating AI applications in
IR, one can try to use these concepts in the development of systems for
information management.
8. The topographical map

Science is the topography of ignorance. From a few elevated points we triangulate vast spaces, inclosing infinite unknown details. We cast the lead, and draw up a little sand from abysses we may never reach with our dredges. (Holmes, 1896, p. 211)

Chapters III–VI have described a number of possible areas of application of AI in IR. Discussion of each area included a summary of IR and AI approaches investigated to date, the "elevated points" from which one can begin to explore the various problems for research which have been identified. From the list of research problem areas given below, it is evident that we have indeed "triangulate[d] vast spaces". The intent of this section is to provide an assessment of those areas of investigation likely to be most fruitful, given the state of the art in other AI applications. Problems for research have been described under the following headings:

Pattern recognition

Automatic indexing as a feature selection problem
  Comparative studies
  Heuristics for feature selection
  Feature selection in short term learning

Pattern classification
  Prototype matching
  Clustering

Measures of similarity
  Similarity based on index terms
  Similarity based on several types of features
  Alternatives to correlation and distance measures
  Interactive pattern recognition

Representation

Internal representations
  Alternative internal representations
  Natural language understanding

External representations
  Internal vs. external representations
  Process description
Problem solving

Question answering as a theorem proving problem
   Adequacy of predicate calculus
   Heuristics to guide search for answers
   Comparative studies

Heuristics in IR
   Forms of heuristics
   Acquisition of heuristics
   Use of heuristics

Ill-structured problems
   Sources of ill-structure
   Man-machine dialogues

Learning

Short term learning
   Incremental adaptation
   Learning with a teacher

Long term learning
   Representation modification
   Representation extension

AI applications in IR can be conceptual or technical. Before reviewing the domains in which "technology transfer" from AI to IR is likely to occur, it is helpful to summarize contributions at the conceptual level which can influence aspects of IR system design. Two types of contribution can be identified:

(1) AI concepts can lead to useful abstractions. For example, the concept of "feature" in pattern recognition is a useful generic term for the diverse elements of a document representation (authors, index terms, citations, etc.).

(2) Conversely, AI concepts can make concrete certain things which formerly were only vaguely defined. For example, although frequent reference has been made to "heuristic searching" of online IR systems, few attempts have been made to operationalize the notion of "heuristic". As seen in the discussion of heuristics in IR in Chapter V, AI concepts
dealing with heuristics are quite helpful in this regard.

At the conceptual level AI thus offers an alternative Weltansicht of IR to that contributed by library science, for example. This new perspective leads to research questions not previously addressed in IR.

Turning to the technical level, it is possible to provide an assessment of those areas of investigation likely to be most fruitful, given the state of the art in other AI applications. A number of different situations can be distinguished:

(1) Few AI techniques are available and/or applicable (i.e. existing techniques are problem-specific). For example, few techniques are available for the design of external representations. While numerous special-purpose heuristics have been developed for various AI applications areas, they are problem-specific and thus unlikely to be applicable to IR.

(2) The feasibility of "scaling up" must be demonstrated if an AI technique is to be applicable in IR. For example, most approaches to natural language understanding deal with "microworlds", i.e. limited subject domains. Whether the same techniques are viable in dealing with more diverse subject areas must be investigated if these techniques are to be transferred to IR systems.

(3) The feasibility of an "appropriate adaptation" must be demonstrated if an AI technique is to be applicable in IR. For example, in question answering as a theorem proving problem it must be possible to supplement automatic theorem provers with heuristics to guide search for answers if they are to work with data bases of facts numbering in the hundreds.

(4) The AI technique can be applied directly without modification. For example, if document and query representations have the form of
vectors, clustering algorithms developed for other AI applications can be used directly.
The above list identifies four different situations, any one of which may confront the IR system designer interested in applying a particular AI technique. Clearly "technology transfer" is most feasible when the technique can be used in its new application area with no modification. At the other extreme there may simply be no technique available at the present time. Intermediate situations are those in which the feasibility of scaling up and/or making appropriate adaptations must be investigated before a definitive judgment can be made as to the applicability of a technique in the new domain.

Determination of the feasibility of transferring technology does not in itself provide an indication of the desirability of doing so. There is a lesson in the notion of "appropriate technology"—technology transfer must be evaluated in light of the objectives of the new application area. This concern underlies the frequent suggestion of comparative studies as one type of problem for research. Once the feasibility of applying certain AI techniques in IR system design has been demonstrated, their contribution to such factors as retrieval performance and ease of system use must also be determined.

In suggesting the potential contribution of AI to IR, we have distinguished between the various types of IR systems—document retrieval, question answering, data retrieval. These are but a subclass of the broader class of information systems which includes systems for teleconferencing, computer-assisted instruction, and decision support. As IR systems change from storage of document representations to storage of full text, it is likely that some of these distinctions will begin
to blur. At present each of the four types of information systems offers one or more capabilities that are not as fully developed in the other systems (Paisley & Butler, 1977): IR systems are capable of storing large files of nonsequential and primarily textual records with many access points per record; teleconferencing systems have a unique capability for interconnecting online users as well as flexible text editing and personal file creation; computer-assisted instruction systems are unique in their response contingent branching and their user-adapted sequence and pacing of information; decision support systems enter further than other types of systems into a user-defined problem, taking in user-supplied data and projecting outcomes of alternative decisions. In addition to the specific problems for research identified on pp. 155-156 and treated more fully in Chapters III-VI, a general problem for research in the future is likely to be how these various specialized information systems can be integrated in support of information work and what role AI techniques can play in this more general information system.
CHAPTER VIII

PROCESSING DOCUMENTS AS QUERIES

In many retrieval situations the inquirer does not come to the system knowing nothing about the literature. When he is already familiar with some citations to relevant documents, the inquirer should be allowed to input the known citations and simply ask the computer to find others "like" them (Lancaster, 1976). In online systems there are available a number of different ways to operationalize "likeness" in terms of the document representation (surrogate) and retrieval rule used. These different operationalizations may lead to different sets of items retrieved when a given document is submitted to the computer as a query. Before choices can be made among alternative operational definitions, there is a need to determine through a comparative study whether such definitions result in different retrieved sets when each is employed. The study reported in this chapter investigates the extent to which the use of alternative internal representations for documents in the INSPEC database yields overlapping sets of items retrieved in response to documents submitted as queries.

A. Abstract model of retrieval

1. Document-term matrix

A retrieval system may be viewed as having three basic components: a set of documents, a set of index terms, and a set of queries. In the situation of interest for this study, the set of queries and the set of documents are identical; any document may be submitted as a query, in response to which the computer must find other similar documents. The
relationship between documents and index terms may be modeled abstractly as a document-term matrix (Lancaster, 1968, p. 40). For example, given two documents \((D_1\) and \(D_2\)) and three terms \((T_1, T_2, T_3)\), a matrix can be drawn, where a 1 indicates assignment of an index term to a document:

\[
\begin{array}{ccc}
T_1 & T_2 & T_3 \\
D_1 & 1 & 0 & 1 \\
D_2 & 0 & 1 & 1 \\
\end{array}
\]

Looking at the rows of the matrix, one finds for each document a record of the presence or absence within its surrogate of each term from the set of index terms. Looking at the columns of the matrix, one can quickly identify all the document surrogates in which a particular term occurs. For any system in which the document surrogate is simply a list of index terms, the document-term matrix is an adequate model of the relations between terms and documents.

Looking at marginal totals across rows and down columns of the matrix provides additional ways to characterize the structure of the file in terms of two design parameters (Travis, 1977): exhaustivity (number of terms per document, given by row totals) and specificity (number of documents per term, given by column totals). Exhaustivity is determined by the person creating a document representation and specificity is determined by the indexing language used (Lancaster & Mills, 1964). Values of these parameters for different representations in the INSPEC data base are given later in this chapter.

2. Extension of the model

With the advent of machine-readable data bases and online systems, it becomes possible to easily search many more portions of document surrogates in addition to index terms. Recognizing this, it is possible to extend the document-term matrix model to better depict the document
representations one finds in most online retrieval systems, as shown below:

\[
\begin{array}{cccccccc}
A_1 & \cdots & A_d & C_1 & \cdots & C_r & I_1 & \cdots & I_s & T_1 & \cdots & T_r \\
D_1 & 1 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 1 & 1 & \cdots & 1 \\
D_2 & 0 & \cdots & 1 & 1 & \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
D_n & 1 & \cdots & 1 & 1 & \cdots & 0 & 1 & \cdots & 1 & 1 & \cdots & 0 \\
\end{array}
\]

This is a document-feature matrix, for the features which make up a document representation may include such elements as authors (A), classification codes (C), index terms selected from a controlled vocabulary (I), and free text terms from title and abstract (T). Another way to extend the matrix model is to allow numbers other than 1 or 0 in the cells, where the numbers can be used to indicate the relative importance of features, for example. While each of the feature types (A, C, I, T) may be thought of as separate components of the document representation, it is also possible to view each type as an alternative internal representation. This has been the approach taken in this study. The question of interest then becomes a comparison of the sets of items retrieved using alternative representations.

A simple example can be used to illustrate two possible situations.

\[
\begin{array}{ccccc}
T_1 & T_2 & T_3 & C_1 & C_2 \\
D_1 & 1 & 0 & 1 & 1 & 0 \\
D_2 & 0 & 1 & 1 & 0 & 1 \\
D_3 & 1 & 0 & 1 & 1 & 0 \\
\end{array}
\quad
\begin{array}{ccccc}
T_1 & T_2 & T_3 & C_1 & C_2 \\
D_1 & 1 & 0 & 1 & 1 & 0 \\
D_2 & 0 & 1 & 1 & 1 & 0 \\
D_3 & 1 & 0 & 1 & 0 & 1 \\
\end{array}
\]

Case 1. Case 2.

In each case assume that \( D_1 \) is the query document and that the pattern of occurrence of features (terms or classification codes) must match exactly for an item in the file to be retrieved. Using alternative representations, the following results are obtained:
Case 1. \( D_1 \) has terms \( T_1 \) and \( T_3 \), so \( D_2 \) is retrieved.  
\( D_1 \) has classification code \( C_1 \), so \( D_3 \) is retrieved.  

Case 2. \( D_1 \) has terms \( T_1 \) and \( T_3 \), so \( D_3 \) is retrieved.  
\( D_1 \) has classification code \( C_1 \), so \( D_2 \) is retrieved.

In Case 1 the two alternative representations are redundant; the same set of items is retrieved using either representation. In Case 2 the two alternative representations lead to different retrieval results.

Recalling the discussion in Chapter V of information retrieval as problem solving, it is clear that one must clarify what is meant by both "search strategy" and "retrieval rule" in the context of the document-feature matrix model. The problem of finding documents "like" the query document is ill-structured until both the search strategy and retrieval rule are defined. Search strategy determines where to look, so in this model the selection of one feature type or some combination of feature types is the search strategy. Given the choice of index terms as the representation of interest, for example, the computer will consider as candidates for retrieval only those documents which have at least one index term in common with the query document. Once the representation to be used has been selected, it is still necessary to specify the retrieval rule. With index terms as the representation of interest, for example, the retrieval rule could be: retrieve a document only if its index terms are identical to those of the query document. An alternative retrieval rule could be: retrieve a document only if it has at least one index term in common with the query document. The result of applying retrieval rules depends on the entries in the cells of the document-feature matrix. To completely specify the retrieval operation, it is necessary to describe both the search strategy and the retrieval rule to be used, i.e., what portion of the document-feature matrix is of interest and how
the data on feature occurrence are to be manipulated. Before discussing the results found for seven different representations in the INSPEC data base, it is of interest to summarize the characteristics of alternative representations in different data bases as reported in the literature.

8. Alternative document representations

1. Representations considered singly

The decision of online system designers to make a number of alternative representations of each document searchable suggests that each is useful in at least some situations. In subject searching the most commonly used representations include classification codes, index terms selected from a controlled vocabulary, and free text (natural language) terms selected from title and/or abstract. Each of these representations has somewhat different characteristics. Classification codes arrange documents according to a logical scheme, often providing a hierarchical grouping of subdivisions of a subject area. In searching, classification codes can be used to define a particular subset of the data base in which relevant documents are likely to occur. Controlled terms assigned by an indexer usually are the most specific terms from the controlled vocabulary descriptive of document content. They are usually more specific than classification codes and provide standard ways of labeling concepts important to a particular subject area. Free text terms, since they are selected manually or automatically from the text of the title and/or abstract, are uncontrolled as to word form and spelling. They can provide access to minor as well as major points of a document, particularly when all free text terms found in title and abstract, except those eliminated by a stop word list, are included in the representation.

The characteristics of each type of content indicator may depend
on the subject area considered. In comparing the relative information content of titles, for example, Buxton and Meadows (1977) found that titles of papers in chemistry and botany had on the average more substantive words than titles of papers in social science and philosophy. Furthermore, most subjects examined showed a significant increase in the number of substantive words appearing in titles between 1947 and 1973. Content indicators are usually supplemented by elements of context, such as authors, journal titles, and citations (Maron & Shoffner, 1969). To the extent that context is related to content, context elements can also be used in subject searching. Studies have shown, for example, that a large percentage of citing papers are related to the subject of the cited paper (Barlup, 1969).

2. Comparisons of representations

Given the properties of representations considered singly, it is of interest to examine some of the relationships which have been found between them. Two types of studies are relevant: (1) those reporting co-occurrence relations between the features of two different types of representations, and (2) those reporting the results of retrieval experiments using at least two different representations. They are, of course, related: co-occurrences of features lead to overlap in the sets of items retrieved using different representations.

Looking at different feature types for patterns of co-occurrence, the most frequent comparison made is between free text terms and assigned index terms. Salton (1973) has observed that when comparing keywords automatically selected from title and abstract with those assigned by subject experts, one normally finds agreement for 60-80% of the assigned terms. Studies comparing title words to assigned subject headings for
Chemical Abstracts (Ruhl, 1964) and Index Medicus (Montgomery & Swanson, 1962) support Salton’s statement. Another type of study deals with relations between the occurrences of content and context elements. Small (1973) found that it was possible to identify relations between words as descriptors and cited references as descriptors, in cases where titles of documents citing the same document contain words in common. What is interesting about these studies is not only that certain patterns of co-occurrence between features used in different types of representations can be identified, but also how the investigators propose to use this information. While more discussion of uses of co-occurrence data is given at the end of this chapter, a summary of conclusions is given here:

1. Where representations tend to be redundant, one may be used in lieu of the other, e.g. the use of free text terms rather than controlled vocabulary terms;

2. Representations may be used in combination, e.g. classification codes to define a subset of the data base, followed by a search on free text or controlled terms strongly associated with that class;

3. Mappings may be established to allow one to switch from one representation to another where the most reliable paths are based on high frequency co-occurrence, e.g. a translation from citations to free text terms closely associated with them.

Any retrieval performance study comparing alternative representations is an interesting extension of the types of studies described above, for the relations among representations, while of some interest in themselves, are ultimately of concern to users in the way they affect retrieval performance. While many studies report comparative performance data in terms of values for recall and precision, that is not sufficient to allow one
to understand the differences among alternative representations. For even if two different representations result in similar retrieval performance as measured by recall and precision, they may not be retrieving the same set of documents. Several studies have specifically addressed the issue of overlap in sets retrieved, however, and the results of some representative ones are summarized below.

The simplest comparison is the case in which one representation is a subset of the second. A comparison using Chemical Abstracts matching search profiles on titles only vs. titles and abstracts found an average loss of 73% of potential output using titles only (Maloney, 1974). Studies comparing retrieval performance of free text vs. assigned terms on the Smithsonian SIE data base (Hersey, Foster, Stalder, & Carlson, 1971) and Nuclear Science Abstracts (Fisher & Elchesen, 1972) found that while assigned terms retrieved a greater percentage of the relevant documents in the data base than did free text terms, each representation retrieved documents not retrieved by the other. For greater recall, therefore, both assigned terms and free text terms should be used in combination. Salton (1971) found that including citations as well as subject terms in document representations led to better retrieval performance than that found when subject terms were used alone. In a study systematically comparing titles, abstracts, controlled indexing terms, and free indexing terms for processing profiles on COMPENDEX (Byrne, 1975), with the results of all four together taken as 100% retrieval, other combinations ranged from 75% (title and abstract) to 21% (controlled terms only). Again the conclusion is that, for complete retrieval, representations must be used in combination. The results of a study on the INSPEC data base presented below also support this conclusion.
C. Data collection

This study utilized an INSPEC data base of 7145 records from Physics Abstracts and Electrical & Electronics Abstracts issued in October 1975. The data base was accessed using the SIRE (Syracuse Information Retrieval Experiment) system (McGill, Smith, Davidson, & Noreault, 1976). The SIRE system was selected because it permits comparison of several alternative document representations using the same system and data base. Since the INSPEC data base consists of records which include title, abstract, classification codes, controlled vocabulary terms, and free indexing terms (selected by an indexer, frequently from the title and/or abstract), it is representative of many other data bases which are commercially available.

Thirty-five documents were selected at random from the data base to serve as query documents. Based on each query document, seven different representations were generated to be processed against the entire data base. The representations, together with their associated retrieval rules, are the following:

1. **Authors**—a list of all authors cited by the query document was created. A data base document was retrieved if it had at least one author from among those cited by the query document.

2. **Classification codes**—the classification codes assigned to the query document were identified. A data base document was retrieved if it had at least one classification code in common with those assigned to the query document.

3. **Controlled vocabulary terms**—the controlled vocabulary terms assigned to the query document were identified. A data base document was retrieved if it had at least one controlled vocabulary term in common with those assigned to the query document.
(4) **Free indexing terms**—the free indexing terms assigned to the query document were identified. A database document was retrieved if it had at least one free indexing term in common with those assigned to the query document.

(5) **Abstract and title**—a stop word list and stemming routine were applied to the text of the query document’s abstract and title to create a list of word stems, together with their frequency of occurrence, which could be treated as a document vector.¹ A database document was retrieved if the cosine correlation between its document vector (based on title and abstract) and that of the query document was ≥ .25.

(6) **Free indexing terms as word stems**—as an alternative to matching free indexing terms with the free indexing term field (as done for representation 4), a stop word list and stemming routine were applied to the free indexing terms assigned to the query document to create a vector of word stems. A database document was retrieved if the cosine correlation between its document vector and the free indexing term vector of the query document was ≥ .25.

(7) **Title**—a stop word list and stemming routine were applied to the text of the query document’s title to create a vector of word stems. A database document was retrieved if the cosine correlation between its document vector and the title vector of the query document was ≥ .25.

As noted above, a single similarity measure has been chosen for each representation, although some representations permit a number of alternatives. Cosine correlation was chosen for representations 5–7 since documents differ in length and in frequency counts for term occurrences.

¹For a detailed discussion of document vectors and correlation measures, see Salton (1968, pp. 236–241). He describes the cosine correlation \( \mathbf{d}_q = \sum d_i q_i \sqrt{\sum (d_i)^2 \sum (q_i)^2} \), the similarity measure used in representations 5–7.
and the formula for cosine correlation takes this into account. The cutoff of .25 gives a manageable number of items retrieved in response to a query document from a database of 7145 documents. Simple overlap was chosen for representations 1-4 since any one record contains only a few author names, classification codes, controlled vocabulary terms, and free indexing terms, and they are not weighted in any way.

The example given below illustrates the seven representations for one of the 35 query documents:

**Document** #3405

**Authors:** Kolers P A. VonGrunau M.

**Title:** Visual construction of color is digital.

**Journal:** Science 187(4178): 757-9, 28 February 1975.

**Abstract:** When disparate shapes are flashed under the appropriate temporal and spatial conditions, the human visual system resolves their disparity smoothly and continuously. No equivalent supplementations are found for color, which the system resolves by abrupt transformation. Shape and color reveal themselves, contrary to some modern theorizing, as properties handled in different ways by the visual nervous system, continuous or analog for shape, abrupt or digital for color.

**Representation 1** (Authors cited by the query document)

<table>
<thead>
<tr>
<th>Anstis S M</th>
<th>Boring E G</th>
<th>Eden M</th>
<th>Fidell L S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foster D H</td>
<td>Goodman N</td>
<td>Hirsch J</td>
<td>Idris I I</td>
</tr>
<tr>
<td>Kolers P A</td>
<td>Long N</td>
<td>Lovegrove W</td>
<td>Mayhew J E</td>
</tr>
<tr>
<td>McCullough C</td>
<td>Murch G M</td>
<td>Over R</td>
<td>Roalof's C D</td>
</tr>
<tr>
<td>Squires P C</td>
<td>van der Waals H G</td>
<td>VonGrunau M</td>
<td>Wertheimer M</td>
</tr>
</tbody>
</table>

**Representation 2** (Classification codes)

A9774 (Colour perception)

**Representation 3** (Controlled vocabulary terms)

- Colour vision
- Neurophysiology
Representation 4 (Free indexing terms)

Digital visual construction of colour
Disparate shapes
Human visual system
Spatial conditions
Temporal conditions
Visual nervous system

Representation 5 (Vector from abstract and title)

<table>
<thead>
<tr>
<th>Abrupt 2</th>
<th>Analog 1</th>
<th>Appropiat 1</th>
<th>Color 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td>Construct 1</td>
<td>Continuous 2</td>
<td>Contrar 1</td>
</tr>
<tr>
<td>Different 1</td>
<td>Digit 2</td>
<td>Dispar 1</td>
<td>Disparat 1</td>
</tr>
<tr>
<td>Equivalent 1</td>
<td>Flash 1</td>
<td>Found 1</td>
<td>Hand 1</td>
</tr>
<tr>
<td>Human 1</td>
<td>Modern 1</td>
<td>Nervous 1</td>
<td>Propert 1</td>
</tr>
<tr>
<td>Resolv 2</td>
<td>Reveal 1</td>
<td>Shap 3</td>
<td>Smooth 1</td>
</tr>
<tr>
<td>Spati 1</td>
<td>Supplementat 1</td>
<td>Tempor 1</td>
<td>Theor 1</td>
</tr>
<tr>
<td>Transformat 1</td>
<td>Visu 3</td>
<td>Way 1</td>
<td></td>
</tr>
</tbody>
</table>

Representation 6 (Vector from free indexing terms)

<table>
<thead>
<tr>
<th>Colour 1</th>
<th>Cond 2</th>
<th>Construct 1</th>
<th>Digit 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disparat 1</td>
<td>Human 1</td>
<td>Nervous 1</td>
<td>Snap 1</td>
</tr>
<tr>
<td>Spati 1</td>
<td>Tempor 1</td>
<td>Visu 3</td>
<td></td>
</tr>
</tbody>
</table>

Representation 7 (Vector from title)

<table>
<thead>
<tr>
<th>Color 1</th>
<th>Construct 1</th>
<th>Digit 1</th>
<th>Visu 1</th>
</tr>
</thead>
</table>

These seven representations for the sample document illustrate some of the characteristics of each representation and the relationships which hold among them. The most obvious is that they differ in length; this aspect of the representations is discussed further in the data analysis section below. Comparing representations 3 and 4, one finds that free indexing terms are more specific than controlled vocabulary terms and cover more of the points presented in the document abstract. Comparing representation 4 with the text of the title and abstract, one finds that the phrases which make up the free indexing terms are either taken directly from the text (e.g. "disparate shapes") or are simple syntactic variations of phrases found in the text (e.g. "digital visual construction of colour" for "visual construction of color is digital"). This last example also
illustrates a problem which arises when spelling is not standardized.

While the title and abstract retain the spelling of the original paper,
other elements, such as controlled vocabulary terms, use British spelling.
Representations 5-7 are closely related, given the way in which they are
generated. Representation 7 is always a subset of representation 5, since
the title is included along with the abstract in generating representation
5. Since free indexing terms are frequently selected from title and
abstract, the stems in representation 6 are also often a subset of repre-
sentation 5. This is true for the sample document, with the exception
of the substitution of "colour" for "color".

Each of the seven representations was processed for each of the 35
query documents to give a retrieved set of documents. For representations
1-4, documents in the retrieved set each have a weight equal to the number
of attributes (authors, classification codes, controlled vocabulary terms,
or free indexing terms) held in common by the query document and the
retrieved document. For representations 5-7, the weight is the value
of the cosine correlation.

D. Data analysis

1. Representations considered singly

Recalling the discussion presenting the document-feature matrix as
the retrieval system model used in this study, it is of interest to begin
by describing the properties characteristic of each representation:
specificity, exhaustivity, and number of items retrieved in response to
a document submitted as a query. Specificity is calculated by looking
at the columns of the matrix, counting the number of document surrogates
in which a particular term occurs. Exhaustivity is calculated by looking
at the rows of the matrix, counting the number of terms which make up
the surrogate for a particular document. The number of items retrieved is determined by comparing a particular representation for the query document with the corresponding representation for document surrogates in the data base and counting those which satisfy the retrieval rule. A distribution of values of these three characteristics for each of the seven representations in the INSPEC data base has been found using 35 query documents. Since most of the distributions are positively skewed, the range, median, and mean of the observed values for each representation are reported.

**Specificity.** The table below gives values for specificity (number of documents per term) as both a number and as a percentage of documents in the data base of 7145 documents.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Range</th>
<th>Median</th>
<th>(% of data base)</th>
<th>Mean</th>
<th>(% of data base)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-3</td>
<td>0</td>
<td>(0)</td>
<td>.2</td>
<td>(0)</td>
</tr>
<tr>
<td>2</td>
<td>1-102</td>
<td>26.5</td>
<td>(.37)</td>
<td>35.1</td>
<td>(.49)</td>
</tr>
<tr>
<td>3</td>
<td>1-114</td>
<td>15</td>
<td>(.21)</td>
<td>18.0</td>
<td>(.25)</td>
</tr>
<tr>
<td>4</td>
<td>1-88</td>
<td>1</td>
<td>(.01)</td>
<td>.02</td>
<td>(.04)</td>
</tr>
<tr>
<td>5</td>
<td>1-1023</td>
<td>150</td>
<td>(2.1)</td>
<td>253.1</td>
<td>(3.5)</td>
</tr>
<tr>
<td>6</td>
<td>1-1007</td>
<td>120</td>
<td>(1.7)</td>
<td>205.4</td>
<td>(2.9)</td>
</tr>
<tr>
<td>7</td>
<td>1-1007</td>
<td>94</td>
<td>(1.3)</td>
<td>190.4</td>
<td>(2.7)</td>
</tr>
</tbody>
</table>

**Table 1.** Specificity values for 7 representations

Specificity values for each representation were calculated using the set of terms from the 35 query documents rather than the entire data base.

For example, the 35 query documents together had 72 classification codes (out of 1300 used in the INSPEC system) and 116 controlled vocabulary terms (out of 4500 in the INSPEC Thesaurus). It is interesting to note that the observed values for specificity are consistent with one's intuitions as to the properties each representation should have. Comparing representations 2-4, it is evident, on the average, a free indexing term
has been assigned to fewer documents than has a controlled vocabulary
term, which in turn has been assigned to fewer documents than has a classi-
Fication code. Comparing representations 5-7, the "typical" word stem
from the title is more specific (in the sense that it appears in fewer
documents) than are stems from free indexing terms or abstracts. Referring
back to sample document #3405, the stems associated with representations
5-7 illustrate this difference. The abstract includes such stems as
"Found" (in 694 documents), "Propert" (in 532 documents), and "Different"
(in 449 documents) which do not appear in either the title or the free
indexing terms. The stem "Cond" (in 550 documents) appears in both the
abstract and free indexing terms but not in the title.

**Exhaustivity.** The table below gives values for exhaustivity (number of
terms/document).

<table>
<thead>
<tr>
<th>Representation</th>
<th>Range</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-59</td>
<td>10</td>
<td>17.3</td>
</tr>
<tr>
<td>2</td>
<td>1-5</td>
<td>2</td>
<td>2.1</td>
</tr>
<tr>
<td>3</td>
<td>1-8</td>
<td>3</td>
<td>3.4</td>
</tr>
<tr>
<td>4</td>
<td>2-10</td>
<td>6</td>
<td>6.2</td>
</tr>
<tr>
<td>5</td>
<td>13-89</td>
<td>31</td>
<td>34.5</td>
</tr>
<tr>
<td>6</td>
<td>4-26</td>
<td>11</td>
<td>11.5</td>
</tr>
<tr>
<td>7</td>
<td>2-10</td>
<td>6</td>
<td>6.1</td>
</tr>
</tbody>
</table>

**Table 2. Exhaustivity values for 7 representations**

Values for exhaustivity illustrate the variations in representation length
already noted in the case of sample document #3405. For representation 1,
when a document had no references, the number of cited authors was of
course equal to zero. The exhaustivity values for representations 2-4
can be compared to those reported as average for INSPEC records (Institution
of Electrical Engineers, 1977). On the average 2 classification codes
and three controlled vocabulary terms are assigned, which is consistent with the medians of 2 and 3 observed in this study. On the average between 7 and 8 free indexing terms are assigned to a document, so the median of 6 found for the query documents in this study is somewhat low. The exhaustivity values found for representations 5-7 indicate the reduction in word stems searched when the representation is limited to stems from free indexing terms and the further reduction when only the title is considered. It should be remembered that exhaustivity values for representations 5-7 count stems rather than words and do not count stop words.

A typical abstract in the INSPEC data base, for example, has 100 words, but after processing using a stop word list and stemming routine the median found for abstracts in this study was 31 stems.

**Number of documents retrieved.** Specificity and exhaustivity both affect the number of documents retrieved when a document is submitted as a query. Considering representation 3, for example, a typical controlled vocabulary term has been assigned to 15 documents (0.21% of the data base). The query document containing this controlled vocabulary term should therefore retrieve 14 other documents which also have been assigned this term. But, according to the exhaustivity value, a query document typically will have three controlled vocabulary terms assigned to it, and each of these is used as a basis for retrieving additional documents from the data base. Values for the number of documents retrieved (excluding the query document) are reported in the table on the following page. Two sets of results are given for representations 5-7: the first set gives the number of items having nonzero cosine correlation with the query document and the second set gives the number of items having a cosine correlation $\geq 0.25$. 
<table>
<thead>
<tr>
<th>Representation</th>
<th>Range</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-13</td>
<td>1</td>
<td>2.4</td>
</tr>
<tr>
<td>2</td>
<td>3-234</td>
<td>52</td>
<td>68.7</td>
</tr>
<tr>
<td>3</td>
<td>1-137</td>
<td>49</td>
<td>52.3</td>
</tr>
<tr>
<td>4</td>
<td>0-87</td>
<td>6</td>
<td>12.1</td>
</tr>
<tr>
<td>5</td>
<td>1720-6493</td>
<td>4363</td>
<td>4378.5</td>
</tr>
<tr>
<td></td>
<td>0-65</td>
<td>4</td>
<td>10.2</td>
</tr>
<tr>
<td>6</td>
<td>362-4259</td>
<td>1594</td>
<td>1813.0</td>
</tr>
<tr>
<td></td>
<td>0-53</td>
<td>5</td>
<td>10.9</td>
</tr>
<tr>
<td>7</td>
<td>38-2643</td>
<td>798</td>
<td>1009.7</td>
</tr>
<tr>
<td></td>
<td>0-80</td>
<td>4</td>
<td>9.5</td>
</tr>
</tbody>
</table>

Table 3. Number of documents retrieved for 7 representations.

Looking at the two sets of retrieval results for representations 5-7, the need for a cutoff on the cosine correlation defining the set to be retrieved is evident. While large differences are observed in the number of items retrieved using the three different representations with no cutoff on the cosine correlation, quite similar results in terms of the number of items retrieved are observed when the same cutoff point (≥ .25) is used for each representation.

2. Comparisons of representations

The seven representations with their associated retrieval rules are the independent variable and may be thought of as "treatments" applied to each of 35 (randomly selected) query documents. They are different ways of operationalizing the system relevance judgment which acts as a filter, passing only a subset of the items in the database to the inquirer in response to his query. The subset of items retrieved using each representation is thus the dependent variable. The above analysis of each representation separately in terms of specificity, exhaustivity, and number of items retrieved indicates that the representations have different
properties. A comparison of representations involves an assessment of whether they differ in the set of items each retrieves in response to the same query document.

Let \( x_{ij} \) denote the set of items retrieved in response to the \( i \)th query document using the \( j \)th representation, where in this study \( i \) ranges from 1 to 35 and \( j \) ranges from 1 to 7. For a particular query document \( i \), the following Venn diagrams display the possible relationships between the sets of items retrieved in response to the same query document using two different representations \( j_1 \) and \( j_2 \):

In case A the two sets are disjoint, \( j_1 \) and \( j_2 \) retrieved entirely different sets, so \( x_{ij_1} \cap x_{ij_2} = \emptyset \).

In case B the two sets overlap, there are some elements in common, so \( x_{ij_1} \cap x_{ij_2} \neq \emptyset \).

In case C the two sets retrieved are identical, so \( x_{ij_1} \cap x_{ij_2} = x_{ij_1} = x_{ij_2} \).

In case D one is a subset of the other, so \( x_{ij_1} \cap x_{ij_2} = x_{ij_2} \).

In this study representations were tested in pairs, using the following measure to quantify overlap:

Let \( x_{ij_1} \) be the set of documents retrieved in response to representation \( j_1 \). Then \( n(x_{ij_1}) \) is the number of items in this set. Overlap between sets retrieved using \( j_1 \) and \( j_2 \) is measured by:

\[
M_{j_1j_2} = \frac{n(x_{ij_1} \cap x_{ij_2})}{n(x_{ij_1})}
\]
Stage 1. Problem identification. The problem is a logical extension of the capabilities of operational information retrieval systems. Acknowledging the fact that inquirers often do know of documents relevant to their interests (e.g. consider the popularity of Science Citation Index as a searching tool), it is natural to ask how one might build computer-based systems capable of finding documents "like" one already known to the inquirer.

Stage 2. Information gathering. The literature comparing retrieval results obtained when using alternative representations to process search profiles or retrospective searches indicates that, for complete retrieval, a number of representations used in combination are needed. Various studies have compared such things as controlled terms vs. free terms and context vs. content elements. There is no reason a priori to exclude any of these alternatives from study.

Stage 3. Idea generation. To build a model, one needs ideas for both representations and procedures. In this case the alternative representations to be compared are relatively simple; lists of features associated with each document, together with frequency of occurrence data for word stems. The retrieval procedure is also straightforward; one of the alternative representations is selected and documents are judged "like" the query document based on simple overlap (representations 1-4) or cosine correlation (representations 5-7). In this study the source of ideas has been primarily a consideration of the data available and how it can be manipulated by machine, rather than an investigation of how inquirers make judgments on what constitute "like" documents and an attempt to mimic the process by machine. The latter is an alternative source of ideas for future studies.
Stage 4. Model building. The model used in this study is the document-feature matrix. Its implementation on the machine includes creation of the alternative document representations available for searching and programming the retrieval rules associated with each representation. Investigation of the inner workings of the model includes study of each representation separately, determining such parameters as specificity and exhaustivity.

Stage 5. Testing. In this study the only tests completed involved comparison of alternative representations to determine overlap in sets of items retrieved. The logical next step is testing for human acceptance, allowing inquirers to interact with a computer having a number of different representations available. Possible modes of experimentation and testing are outlined below.

1. Representations considered singly

The examination of individual representations in this study has focused on two parameters which are basic to the document-feature matrix model: specificity and exhaustivity. The building blocks of six of the seven representations examined are terms, whether they be words or codes, assigned to texts or occurring in texts. But they do have different properties from the point of view of retrieval, though without further analysis it is not always clear why, as in the case of representations 5–7 of this study. Our understanding of which representations are most appropriate to which types of queries has not really advanced very far beyond the situation described by Bar-Hillel (1957) twenty years ago:

An index set is a tool whereby a document is to be caught whenever it is pertinent to a certain topic and should be judged accordingly. The hook whereby a fish is caught is not, in general, supposed to be a miniature or condensed fish. . . . There is no intrinsic reason why an index should be a miniature,
or condensed, document. That an index . . . is often, but by no means always, a word or phrase occurring in the document or in its title is probably only a transitory phenomenon, the result more of convenience and conservatism than of any inherent reason (p. 106).

In the period since Bar-Hillel wrote the above passage, with the advent of machine-readable data bases and computer-based retrieval systems, there has been even more use of "free text" terms without a detailed study of their structure and function vis-à-vis controlled terms. Now that computer-based retrieval systems make it relatively easy to implement a number of alternative representations for documents, it is particularly appropriate to initiate studies aimed at describing the characteristics of alternative representations and relating these to the possible functions each might serve. Studies which could not have been conducted a few years ago are feasible now using the large dictionary files associated with operational online systems. Such files include searchable features (authors, terms from title and abstract, controlled vocabulary terms, etc.) together with postings data (the number of documents with which each term is associated and their identities). This data of course gives specificity values for features directly and can also be manipulated using logical operators to identify relations among sets of documents associated with particular features. These dictionary files are clearly a resource which can be exploited in the future in large scale studies of the properties of document representations.

2. Representations in combination

Although searchers have for some time made use of representations in combination when developing SDI profiles and retrospective searches, their approach has been primarily a "logical" one, because all terms to be searched must be included explicitly as part of a search strategy.
with terms linked together by logical operators. Doyle (1965) has suggested an alternative approach:

> Information retrieval technology in the 1950's was based largely on principles of logic, an emphasis which was perhaps a "logical" result of the emphasis on use of computers in information retrieval. Computers are (above all) logical. Then a well-known logician [Bar-Hillel, 1957] said that logic was at least being grossly misapplied or at worst nearly useless in the information retrieval field.

Judging by the trend of interest in statistical approaches in general and associative indexing in particular, the 1960's will see information retrieval based more and more on principles of redundancy. This is more appropriate because, as we are often painfully aware, the literature is quite redundant and not very logical (p. 15).

There are a number of approaches to retrieval based on redundancy rather than logic. These approaches recognize that redundancy in information retrieval is not apt to denote systems in which one representation is completely substitutable for another, but rather systems which make use of alternative representations in combination as well as data on term co-occurrence (the statistical approaches mentioned above by Doyle).

These approaches are briefly outlined below, each falling within one of the problem areas for AI in IR.

1. Multistep processing of large files (heuristics in IR)—the choice among alternative representations and their use in sequence or in combination depends on the existence of heuristics to guide the choices. Alternative representations and retrieval rules may be thought of as screens of varying coarseness. Multistep processing of large files uses different representations in distinct steps to process a single search request (Becker & Pyrce, 1977). The steps are arranged so that the first process is most appropriate for a large file, the second step operates on the subfile identified by the first step to further refine the output file,
and so on until the final retrieved set to be shown to the inquirer is identified. For example, classification codes could be used to identify subsets of the file which would then be processed using the text of the abstract. The development of such heuristics for multistep processing would be facilitated by an understanding of the relationships between structure and function for different representations.

(2) User-directed relevance feedback (short term learning)—one situation in which the machine has available the identities of documents relevant to the user's query is in an information retrieval system employing relevance feedback. Once the user has identified relevant documents, there are a number of possible approaches to modifying the retrieved set. In an approach to relevance feedback under investigation by Noreault (1978), the system uses a document identified as relevant by the inquirer as a prototype and searches the data base for similar documents to be included in the retrieved set. While the initial test of this approach is limited to comparison on text terms from title and abstract, an obvious extension is the utilization of multiple representations (e.g., both content and context elements), perhaps leaving selection of which representations to use under the control of the inquirer.

(3) Associative interactive dictionary (external representations)—to take advantage of the redundancy inherent in term co-occurrences without storing co-occurrence data explicitly, the computer must automatically generate lists of terms closely associated with terms in a query at the time the query is processed. These associated terms can then be displayed to the inquirer, who can select additional terms to be incorporated in his query. Doszkocs (1978) has shown the feasibility of identifying associated controlled terms and free text terms at the time
of query processing in operational systems by analyzing the frequency
of occurrence of terms in document representations retrieved in response
to the initial query. A possible extension would be to include such ele-
ments as citations and authors in the associative dictionary which is
generated and displayed for the inquirer.

3. Conclusion

In order to put the study reported in this chapter in perspective,
the distinction between orders of questions made by Kessler (1965) is
helpful:

The question—How do two processes compare with
regard to their ability to form groups of technical
papers?—and the question—How do two processes
compare with regard to solving the problems of
technical communication?—are two different orders
of question from the logical point of view (p. 223).

Only the first type of question has been addressed in this study. It has
looked at the mechanics of group formation rather than at the relevance
of groups formed using different representations. If one finds an
equivalence between two groups, they may be equally good, equally bad,
or equally indifferent as far as a particular inquirer is concerned.
The next phase of study must try to assess the "discriminating power"
of alternative representations (Newman, 1972), the capability of all
features or particular subsets to differentially separate documents
relevant to a particular query from the remainder of documents in the
data base. This is essentially a pattern recognition problem. The suc-
cess of this next phase in identifying representations which are good
discriminators from the point of view of an inquirer will be a major
factor determining human acceptance of computer-based information
retrieval systems.
CHAPTER IX
QUERY FORMULATION AS PROBLEM REDUCTION

In most computer-based systems providing retrospective and/or current awareness searches, the query formulation process begins with the completion of a search request form by the inquirer. The data recorded on this form are often referred to as an interest statement, from which an intermediary can develop an SDI profile (for current awareness searches) or a search strategy formulation (for retrospective searches). The task of transforming the interest statement into computer-processable form may be viewed as a problem to be solved. One approach to problem solving is problem reduction—dividing the initial problem into subproblems more readily solvable than the initial problem. To model SDI profile development as problem reduction, for example, it is necessary to enumerate the subproblems whose solutions can be used to solve the main problem of translating the inquirer's interest statement into a profile. The study reported in this chapter investigates the processes used by intermediaries in solving the problem of profile development. The intent is to determine what parts of the process could be accomplished directly by machine or through a dialogue between inquirer and machine, as well as to identify intermediary expertise which would be difficult for the machine to duplicate.

A. Process models

Two types of process models found in the library literature are relevant to this study of SDI profile development: models of inquirer-
intermediary interaction and models of man-machine interaction. Representative examples of both are described below to provide illustrations of differences in form and content. An awareness of models already proposed in the literature provides a context for discussion of the problem reduction model developed in this study.

1. Inquirer-Intermediary Interaction

The interface process in a computer-based information system is basically analogous to reference service in a library, each constituting a mediation between the information needs of the inquirer and the information resources of the system. The intermediary or reference librarian plays the role of system specialist with a knowledge of how the system works and what it contains, and the responsibility for assisting inquirers in making effective use of available resources. Three models of the interaction which takes place between inquirer and intermediary are discussed briefly below. They illustrate differences in the medium for model construction, including: (1) natural language; (2) flowchart; (3) state transition graphs and matrices.

Taylor's (1968) study of question negotiation may be viewed as a natural language model of a particular component of the reference process of answering questions. The question negotiation process, by which the statement of the inquirer's information need is translated into a form interpretable by the information system, is an important case of the intermediary interacting with the inquirer. Taylor has suggested that this represents a transition from $Q_3$ to $Q_4$ in the stages of question development, where $Q_1$-$Q_4$ can be interpreted as four levels of information need with associated query configurations:

- $Q_1$: actual, but unexpressed, need for information (the visceral need)
$Q_2$—conscious, within-brain description of the need (the conscious need)

$Q_3$—formal statement of the need (the formalized need)

$Q_4$—question as presented to the information system (the compromised need)

The skill of the reference librarian is to work with the inquirer back to the formalized need ($Q_3$), possibly even to the conscious need ($Q_2$), and then to translate these needs into a useful search strategy. Taylor identifies five filters through which a query passes, and from which the librarian selects significant data to aid him in his search: (1) determination of subject; (2) objective and motivation; (3) personal characteristics of the inquirer; (4) relationship of inquiry description to file organization; (5) anticipated or acceptable answers. These five general types of information necessary for search definition are not mutually exclusive categories. While the listing is approximately in order of occurrence, they may occur simultaneously, i.e., relevant data for several filters may be embedded in a single statement by the inquirer. This natural language model has introduced certain constructs (levels of information need and filters through which a query passes) to clarify the process of question negotiation. Taylor suggests that one may work back and forth through the levels of information need and that the filters do not have a fixed order of occurrence, but the natural language model is not precise enough for one to diagram the question negotiation process in detail.

A model proposed by Norman Crum (1969) extends the description of the reference process to include the search strategy and the evaluation of system responses. The intermediary plays a part in interview and negotiation, reformulation of the query, search strategy development,
and evaluation of the initial response of the system. The flowchart shows basically a sequence of events with possible iterations indicated by arrows looping back to previous steps which may be modified based on system response.

A flowchart is seen to offer a convenient means with which to describe simple processes, basically sequential but allowing some backtracking.

In contrast to the above model of the reference process showing a sequence of events, a recent analysis of SDI profile development with the assistance of intermediaries at the University of Georgia and UCLA has suggested that the process is not linear (Carmon, 1975). The three predominant characteristics of the inquirer-intermediary interaction process found in this study were that it is nondeterministic (i.e. not a fixed sequence of events), highly adaptive to the perceived needs of inquirers, and has a large instructional component. Where the process is nondeterministic, as was found in the UGa-UCLA study, state transition matrices or graphs can be used to model the process. In this case categories of events in the interview process were identified and frequency
of transitions between events was tabulated. Picking the five most frequent events, the percentage of transitions between them can be illustrated using the state transition matrix given below. For process data from several cases which do not conform to a single sequence of events, such a model can be used to represent the most frequently observed subsequences (where each entry in the table below equals the percentage of transitions from event A to event B).

<table>
<thead>
<tr>
<th></th>
<th>to 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>from</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
</tr>
</tbody>
</table>

Events
1 = Profile Construction
2 = Clarification of Request Statement
3 = Request Statement Negotiation
4 = Search Strategy Formulation
5 = Tutorial Activity

It is evident from the numbers in the cells of the matrix that activity centers on profile construction, since it is the event which most frequently precedes and follows several of the other events.

The three models of the inquirer-intermediary interaction process suggest that choice of a medium depends on how one can best structure available observations. Natural language is only one language among several available; flowcharts and transition matrices can be used for studying processes in some detail.

2. Man-machine interaction

Searching an online computer-based system is a problem solving situation in which a record of the man-machine dialogue forms a trace of the problem solving process. A number of investigators have already proposed techniques applicable to the analysis of this data. Each focuses on somewhat
different content, answering different questions.

Mittman and Dominick (1973) simply consider parameters to be monitored such as total real time per user access, real time and central processor time per query, number of records searched, number of records retrieved, number of search terms used, type and size of reports generated, number and type of language or syntax errors made by the user. While such data may be helpful to the system designer to suggest needed improvements, it gives relatively little insight into the problem solving process employed by users of the system.

Penniman, using data collected at the same level (trace of user actions and system responses recording type of commands but not content), has constructed a model giving a process description of patterns of searching in online retrieval systems (Penniman, 1975). He proposes analysis of the man-machine dialogue in terms of the following states: begin session, request data base, search index, logic formation, offline print, online display, review search, review commands, set parameters, exit data base, and session. The pattern of movement through states is used as a map of user behavior. In conjunction with each action is the time it was initiated. State/time pairs form the basic data for analysis of user behavior. These are used to build state transition matrices, recording the frequency of state to state transitions. When the data for several searches made by many individuals are aggregated, one has a composite matrix which serves as a stochastic model of state transitions in the searching process. The transition matrix for any session can be compared to the composite matrix for determining similarity using goodness of fit tests. Penniman's ultimate objective is to use composite data as the basis for an adaptive prompting module which could present diag-
nostics and search suggestions based upon comparison of the user's past behavior at the terminal during a particular session with composite data on likely sequences of steps.

Finally, Rosenberg (1973), while not presenting detailed experimental results, proposes monitoring user behavior at the terminal in two ways. The dialogue is monitored by creating a printout of the user's input and the system's output. A time log is kept in the margin of the computer printout to indicate the total elapsed time at each stage. A second monitoring is accomplished by the use of tape recorders. The user articulates his thoughts as he searches the system. Verbally he asks a question before putting it to the computer, describes the search strategy as it is carried out, and evaluates the system output. The result is a tape recording of the user's thoughts as he works with the system. In the margin of the written transcript created from the recording, elapsed time is indicated. The transcript is compared with the printout by matching timing marks, coordinating the user's thoughts with his actions and the computer's responses. Rosenberg suggests that such a large volume of data is not easily reducible for analysis, but it does give some insight into the search process and where the man-machine dialogue does or does not match the human problem solving process.

3. Problem reduction model

The specific problem solving process of interest for this study is that of transforming an inquirer's interest statement into a profile that can be implemented on a computer-based retrieval system. At present in many systems intermediaries begin with the interest statement and interview the inquirer--to clarify and negotiate the request, to select vocabulary, and to formulate a search strategy. Given the interest statement and
information gathered during the course of the interview, the intermediary constructs a profile which can then be processed to retrieve references likely to be relevant to the inquirer's interest area. The presearch interview for an online search is similar to that used to aid in SDI profile development.

Investigators in the UGa-UCLA study cited in the previous section (Carmon, 1975) studied SDI profile development by examining interest statements, transcripts of interviews, and the resulting profiles for several inquirers. The investigators chose to analyze the interview by developing a detailed characterization of the different events occurring over time. As shown above in the state transition matrix found for events in the study, the sequence of events is nondeterministic. The interview interaction is a flexible process which takes its structure from the unique characteristics of each intermediary, inquirer, and search request. It would therefore be difficult to duplicate the pattern of events in the interview with a man-machine dialogue.

There is an alternative approach to building a model of this problem solving process which can serve as a basis for at least partial automation. The emphasis is on examining the interviews to determine what intermediaries do and why, without attempting to duplicate how they do it in terms of an exact replication of the events of the interview. While the UGa-UCLA analysis suggests that the intermediary is an essential part of the profile development process, there is still a need to analyze the process into steps which are routine and steps which require expertise. This would permit one to identify what parts of the process could in fact be delegated to a machine and what tasks the intermediary is uniquely equipped to perform. The study reported here involves reanalysis of the
UGa-UCLA data, developing a different model which aids in the identification of those parts of the process which can be accomplished without the aid of the intermediary, either by delegating them entirely to the machine or accomplishing them by means of a dialogue between inquirer and machine.

As suggested by the process models described in the previous section, there are many approaches possible in defining model form and content. A model is a simplification of reality and the questions to be answered in a particular study determine the form it will take. Since the ultimate goal of the present study is to determine which portions of the profile development process may be mechanizable, the problem reduction formulation of problem solving (dividing the initial problem into subproblems) has been selected as the basis for model development. By using problem reduction, the knowledge and operations required to solve particular subproblems can be isolated and analyzed.

1. Related models

The problem reduction model is related to two recently proposed models (one for reference and one for profile development) which stress decision making. In general, decision making processes are steps in problem solving processes where complexity of the problem solving process depends on the number of and relationship among component decision making processes. The reference process has been characterized in terms of six decision making steps (Jahoda, Braunagel, & Nath, 1977):

1. question negotiation--identification of the inquirer's real information need and elimination of ambiguity

2. message selection--identification of the subject of the information need and the information desired about the subject

3. selection of categories of answer-providing tools (e.g., dictionaries, handbooks)
4. selection of specific answer-providing titles

5. selection of access points within a specific answer-providing title

6. selection of the answer from a specific answer-providing title

The process is basically sequential, although backtracking may be required if the decision made at one step leads to failure at a later step, e.g. the subject headings selected (access points in step #5) do not lead to relevant articles in a periodical index (the answer in step #6), in which case new subject headings (#5) or perhaps a new periodical index (#4) would have to be selected.

A model of the profile development process constructed by Park (1977) also emphasizes the decisions to be made. She assumes that the process of preparing SDI profiles is rational and can be described in terms of four components of a decision theory model: decisions to be made, set of possible alternative outcomes, information (or data) required to assign relative weights to possible outcomes, and contingency relationships among decisions. On this basis she has developed an inquiry process model (IPM) which retains the inquirer as decision maker. The model includes five major functions:

1. CAI module—provides instruction

2. Retrieval specifications development module—includes all the functions necessary to create a profile and its corresponding retrieval specifications for input to online and batch retrieval systems

3. Search module—provides interface to an online search system for creation of a bibliography

4. Bibliography module—formats the bibliography for display at a terminal and subsequent analysis of the profile performance to facilitate profile revision

5. Management module—controls sequencing, error and exception handling, administrative data collection, and interactive session evaluation functions.
The model developed by Park is at points more comprehensive and at points less inclusive than the model developed in this study. While both deal with initial development of a profile, Park also includes consideration of profile revision. Her model is limited, however, to profile development for the ERIC database, while the study reported here includes consideration of profile development for multiple databases and for databases which depend on free text rather than controlled vocabulary searching.

2. Model development

The technique of mastering complexity has been known since ancient times: *Divide et impera* (Divide and rule). (Dijkstra, 1965, p. 214)

The interest statement completed by the inquirer is an ill-structured problem from the machine's viewpoint, since it has open constraints (Reitman, 1965), such as no indication of the databases which are to be searched. Open constraints are aspects of the problem not specified in the problem statement, and in the problem of profile development closing these open constraints is accomplished by the intermediary. This process can be viewed as the solution of subproblems.

A number of sources (in addition to the process models described in the previous section) were used in the development of a tentative list of subproblems: Jahoda and Olson's (1972) comparison of several models of the reference process, Jahoda's (1974) study of reference question analysis and search strategy development, Benson and Maloney's (1975) description of the search process, and Atherton and Jensen's (1976) explication of events in the presearch interview. The original list included nine subproblems:

I. Identify databases to be searched
II. Identify concepts
III. Develop search terms
IV. Identify constraints  
V. Refine terms  
VI. Develop search logic  
VII. Decide which item names to search  
VIII. Determine size of answer required  
IX. Determine level of answer required  

The solution to each of these subproblems is contained in the SDI profile which the intermediary prepares.  

In order to study the process by which the intermediary goes from the interest statement to the profile with all the subproblems solved, it is necessary to examine the portions of the interest statement and interview transcript relevant to subproblem solution. The data available for analysis are sets of forms (interest statement, interview transcript, and the resulting profile) for 20 cases from UGA and 46 from UCLA. Approximately 2/3 deal with topics in the natural sciences and 1/3 in the social sciences and they cover a variety of subjects and data bases, using both free text and controlled vocabularies. Development of the problem reduction model depends on both content analysis and protocol analysis and is described in detail in the following two sections on data collection and data analysis. Content analysis is used to determine which portions of the interview transcripts are related to particular subproblems. Once the text related to a particular subproblem has been identified, protocol analysis is used to infer the processes used by the intermediary in developing a solution to the subproblem.  

C. Data collection  
1. Content analysis  

In a study involving content analysis, the interest lies not in the content analysis itself, but in the relationships uncovered by the larger study in which content analysis is embedded. In the present study
content analysis is the first step in analysis of interview transcripts. Holsti (1968) has defined the characteristics of content analysis:

1. Content analysis is objective because it is carried out on the basis of explicitly formulated rules which enable two or more persons to obtain the same results from the same documents, since categories of analysis used to classify content are clearly and explicitly defined.

2. Content analysis is systematic because inclusion or exclusion of content is done according to consistently applied criteria of selection to methodically classify all relevant material in the sample. Sources for the establishment of content analysis categories are the investigator's research hypotheses and the material itself. To increase the reliability of classification of portions of the text into categories, it is necessary to specify clearly the characteristics of statements to be placed in a given category and to use examples drawn from the material being analyzed to illustrate what kinds of statements are to be considered as belonging in a given category (Selltiz, Wrightsman, & Cook, 1976, p. 396). There are three primary requirements for categories (Budd, Thorp, & Donohew, 1967, p. 39): (1) adapted to the needs of the study; (2) exhaustive (relative to the problem); and (3) mutually exclusive.

An investigation related to the present study used content analysis for categorization of questions in reference interviews (Lynch, 1977). Lynch had audio recordings of 309 reference interviews which were divided into four types of transactions--directional, holdings, substantive, and moving. She developed separate sets of content analysis categories to be used in classifying the types of questions asked by reference librarians in holdings transactions (requests for information about whether specific documents are owned by the library) and substantive transactions (requests
for factual information, help in interpreting something in an information source). She identified ten categories of questions for holdings transactions and twenty categories of questions for substantive transactions, and developed detailed category schemes including definitions and examples for questions in both types of transactions (Lynch, 1979, pp. 138-141). The instructions for content analysis developed in the present study (see Appendix A) are similar in form (though not in content) to those developed by Lynch.

2. Content analysis category development

Content analysis of interview transcripts to gather data for study of the profile development process depends on adequate definition of the categories to which portions of text are to be assigned. Since this study models the profile development process as one of problem reduction, the categories for analysis are the subproblems which make up this task. The original list of nine subproblems given in the previous section was derived from previous studies of the reference interview and the profile development process. This list was regarded as tentative, subject to revision based on a review of some of the interview transcripts.

The cases available for this study included 20 from UGA and 46 from UCLA. Half of these (sample A) were used to develop criteria for assigning statements to each category and the remaining half (sample B) were then coded to yield the data necessary to study the profile development process using protocol analysis. Sample A was drawn by choosing (without replacement) a random sample of ten of the UGA cases and 23 of the UCLA cases, leaving ten UGA cases and 23 UCLA cases as sample B.

In content analysis category definition there is a need "to distinguish between programmable units, whose limits in the present state of
the art can be specified unambiguously and hence can be identified by a digital computer and semantic units which . . . cannot be rigorously defined" (Markoff, Shapiro, & Weitman, 1975, p. 11), and thus require the use of human coders. If it were possible to list all the variations of content which indicate a given category, such a list would provide a complete operational definition of the category. In actual practice most categories with which social scientists deal cannot be defined by an exhaustive listing of indicators. Instead of attempting to construct a complete list, which could be used by a computer, it is necessary to rely upon the ability of trained human coders to respond to content indicators in a systematic way (Cartwright, 1953).

In this study a category is defined using a brief descriptive statement and actual examples from the text of interview transcripts. Until an effort is made to apply them, there is no guarantee that a theoretically derived set of categories, such as the original list of nine sub-problems, is in fact applicable to a particular set of documents. In order to develop an adequate set of categories, the transcript from each of the 33 cases in sample A was read by this author with three purposes in mind:

(1) to determine if there were categories lacking from the original list which should be added;

(2) to identify categories on the original list which failed to appear in the transcripts and should be deleted;

(3) to select passages which could be used as part of category definition.

As a result of this analysis, eight categories were defined and instructions for human coders were developed. Each category corresponds
to one subproblem in the profile development process.

3. Content analysis category revision

The lists of content analysis categories given below summarize the evolution of the content analysis category scheme used in this study.

The set of categories is shown at three stages:

(1) original set based on a review of studies by other investigators which analyzed steps in the reference process and the profile development process;

(2) first revision based on the author's analysis of cases in sample A;

(3) second revision based on the outcome of coding reliability tests (described below).

<table>
<thead>
<tr>
<th>Original Set</th>
<th>First Revision</th>
<th>Second Revision</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Identify data bases to be searched</td>
<td>I. Identify data bases to be searched</td>
<td>I. Identify data bases to be searched</td>
</tr>
<tr>
<td>II. Identify concepts</td>
<td>II. Identify concepts</td>
<td>II. Identify concepts</td>
</tr>
<tr>
<td>III. Develop search terms</td>
<td>III. Develop content search terms</td>
<td>III. Develop content search terms</td>
</tr>
<tr>
<td>IV. Identify constraints</td>
<td>IV. Develop context search terms</td>
<td>IV. Develop context search terms</td>
</tr>
<tr>
<td>V. Refine terms</td>
<td>V. Refine terms</td>
<td>V. Refine terms</td>
</tr>
<tr>
<td>VI. Develop search logic</td>
<td>VI. Develop search logic</td>
<td>VI. Develop search logic</td>
</tr>
<tr>
<td>VII. Decide which item names to search</td>
<td>VII. Decide which item names to search</td>
<td></td>
</tr>
<tr>
<td>VIII. Determine size of answer required</td>
<td>VIII. Determine size of answer required</td>
<td></td>
</tr>
<tr>
<td>IX. Determine level of answer required</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The original set of categories was based primarily on studies concerned with modeling the reference process, so the possibility of a need for revision, deletion, and/or addition of categories was anticipated. The intent was to develop categories capable of describing the subproblems
in the profile development process for systems which have multiple data
bases available and which use Boolean logic as the basis for combining
elements in the profile and for determining which records to retrieve.
The final set of categories represents a reduction in number and some
changes in label from those found in the original set. The rationale
for the revisions can be given as follows:

**First revision.**

Category IX was deleted because it was not applicable to the user
group at either UGa or UCLA. "Level of answer" relates to the level of
expertise required to utilize material retrieved, e.g., material suitable
for a non-scientist is different from that required by an expert. The
users at both centers were graduate students, faculty, and researchers,
already quite knowledgeable about the subject areas which their search
requests addressed. "Level of answer required" was therefore not a fac-
tor in profile development in the cases available for study.

Labels for categories III and IV were revised to better indicate
the scope of each category. "Develop search terms" became "Develop
content search terms" to point out that this category is limited to
discussion related to selection of subject descriptors in whatever form--
controlled vocabulary terms, free text, subject codes. "Identify con-
straints" became "Develop context search terms". Context elements, in
contrast to the content terms which describe subject, include such
attributes of a document as its author, the journal in which it is
published, and the language in which it is written. Such items may be
used as "constraints" on the subject portion of the profile, i.e. select
a document on this subject only if it has been written by a certain
author, published in a certain journal, etc. In other cases, however,
an inquirer may be interested in retrieving all articles written by a particular author. Also in citation index data bases, the citation is the principal mode of searching. Category IV was therefore renamed to suggest that search on context elements can supplement a search on content terms.

Second revision.

Category VII was deleted because decisions on which item names (e.g. author, controlled vocabulary terms) to search are made as part of subproblem solution in categories I, III and IV. Category VII was therefore redundant and not needed as a separate category. Selection of data bases (Category I) dictates what item names are available for searching. If authors or journal CODEN are selected for inclusion in the profile (Category IV), then the appropriate item names (author, CODEN) would be selected for searching. A similar situation exists for controlled vocabulary and free text terms (Category III). The results of the coding reliability test showed that Category VII was used very infrequently; where it was used, the discussion was actually tutorial, explaining to the user which item names could be searched on particular data bases (e.g. which data bases had abstracts). Deletion of Category VII thus represents elimination of redundancy and of a source of confusion to coders.

Category VIII was deleted after an analysis of the results of the coding reliability tests, where it was observed that this category was assigned inconsistently. Discussion of "size of answer required" can take two forms:

(1) use of broader or narrower subject terms (which is part of Category III);
(2) selection of more or fewer concepts to tighten or relax conditions for retrieval (which is part of Category II).

Since specification of search requirements (broad or narrow) takes place in previously defined categories, there is no need to retain Category VIII as a separate category.

The instructions for content analysis including definitions of the resulting six categories, together with examples of tutorial material and material to be excluded from analysis, are given in Appendix A. This was used by the author in coding the cases in sample 3. The results of this content analysis served as the raw material for the protocol analysis described in the data analysis section of this chapter.

4. Reliability analysis

While this author used the instructions for content analysis given in Appendix A to code all cases in sample 3 prior to protocol analysis, results of an investigation of intercoder reliability are of interest as an indication of whether other coders using the same instructions would replicate category assignments made by the author. Since the units of interest in the text for this study are semantic rather than programable, some variation is to be expected when different coders are used for the content analysis. Carmon (1975) has noted this difficulty with respect to event analysis of the interview transcripts using a single coder:

Translation of the actual events occurring in the interview to the stylized expected events required considerable subjective judgment on the part of the analyst, frequently requiring inferences as to the rationales behind the users' and intermediaries' questions. While identical results would probably not have been obtained from two (or more) analysts, the results of this task can be used as a reasonable approximation to the events which are occurring. (p. 21)
In order to assess the reliability of coding in this study, three transcripts which had been coded by the author were given to four other people for independent coding. The coders selected were familiar with IR systems and the task of profile development but had never worked as intermediaries. The transcripts chosen (one from UGA and two from UCLA) were selected because their length was closest to the mean for all cases. These transcripts were sufficiently long to contain examples of all of the content categories, but not so long as to require excessive time for coding.

The reliability analysis had two phases. Coders A and B used the list of eight categories resulting from the first revision and coders C and D used the list of six categories resulting from the second revision (as discussed above). Since a category number may apply to a single statement, a paragraph, or even a long conversation between intermediary and inquirer, the number of category assignments made by each coder may be different. Reliability was therefore measured by counting the number of lines in a transcript labeled the same by a coder and the author, and determining the proportion where there was agreement. This proportion was calculated for each of the four coders on the three transcripts as shown in Table 5 below.

<table>
<thead>
<tr>
<th>Case</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.69</td>
<td>.75</td>
<td>.61</td>
<td>.92</td>
</tr>
<tr>
<td>2</td>
<td>.50</td>
<td>.46</td>
<td>.77</td>
<td>.72</td>
</tr>
<tr>
<td>3</td>
<td>.71</td>
<td>.79</td>
<td>.65</td>
<td>.89</td>
</tr>
</tbody>
</table>

Table 5. Proportions for reliability analysis

The category assignments of coders A and B were examined to determine whether some change in category description could have prevented disa-
greement, resulting in the revised set of categories used by coders C and D. While the assignments made by coder D showed better agreement with the author than those made by coders A and B, those made by coder C did not. The two main sources of this disagreement suggest the need for possible additions to the instructions given in Appendix A:

(1) Note under Subproblem V: Subproblem V refers only to term refinement (spelling, truncation, weighting). Discussion of term selection should be classified under Subproblem III (for content terms) or Subproblem IV (for context terms). (This revision is needed because coder C classified discussions of term selection using a thesaurus as V rather than III.)

(2) Note under Tutorial: Tutorial refers only to instruction of the user by the search analyst. Instruction of the search analyst by the user should be classified under the subproblem to which it is related, such as Subproblem III for an explanation of the meaning of terms. (This revision is needed because coder C classified instruction of the search analyst by the user as tutorial rather than under one of the subproblems.)

The reliability analysis also suggests that certain cases, such as case 2 from among the three used, may be more difficult for coders to interpret and hence lead to more unreliable coding. An examination of case 2 shows that much of the discussion between inquirer and intermediary takes place while they are examining issues of Chemical Abstracts and Biological Abstracts. From the verbal transcript it is often difficult to infer exactly what is the focus of discussion, resulting in wide variation in the way several passages were categorized by different coders. This illustrates one limitation of interview transcripts, another being that they contain no record of nonverbal communication such as pointing, shaking one's head, etc. Where verbal transcripts are insufficient to allow reliable categorization, they are also insufficient for inferring problem solving processes in protocol analysis. Situations where this occurs are noted in the data analysis section below.
The results of the reliability analysis should be interpreted while taking into account the following observation by Holsti (1968):

An acceptable level of reliability is one of many issues for which there is no ready definition. The question can only be answered in the context of a given research problem. That high reliability can be achieved for simple forms of content analysis, in which coding is essentially a mechanical task, is amply documented in the literature. Conversely, as categories and units of analysis become more complex, they are likely to become both more useful and less reliable. In formulating and testing a content-analysis research design, the analyst may thus be forced to strike some appropriate balance between reliability and problem significance. (p. 660)

While additional reliability tests may yield other ideas for improving the instructions for content analysis, the level of agreement was judged to be satisfactory for the purposes of this study. It was evident from the reliability analysis that familiarity with the task of profile development is a necessary qualification of the coder, since assignment to categories involves more than simple identification of keywords. Assignment depends on an understanding of the context in which statements are made, as the following example illustrates:

Just for curiosity, let's see if there is anything on educational planning in Psych Abstracts...they have educational background, guidance, measurement, and psychology, so that's not exactly what you want is it? O.K., the word planning does not even appear, so that would leave it out, if the keywords are not there, there's no use to go on with the others.

While superficially this passage looks like a discussion of subject term selection (Category III), it actually represents a decision on whether to search a particular data base (Category I). Use of the content analysis categories developed, therefore, depends on the availability of coders knowledgeable about the profile development task and coding cannot be done automatically.
D. Data analysis

1. Protocol analysis

In a paper urging librarians to view reference as a special case of problem solving, Gardiner (1969) observes that what distinguishes the reference librarian from other people is the way he has learned to encode the task domain over which he works. One approach to studying this encoding makes use of the thinking aloud method in order to get protocols from which one can infer the kinds of cognitive strategies that operate in complex tasks. In addition to those situations where a person is asked to "think aloud" to generate a protocol, protocol analysis can be used where external responses are an intrinsic part of the problem solving process. In the profile development process, the problem solving processes of the intermediary are verbalized to some extent when he develops a profile in the presence of the inquirer for whom the output of the profile search is intended. The intermediary tries to verbalize the rationale for various decisions so that the inquirer can have a better understanding of the process used to develop a profile from the original interest statement.

When applied in cognitive simulation studies, protocol analysis is frequently used to create a problem behavior graph. The following guidelines can be given for developing a problem behavior graph from a subject's problem solving protocol (Saylor, 1973):

(1) Protocol is divided up into segments based on pauses in the flow of speech or on what appears to constitute a single task assertion.

(2) Problem solving is assumed to take place in a problem space or set of problem spaces. Elements of the problem space are used to define states of knowledge through which the subject is assumed to pass in the
course of solving problems. To effect passage from one state to another, the subject must apply an operator. The subject sets goals to move toward new states of knowledge.

(3) The problem solving process is sequentially described in terms of a problem behavior graph. Nodes correspond to the contents of the subject's states of knowledge and arrows to the right point at successor states and are labeled by the operators applied by the subject. If the problem solver returns to a given state to try a new operator, the state is redrawn below its old appearance. Numerous examples of protocols and the problem behavior graphs resulting from their analysis can be found in Newell and Simon (1972).

While protocol analysis in the present study has not been used to produce a problem behavior graph, the constructs of knowledge state, goal, operator, and backtracking are useful in describing the problem solving process of the intermediary. In the profile development process the initial knowledge state for each subproblem is made up of relevant data from the interest statement plus the solutions of any other subproblems which have already been determined. The goal state is the solution of the subproblem as it is embodied in the profile. For example, in working on Subproblem IV (Develop context search terms), the intermediary has initially the lists of authors, journals, and languages of publication which are of interest to the inquirer. In addition, the intermediary has the solution to Subproblem I, i.e. which data bases are to be searched. To reach the goal, the decision as to which context elements are to appear in the profile, the intermediary may perform a number of operations, such as asking the inquirer which authors are of special interest and checking a reference manual to determine if the language field is searchable in
a particular data base. Backtracking may also occur. In solving Sub-
problem III, for example, the effort to translate free text terms into
controlled vocabulary terms may be abandoned when it is found that the
thesaurus for a particular data base contains no appropriate terms. Thus
while the product of protocol analysis in the present study is not a
detailed problem behavior graph, the purpose is still to try to infer an
intermediary's knowledge states and associated operators in order to
develop an understanding of the problem solving process in profile devel-
opment.

Development of the approach to be followed in this study has drawn
on previous studies by Montgomery and Swanson (1962) on machinelike in-
dexing by people and Jahoda (1974) on reference question analysis and
search strategy development by man and machine. Montgomery and Swanson
observe that indexing as an intellectual process is profound in principle
and difficult to formalize, yet indexing in practice may be of a quite
different nature. The present study was undertaken with the expectation
that the same may be true of the profile development process. Montgomery
and Swanson concluded that most of the articles in their sample from
Index Medicus could have been indexed by machine on the basis of machine
inspection of article titles alone, with word and synonym match of title
words to subject headings. Two other aspects of the Montgomery and
Swanson study are paralleled in the present study:

(1) They were not concerned with how medical literature could have
or should have been indexed, but only with how it was indexed. The
approach thus does not involve evaluation in the sense of measuring
retrieval effectiveness, but rather is based on the question: To what
extent can the human indexing operations that take place in an existing
system be simulated by machine? Similarly in the proposed study, the intent is not to determine what intermediaries ought to be doing to develop profiles adequate for retrieval purposes, but rather what they are doing. The criterion for proposed machine processes is to develop the same profile as that developed by the human intermediary, beginning with the completed interest statement.

(2) While Montgomery and Swanson found that mechanizable processes could be postulated which would lead to 86% of the subject headings assigned by humans, the remaining 14% would not have been generated by the proposed machine procedures. It is therefore of interest to do a "failure analysis" of the 14% to determine why they are not readily mechanizable. In the present study failure analysis is used to identify intermediary expertise (such as interpersonal skills, system knowledge, data base knowledge) for which there is no mechanical substitute.

In Jahoda's study each of 28 reference questions was first manually (i.e. intellectually or non-mechanically) analyzed in terms of a sequence of identified steps in the process of answering reference questions. Results of the manual analysis of each step were examined to determine how each step was performed and whether mechanical rules for performing the step could be developed. In the analysis for steps after the first one (selection of indexable information), correct (manually selected) results were assumed for the previous step. Similarly in the analysis of profile development performed in this study, each subproblem was examined separately. If solution of one subproblem depended on solution of another, correct results were assumed for the first in proposing machine procedures for the second (e.g. choice of data base determines whether terms to be selected are controlled or free indexing terms).
Results of the analysis using the 33 cases in sample B (including the interest statement, interview transcript, and profile) are summarized below by subproblem. A copy of the form for completion of an interest statement (called the Information Form for User in the UGa-UCLA study) has been included as Appendix B. Two sample cases, one each from UGa and UCLA, have also been included in full: Appendix C lists sample data from the completed interest statements; Appendix D contains interview transcripts; and Appendix E has the final profiles developed by the intermediaries on the basis of the interest statements and interviews. Examples given in the discussion of subproblems below denote data from Appendix C by IFU, data from Appendix D by TRA, and data from Appendix E by PRO. In analyzing each case, each subproblem was considered separately, with the final profile representing the combination of solutions to the identified subproblems. The portions of the interest statement (IFU) and interview transcript relevant to the solution of each subproblem were used to infer the processes used by the intermediary. The entire profile development process cannot be automated if one or more subproblems cannot be solved automatically, but the analysis by subproblem given below yields:

(1) identification of subproblems which can be solved without the aid of the intermediary either by delegating them completely to the machine (automatic solutions) or accomplishing them by means of a dialogue between inquirer and machine (possible interactive solutions);

(2) identification of the problem solving process used by the intermediary (intermediary solutions) and enumeration of the specific skills of the intermediary which cannot presently be duplicated by machine (intermediary expertise).

Before discussing in detail the findings for each subproblem, it is appropriate to present some general remarks relating the problem solving
task under study (profile development for batch processing systems) to the task of developing strategies for retrospective searches of online systems. The criterion for evaluating the mechanizability of the problem solving process is production of a profile identical to that produced by the human intermediary. This is a rigid criterion for two reasons:

1. Profiles of all intermediaries are considered to be solutions which machines should try to duplicate. Since the profiles varied considerably (in terms of such factors as number of concepts, number of terms to represent each concept, utilization of particular system features), any proposed automatic solution is likely to match only some percentage of intermediary solutions due to the variation among intermediaries in the form their solutions take. These differences are illustrated in the discussion of individual subproblems below.

2. In general, SDI profile development for a batch processing system is a more complex problem solving task than developing a search strategy for an online retrospective search. SDI profiles tend to cover many subtopics within an individual's area of interest, while a typical online retrospective search deals with a narrow topic. Where search topics are focused, covering fewer subtopics and requiring the selection of fewer terms, it is likely that the percentage of failures for both automatic and interactive (inquirer-machine) solutions would be smaller than that found for the process of SDI profile development studied here. Complexity of SDI profiles is also due to the efforts of intermediaries to cover all possible approaches to a subtopic, through the inclusion of many terms and possibly complex search logic. The purpose of a retrospective search, on the other hand, may vary from finding a few relevant references to performing an exhaustive search. Where the goal is to be selective, automatic and interactive solutions are more feasible than in the case of SDI profiles.
Problem to be solved.

Solution in the profile.

Relevant data from the Information Form for User.

Relevant data from the transcript.

Example.

Overview.

Automatic solutions.

Intermediary solutions.

Possible interactive solutions.

Intermediary expertise.

2. Subproblem I

Problem to be solved. Identify data bases to be searched.

Solution in the profile. Data bases searched.

Relevant data from the Information Form for User.

19 - publications consulted in manual literature search

22 - subject areas to which the question is directly related

Relevant data from the transcript.

Discussion relates to selection of the data bases to be searched from among those available on magnetic tape at the information center.

Example (Case 1).

IFU: 19 - Biological Abstracts, Index Medicus
     22 - biology, medicine

TRA: A: I discussed and I think you were right that the BA Biological Abstracts would probably.
     U: Fit my needs better.
     A: Yeah.

PRO: Searches Biological Abstracts.

Overview.

Since the search centers at UGA and UCLA had different data bases
available for searching, this must be taken into account in analyzing the solutions to this subproblem. UCLA had five data bases available and UGa had sixteen data bases available, of which five were limited to retrospective searching. All but one of the data bases available at UCLA (Social Sciences Citation Index) were also available at UGa. Of the five data bases at UCLA, three covered the natural sciences and two covered the social sciences. Of the sixteen data bases available at UGa, twelve covered the natural sciences, three covered the social sciences, and one (Government Reports Announcements) was general in coverage.

Automatic solutions.

In solving this subproblem using only the relevant data from the IFU, the simplest automatic rule is to search those data bases checked in item 19 of the IFU (publications consulted in a manual literature search) which are also available in machine-readable form at the search center (UCLA or UGa). This would be successful in 58% of the cases in the sample. An extension of this rule involves simple inferences using data from item 22 of the IFU (subject areas to which the question is directly related) to add to or modify the list created by considering only item 19 (e.g. search Social Sciences Citation Index if "social science" is checked in item 22, search only the biochemical section of Chemical Abstracts if "biology" or "medicine" is checked in item 22). This would be successful in an additional 27% of the cases. The remaining 15% cannot be solved by simple automatic means.

Intermediary solutions.

Selection of data bases was mentioned in all the transcripts, although in some cases the intermediary simply repeated what was checked on the IFU and agreed that those data bases would be searched. Only
occasionally did the intermediary make a decision to search a data base other than those checked without first consulting the inquirer. These consultations took two forms:

(1) where the inquirer was familiar with the data base, the intermediary asked whether its printed counterpart frequently included articles on this topic;

(2) where the inquirer was not familiar with the data base, the intermediary first explained subject or type of material (e.g. government reports) coverage, then asked whether the inquirer would like to search the data base.

When the intermediary selected a data base without consulting the inquirer, subject coverage was given as the basis for selection most frequently, although time lag in indexing material and type of material covered were also mentioned. There was insufficient data in two cases to determine the basis for subproblem solution by the intermediary.

Possible interactive solutions.

A dialogue could begin by asking questions 19 and 22 from the IFU. With this data the machine could suggest data bases from those available and allow the inquirer to confirm their selection. Where the inquirer is unfamiliar with a data base, he could query the machine for a brief description of the subject coverage, types of material included, and time lag characteristic of a particular data base prior to making a decision as to whether it should be searched.

Intermediary expertise.

There are two types of intermediary expertise which would be more difficult to duplicate either automatically or interactively:

(1) Detailed knowledge of data base subject coverage. Occasionally
the intermediary used knowledge of subject coverage of data bases not immediately apparent from their titles (e.g. ERIC covers linguistics). This knowledge is especially important when the inquirer is unfamiliar with the printed counterpart.

(2) Ranking data bases. In an online system, the ability to rank data bases, determining sequence of search, is particularly important. While automatic selection of all data bases to be searched can be accomplished automatically in many cases, the decision as to which should be searched first is more difficult. The intermediaries frequently told inquirers which data bases of those to be searched were likely to contain the most citations.

While the present study provides no indication of how these types of expertise might be duplicated automatically, the results of a study by Williams and Proce (1977) which is currently in progress should provide an indication of whether automatic data base selection is feasible.

Their data base selector operates on query terms to provide a relative ranking of data bases according to their applicability to the query, and it is currently being evaluated.

3. Subproblem II.

Problem to be solved. Identify concepts.

Solution in the profile. Groups (UGa) or term sets (UCLA).

Relevant data from the Information Form for User.

11 - profile title

12 - prose statement of the search request

Relevant data from the transcript.

Discussion relates to identification of the concepts which will form the basis for profile development. The concepts are often the major nouns or groups of nouns in the inquirer's description of his area of interest.
Example (Case 2).

IFU: 12 - What are the effects of dietary lipids on channel catfish (Icturus punctatus)

TRA: A: O.K., what we are going to do here is to essentially as you have indicated set up these two groups of terms--your lipid and fatty acids terms we will designate as one set of terms and then the fish terms as the second set.

PRO: Two groups of terms, GO01 = fish terms and GO02 = lipid terms.

Overview.

The solution of this subproblem is particularly important in retrieval systems which employ logical operators since concepts form the basis for profile development. In general the greater the number of concepts linked by AND, the narrower the search, and the greater the number of concepts linked by OR, the broader the search. In the sample profiles in Appendix E, Case 1 has four concepts which are searched on title, terms, and abstract (CHEMTERMS, MICROTERMS, PITTERMS, and GLANDTERMS), while Case 2 has only two concepts (GO01 and GO02 representing fish terms and lipid terms). These profiles are typical of other cases in that those from UCLA tend to contain several concepts while those from UCSC tend to contain two or three. While instructions for profile preparation often state that concepts are "the major nouns or groups of nouns in the inquirer's description of his area of interest", identification of concepts in text cannot be done by any simple parsing routine. To give two examples: (1) for the search request "peptide and amino acid transport and utilization" the two concepts are "peptide and amino acid" and "transport and utilization"; (2) in "rehabilitation counseling techniques used to rehabilitate juvenile offenders" the three concepts are: "rehabilitation counseling", "techniques", and "juvenile offenders".

Automatic solutions.

In almost all cases it would be difficult to automatically process
the text of items #11 and #12 of the IFU to identify the same concepts as those identified by the intermediary. Picking out nouns or noun phrases is relatively straightforward, but it is the fact that concepts are often groups of nouns that makes it difficult to perform the task automatically. There are two types of failure for automatic techniques: (1) all the concepts are in the IFU but they cannot be identified automatically (e.g. "measuring change in self esteem" became a single concept search dealing with "self esteem") or (2) some of the concepts used in the final profile are added as the intermediary learns more about the inquirer's area of interest during the interview (e.g. a request for "techniques and experimentation in phonetics" was expanded to include a set of organ terms). In 27% of the interviews concepts not in the original IFU were added during the course of the interview or by the intermediary in preparing the profile. It should be noted that all of these were among the cases from UCLA which tended to include a large number of concepts in profiles. Only two searches from among those examined were single concept searches. The remaining were multiconcept searches where some parsing of the text in items #11 and #12 of the IFU was required to identify the concepts.

Intermediary solutions.

While intermediaries identified concepts before working with the inquirer to create the list of terms which was to be associated with that concept, they infrequently discussed their decisions on concepts with the inquirer. Some of them would explain their choices to the inquirer as part of a tutorial on the profile development process, but they were able to make decisions on what constituted concepts without the aid of the inquirer. Since the basis for selection was seldom verbalized, one must infer it from a comparison of the IFU and the final profile. In addition to the use of knowledge of English and of subject areas to identify concepts,
intermediaries applied knowledge of data base characteristics in two ways. In data bases with a controlled vocabulary such as ERIC, many terms are precoordinated and can be used as labels for single concepts, whereas to search a free text data base there would have to be two concepts linked by AND (e.g. if "rehabilitation counseling" is a single term then one need not search on "rehabilitation" AND "counseling"). The other type of data base knowledge is related to what one might think of as "implicit search terms", which if made explicit in the profile would be redundant. For example, for a search related to the geology of a region in a geological data base, it is sufficient to search the name of the region. If an engineering data base were searched, on the other hand, it would be necessary to add another concept for "geology".

Possible interactive solutions.

Proposals given here for interactive solutions are more tentative than those given for other subproblems, because they presuppose that it is possible to convey to the inquirer a sufficient understanding of the structure of profiles using logical operators that he can identify the concepts which make up his area of interest. While 32% of the cases examined involve profile preparation for first time system users, five had previously had from 1-5 searches done for them and one inquirer had had more than 10. This last individual had labeled groups of terms in item #13 of the IFU as G1, G2, G3 corresponding to the concepts in his interest statement, suggesting that he viewed the profile as an intermediary does. Whether the introduction to the process of translating an interest statement into a set of concepts is best accomplished by means of personal instruction by intermediaries, manuals, CAI, or some other method should be the subject of further study. Moghdam (1975) describes the differing features of printed manuals, live help, AV presentations, and CAI as alternative training media. Since searching
of operational online systems also requires initial identification of concepts before a search strategy is developed, the issue of training the inquirer is important for the use of these systems as well. Assuming that the inquirer is capable of translating his interest statement into separate concepts, dialogue with the machine may still be useful in those situations where the number of concepts to be searched is data base dependent, such as cases in which there is a controlled vocabulary or for databases which are limited to a particular subject area as illustrated by the examples given above. Such a division of labor between inquirer and machine is likely to generate a workable set of concepts more quickly than would be possible if the machine had to begin with a natural language interest statement and "guess" concepts which could then be accepted or rejected by the inquirer as the basis for the profile.

Intermediary expertise.

Solution of this subproblem depends on knowledge of the English language (to parse interest statements into candidate concepts), knowledge of subject areas (to determine what are distinct concepts for searches in a particular subject area), knowledge of data bases (to modify concept sets by determining which are appropriate for searching particular data bases), and knowledge of the search system (to understand the principles of concept identification and coordination using logical operators).

The first two types of knowledge are also possessed by the inquirer. It is desirable to investigate ways in which he can also acquire knowledge of the search system, for without this he must be dependent on an intermediary to aid him in profile development. How best to incorporate knowledge of data bases into the machine so that it can assist in the solution of the concept identification subproblem remains to be inves-
tigated. The cases examined here suggested certain ways in which the
decision to search a particular data base affects the choice of concepts;
there may be others.

Those cases in which additional concepts were identified during
the course of the interview illustrate another type of intermediary
expertise: interpersonal communication skills. By asking the inquirer
to describe his area of interest in greater detail, the intermediary was
able to identify another facet of the search not expressed in the original
interest statement, but which was included in the final profile. For
example, a search on the "anatomic and physiologic basis for acupuncture"
included terms on the "martial arts", following this discussion between
the intermediary and the inquirer:

A: O.K. You mentioned pressure points. Is this
something you want to pursue at all in relation
to this? Is there any reasonable way to pursue it?

U: I haven't thought of any yet. I have a little
book at home that has the title, "First Aids At
Your Fingertips" and it is the use of digital
pressure on points that have either been used
for needling or in the martial arts.

The intermediary's question and inquirer's response remind one to some
extent of a conversation with ELIZA (Weizenbaum, 1966), a major difference
of course being that the program has no real understanding of what a
user says. The following example illustrative of a conversation with
ELIZA is taken from Raphael (1976), where lines in capital letters are
generated by the computer:

DO COMPUTERS WORRY YOU?
Not at all. I think they can be very useful.
CAN YOU ELABORATE ON THAT?
Computers play a vital role in the operation of our society.
WHY DO YOU MENTION COMPUTERS?
Because that is what the book is all about.
IS THAT THE REAL REASON?
Of course. (p. 199)
While an intermediary can make effective use of such dialogues to learn more about the inquirer's area of interest, their use in interactive solution of IR problems cannot be fruitful until machines can understand what the inquirer is saying in the sense that they can use the new information gathered to modify the profile or search strategy development process. For the short term, rather than try to mimic the intermediary's role entirely, the machine can be employed as an assistant, providing the inquirer with the information about data bases needed to make decisions on concepts to be used in searching. It should also be noted that for retrospective searching rather than SDI, the concepts in a strategy are likely to be few in number because the topic is usually more focused than is the case for SDI profiles used for current awareness. The use of an extended dialogue to identify concepts therefore need not occur as frequently.

4. Subproblem III

Problem to be solved. Develop content search terms.

Solution in the profile. All subject terms included (controlled, free text, and subject codes).

Relevant data from the Information Form for User.

13 - suggested terms

18 - citations of relevant articles

Relevant data from the transcript.

Discussion relates to generation and selection of lists of subject terms which are to be associated with each concept. Such terms can be selected from a controlled vocabulary, expressed as words and phrases to match free text, or chosen from lists of subject codes for use in searching a particular data base.

Example (Case 1).

IFU: 13 - electron microscopy, ultrastructural studies, microprobe analysis
TRA: U: So, you might come across microprobe studies.

A: O.K.

U: Or x-ray analysis studies that would also be the kinds of things I'm after.

PRO: MICROTERMS: (ELECTRON MICROSCOP*)/ULTRASTRUCTUR*/(MICROPROBE ANALYS*)/(X RAY ANALYS*)

Cf: 01058$ "electron microscopy"

Overview.

The solution to subproblem III constitutes the major part of a profile in most cases and is very much dependent on what data bases are to be searched. Records of a particular data base may be searchable by free text on the title and abstract, by controlled vocabulary terms on major and minor descriptor fields, and by subject code on special code fields. If free text is to be used one must think of synonyms and related terms; if controlled vocabulary is to be used one must find the controlled vocabulary terms in the thesaurus for each data base that is to be searched; and if subject codes are to be used one must identify those equivalent to concepts in the profile. The solution of this subproblem occupied the bulk of time in most of the interviews, particularly when multiple data bases were to be searched and alternative sets of terms had to be generated for each one.

Automatic solutions.

Automatic solution of this subproblem is possible whenever all terms used in the final profile are listed in item #13 of the IFU in groups corresponding to the concepts used in the profile. Automatic solution is also possible for data bases using controlled vocabularies and/or subject codes if all controlled terms and codes used in the profile can be selected by simple matches with free text terms listed in item #13.
Using these criteria 15% of the cases examined contain sufficient information to solve subproblem III automatically.

Intermediary solutions.

Contrary to the situation found for concept identification (subproblem II), inquirers were actively involved in the selection of content search terms. The starting point was usually the terms listed in item #13 of the IFU and intermediaries would prompt the inquirer to think of other terms which should also be included in the profile by asking such questions as:

Can you think of any synonyms?

Remember the computer is searching titles. What words do authors use in titles?

Are there any other related terms?

Do you want to search on specific ones or just the more general term?

Are there aspects of the topic you wish to exclude?

When selecting terms from a controlled vocabulary, the inquirer was also actively involved. In using the ERIC thesaurus, for example, the intermediary would often look in the rotated descriptor display under a term in item #13 of the IFU and ask the inquirer to select terms of interest. These would then be checked in the alphabetical portion of the thesaurus to see if other related terms could be identified for inclusion in the profile. While subject codes (e.g., cross codes and taxonomic codes in Biological Abstracts) were generally chosen by the intermediary, their selection was usually accomplished by a simple match with terms already in the profile. The example given above from Case 1 illustrates this: the term "electron microscopy" was selected as a search term and the cross code "01058" (electron microscopy) was added by the intermediary.
In a few cases the intermediary used knowledge of the subject area as a basis for suggesting terms which the inquirer agreed to include (e.g., names of endocrine glands in Case 1), but most terms were suggested by the inquirer in response to more general prompting by the intermediary (e.g., A: What else? U: Like namely, like pancreas). There were several sources of possible terms consulted during the interview in addition to thesauri and lists of subject codes: reference books for scientific names, a list of a certain class of chemical compounds, indexes of the issues of the printed counterparts to data bases (Chemical Abstracts, Biological Abstracts), a search done for someone else on a similar topic. While the source of all terms in most profiles could be identified as coming from the IFU or having been selected during the interview by one of the means described above, occasionally an additional related term was found in a profile and it was not possible to infer how it had been selected. In the profiles prepared by one intermediary in particular, however, there tended to be many terms which were not in the IFU and had not been mentioned in the interview. This intermediary took the interview only as a starting point, consulting thesauri and possibly other sources to select terms. In these cases there were therefore many terms included in the profile for reasons which could not be completely determined given the evidence available.

Possible interactive solutions.

Subproblem III is well suited to solution in an interactive mode. The intermediary's approach to its solution provides many examples of things which could also be done readily online. In developing free text terms the machine can prompt the inquirer to think of synonyms and related terms. If thesauri are stored in online files, the machine can match free terms already selected and display appropriate portions of the
thesauri so that the inquirer can select appropriate controlled terms. It should be noted that there is a considerable amount of knowledge embedded in thesauri which allows an inquirer to rely on recognition rather than memory to select many terms. In a display of the hierarchical portion of the ERIC thesaurus, for example, the inquirer interested in "cognitive processes" is reminded that many narrower terms such as "learning processes", "memory", and "perception" may be of interest and that these are also further subdivided to give terms on a very narrow aspect of the starting point "cognitive processes". Subject code lists can be checked automatically by machine and any resulting matches displayed for verification by the inquirer.

Two sources of content terms which were not easily exploited by the intermediaries in the cases studied but which become feasible in an online environment are relevant references already known to the inquirer and search profiles already completed. In 42% of the cases examined the inquirer included at least one reference under #18 of the IFU. Provided the citation is included among those online in the data base to be searched, it can be retrieved by the machine and displayed for the inquirer who can view how it has been indexed. Any index terms likely to be used in indexing additional relevant articles can then be selected by the inquirer. Search profiles already completed and available for online retrieval could serve as a profile building resource to supplement those already mentioned. Profiles are probably not best stored and retrieved in their entirety, but rather as miniprofiles, sets of "searchonyms". Searchonyms are not strict synonyms, but terms which can be grouped together because any one of them has a good chance of retrieving documents relevant to a particular concept (Attar & Fraenkel,
1977). Consider group 2 in Case 2 of Appendix E, for example: a new inquirer interested in "dietary lipids" could use this group as a basis for this concept in his own profile, with the machine allowing him to add and/or delete terms as needed to tailor it to his particular interest. But he would not have to start from scratch to generate an entirely new list for his own search.

Intermediary expertise.

While the intermediary played an active part in developing content search terms, prompting the inquirer to think of related free text terms and utilizing appropriate search aids to identify controlled terms and codes, it should be feasible to replicate the solutions to this subproblem through a dialogue between inquirer and machine, since the intermediary in 73% of the cases allowed the inquirer to suggest and select terms. This leaves to the machine the task of prompting the inquirer and assisting where knowledge of controlled vocabularies and subject code lists is needed. Where the intermediary used subject knowledge to suggest terms, the same activity cannot be accomplished easily by machine. But if the machine has available both search aids (controlled vocabularies, subject code lists) and search fragments or miniprofiles, it should have the tools with which to assist the inquirer by providing aids to his memory, rather than being limited to prompts such as "Are there any related terms?". Research on how best to represent these search fragments internally is needed, as well as on how to select fragments most likely to be useful to a particular inquirer. It is not likely that storing all search fragments would be either desirable or feasible, for they are apt to be partially redundant, to vary in quality, and to vary in the frequency with which they prove helpful to other inquirers. The
development of search tools and procedures for their use is a particularly interesting area for AI applications, for it provides an opportunity to explore alternative representations, learning mechanisms, and heuristics for use of the tools developed.

5. Subproblem IV

Problem to be solved. Develop context search terms.

Solution in the profile. All context terms included (authors, journals, language, citations).

Relevant data from the Information Form for User.

14 - acceptable languages
15 - authors of interest
16 - journals of interest
17 - journals not to be included
18 - citations of relevant articles

Relevant data from the transcript.

Discussion relates to the use of context portions of a database record as part of the search profile. Context attributes of a document include its author(s), the journal in which it appears, the language in which it is written, and its citations.

Example (Case 2).

IFU: 15 - J D Castell, D J Lee, R O Sinnhuber, J W Andrews, W G Knipprath, J F Mead, Nicholas Nicolaides, A N Woodall, Masamichi Toyomizu, J H Wales, J Halver

TRA: A: Or if we get all of the papers on fatty acids and lipids in fish will that automatically include those papers by these authors that you are interested in?

U: That should include papers by them.

A: And anything else that these people may have written that's not on that subject you are not particularly interested in, right?

U: Right.

PRO: No authors included.
Overview.

Solution to subproblem IV may be dependent on the solution to subproblem I. Certain context elements such as authors can be searched in any database, but others such as citations and language of publication are searchable in only some databases. For the databases available at UCLA and UGA, for example, citations are searchable only on the Social Sciences Citation Index. In general in the cases examined, citations and authors were used to supplement retrieval by content terms while language was used to restrict the set of documents retrieved using content terms alone. Of the 33 cases, 48% searched authors, 18% searched citations, 15% searched journals, and 12% searched languages.

Automatic solutions.

Considering the frequency with which context elements are searched in the profiles, a simple automatic rule using only the relevant data from the IFU could be formulated as follows: (1) search all authors listed in item 15 of the IFU; (2) search all citations listed in item 18 of the IFU if Social Sciences Citation Index is one of the databases to be searched; (3) disregard any entries in items 14 (language), 16 and 17 (journals). This rule would be successful 48% of the time. Reasons for failure are of two types: (1) profile searched language and/or journal; (2) profile searched only part of the list of authors or citations included on the IFU.

Intermediary solutions.

In solving subproblem IV the intermediary plays a passive role, since in none of the cases examined does the intermediary suggest context elements to search in addition to those already included on the IFU. The method of solution, where it can be inferred, is in all cases
straightforward. Language was included in profiles prepared by only one intermediary who never stated a reason for doing so, but simply used the restrictions indicated in item #14 of the IFU. Other intermediaries occasionally explained to inquirers why language is not a particularly helpful restriction, e.g. it is not used consistently (or at all) in a particular database, English language abstracts for foreign language articles are included on the printout. Inclusion of authors was usually discussed with the inquirer who could restrict the list to those authors felt to be of primary importance or make the decision not to search authors at all. Occasionally the author list was simply used in the profile without any discussion in the interview. Citations were searched on Social Sciences Citation Index whenever a list was provided on the IFU. The reasons for including (or not including) journals cannot be inferred. While 17 IFU's included journal names in items 16 and/or 17, only five profiles searched journals and no justification was given for inclusion or exclusion except in one case where the inquirer wished all journal titles listed to be included as a current awareness service.

Possible interactive solutions.

Solution of this subproblem is well suited to a dialogue between machine and inquirer. Items 14-18 on the IFU are probably not the best format for gathering data on context elements since the IFU does not clearly indicate how they will be used in profile formulation. For authors, for example, the inquirer could be asked to list any authors whose papers should always be retrieved. For citations, the inquirer should be encouraged to provide as many relevant references as possible if Social Sciences Citation Index is one of the data bases to be searched, since citations are the primary retrieval mechanism in this data base.
In the cases examined, journals and language were infrequently used as profile elements. Should they prove to be helpful in other retrieval environments, the inquirer could be questioned by machine as to which journals or language elements should be included when they are searchable on the selected data bases.

Intermediary expertise.

Analysis of solutions for this subproblem did not reveal any types of intermediary expertise which could not be duplicated using a dialogue between inquirer and machine. In fact, many of the intermediaries did not seem to have a clear understanding of when such context elements as journals and language might be useful in retrieval. The role of context elements vis-à-vis content elements deserves increased attention, as suggested by the results of the study of alternative representations described in Chapter VIII. Only through additional research can data be gathered which would be useful to an intermediary in solving this subproblem. For the present, solution is probably best left under the control of the inquirer.

6. Subproblem V

Problem to be solved. Refine terms.

Solution in the profile. All uses of truncation, weighting, alternative spellings, and abbreviations.

Relevant data from the Information Form for User.

No item is always useful. Any source of terms (11-16) may prove useful in particular situations.

Relevant data from the transcript.

Discussion relates to such activities as truncation of terms, assignment of weights, and inclusion of variant spellings or abbreviations for both content and context terms.

Example (Cases 1 and 2).
PRO: Truncation

Blank or 1 letter Secretary tissue$

0 or more letters Lipid*

Abbreviations

Salmo gairdneri
S. gairdneri

Alternative spelling

Cottonseed oil
Cotton seed oil

Overview.

Refinements made depend on the prior solution of a number of other subproblems; data bases to be searched determine the form such terms as author names should take and the content and context terms to be searched determine the set of terms which may be modified in some way in solving subproblem V. Weighting may be used either to limit retrieval of citations to those whose weight is above some threshold or to sort retrieved citations into logical groups. In the cases reviewed only the latter use was found. Truncation possibilities are determined by the software of a particular search center: UCa allows left and right unlimited truncation (e.g. *aryl*) while UCLA also provides for restricted truncation which may be embedded (e.g. wom*n). The need to consider variant spellings and abbreviations depends on the data bases to be searched, since some may not standardize spellings (allowing both British and American) or may make special use of abbreviations. Concern for truncation, abbreviations, and alternative spellings plays a part where free text rather than controlled vocabulary terms are searched. Of the cases examined, 91% make use of truncation, 27% make use of weighting, 52%
contain abbreviations, and 58% contain variant spellings.

Automatic solutions.

Assuming that the solutions to subproblems I, III, and IV are already known, the solution of subproblem V by automatic means can be partial at best. This is somewhat surprising given the seemingly "automatic" procedures of truncation, abbreviation, and identification of alternative spellings, but a careful examination of profiles leads to the conclusion that it would not be possible to duplicate the solution to subproblem V automatically except in those cases where the profile is made up solely of controlled terms and no refinement of terms is necessary (5% of the cases in this study). Truncation can be applied automatically only in the case of author names, for which the entry "Glalock H* M", for example, will match on both "Glalock H M" and "Glalock Hubert M", needed where the form of entry for author names is not standardized in the data base. The application of truncation to content terms cannot be done automatically to duplicate the use of truncation by intermediaries because any simple rule has numerous exceptions. § is often used to match both singular and plural (college§, universit§§§), but also is used to match variant word forms (priest§ to match "private" and "privacy"). The use of * is even more difficult to duplicate automatically because it can replace any number of letters (regulate*, synth*, implant*, immobilize*). And the same decision is not always appropriate (metabolism in one profile vs. metabolism* in another; education§§ in one profile vs. education§§§ in another).

Abbreviations and variant spellings are more easily generated automatically from context and content terms. Again the task is easier for context elements--choosing standard codes for languages to be searched, choosing CODEN's or standard journal abbreviations for journals to be searched.
Since the use of a particular set of abbreviations is data base dependent, the task of generating appropriate abbreviations is well suited to the machine, removing the need for the intermediary or the inquirer to remember the conventions of each particular data base. Abbreviations of content words were of three types: (1) given in item #13 of the IFU by the inquirer (DNA/deoxyribonucleic acid, P.L./Public Law, CD/circular dichroism); (2) common ones which could be stored in a system dictionary (United States/U.S., U.S.A.); (3) standard alternate form of entry for scientific names (Icterus punctatus/I. punctatus, rubidium-strontium/Rb-Sr). All of these could be handled automatically. Spelling variants were almost entirely alternative spacings (di methyl/dimethyl, 3 12/3 12, x ray/xray) which for certain classes of terms, such as chemical, could be generated by looking for standard prefixes and suffixes.

Weighting was used to sort the output of profiles in which there were a number of subtopics or in which one aspect was of special interest to the inquirer. Since weighting was used by intermediaries in only some of the profiles having these characteristics, there is no way to make the decision to weight certain terms and/or term groups automatically that would duplicate results obtained by intermediaries.

Intermediary solutions.

Abbreviations, variant spellings, and truncations were infrequently discussed in the interview, so the basis for the intermediaries' solutions must be inferred from the contents of the profiles. The knowledge used to make these decisions appears to be of two kinds: knowledge of the English language, such as the plural forms of words, and knowledge of the use of terms in a particular subject area, such as what forms of words are likely to be used to represent a particular concept (e.g., only
"transport" and not "transportation" is used with "electron" while the two forms might both be of interest in a search related to transport of goods and people). Use of abbreviations for context elements such as journal titles reflected the intermediaries' knowledge of data base conventions. This knowledge is also evident in decisions on variant spellings since certain data bases may have standard ways of handling such things as British vs. American spelling and punctuation marks which appear within terms (apostrophes, hyphens, periods).

The intermediaries occasionally explained the use of truncation to inquirers, but only in one search dealing with a chemical topic did the intermediary ask the advice of the inquirer on how best to truncate terms. The situation was different with weighting which was used to sort retrieved citations into subtopics in 9 cases. In 7 of these cases the intermediary asked the inquirer whether he would like output sorted, but in all 9 cases the intermediary assigned weights. In the remaining cases weighting was neither discussed in the interview nor used in the profiles, so it was not possible to determine why it was not considered appropriate.

Possible interactive solutions.

Solution of this subproblem offers a good example of an appropriate division of labor between inquirer and machine, particularly in the case of an online system where the dictionary file of a data base is available for display. Rather than trying to develop rules for automatic truncation, the machine can either ask the inquirer to indicate how terms already selected should be truncated or it can present a display of terms from the dictionary file which are alphabetically adjacent to a term to be searched. In the latter case the inquirer can select any related terms of interest and the machine can truncate the original term or add the
newly selected terms to the list of terms already developed, so that all terms of interest to the inquirer will be matched. A limitation of the dictionary display approach is that it cannot deal with left truncation. This was used in three profiles in the cases examined, and it is particularly helpful in searching on chemical compounds (*benz*, *naphth*, etc.). In the case of chemical searches, therefore, the inquirer should be asked to indicate whether any of the terms should be refined using left truncation. The inquirer should be prompted to suggest variant spellings and abbreviations for terms except those (like the names of chemical elements) for which abbreviations can be stored in the system dictionary and added automatically. With respect to weighting to sort output, the inquirer can be asked if he would like the results of alternative strategies to appear separately in the output and if so, the machine can apply weights to accomplish this. Weighting of terms for retrieval of citations above a threshold would involve more work on the part of the inquirer, for he must indicate specifically what term combinations are to occur in order for a citation to be retrieved. The machine could refine context terms automatically, choosing the appropriate format for language codes, journal names, and author names for the data bases to be searched.

Intermediary expertise.

In this case there are three types of intermediary expertise: knowledge of the English language, knowledge of the ways terms are used in particular subject areas, and knowledge of data base conventions. As the above discussion of possible interactive solutions suggests, the first two are kinds of knowledge shared by the inquirer and the third can be embedded in the machine. Solutions of the term refinement
subproblem are possible only interactively at present. While automatic stemming algorithms (e.g. Lovins, 1968) may eventually prove useful in lieu of truncation decisions by the inquirer, the cases examined here suggest that truncation decisions are very context dependent. Retrieval performance of terms refined by machine vs. by the inquirer must be investigated to determine the most effective division of labor between man and machine in the solution of this subproblem.

7. Subproblem VI

Problem to be solved. Develop search logic.

Solution in the profile. Logic operators (AND, OR, NOT) in the profile.

Relevant data from the Information Form for User.

No item is always useful. Any source where logic is implicit may prove useful in particular situations. Items used for concept identification (11 and 12) relate to the use of AND operators between concepts. Item #13 (suggested terms) relates to the use of OR and NOT if the inquirer includes synonyms and excluded uses of terms.

Relevant data from the transcript.

Discussion relates to analysis of the logical relationships between concepts and terms within concepts using the operators OR, AND, NOT.

Example (Cases 1 and 2).

IFU: 12 - Secondarily I'm interested in any literature on electron microscopy of the pituitary gland.

TRA: A: I looked manually in Biological Abstracts yesterday in preparation for this interview just to see what kind of articles and yes, there aren't that many.

U: Right.

A: I looked under pituitary and then I looked under EM and, you know, I can't coordinate too well by eye.

PRO: SELECT IF C1 and PITTERMS; (where C1 is the code for EM)

IFU: 12 - What are the effects of dietary lipids on channel catfish

TRA: A: And specify that there must be one term from each group present in the keyword list and so forth if the paper on lipids in order for it to be printed out.

PRO: G001&G002 (where G001 is the fish group and G002 is the lipid group)
Overview.

Development of search logic is appropriately discussed last, for partial solutions of this subproblem are implicit in the solutions to subproblems II-V. In a retrieval system in which search strategies or profiles are constructed using logical operators, choice of search logic determines the criteria which a document surrogate must satisfy for it to be retrieved. Solutions to subproblem V (abbreviations, variant spellings) lead to alternative forms of a particular term which can be linked by OR's in the profile. Solutions to subproblem III give for each concept a set of searchonyms linked by OR's. If context elements are selected for inclusion in the profile through solution of subproblem IV, individual terms within the set of authors, journals, citations, or languages are linked by OR's. With regard to their relation to the profile as a whole, the search on any set of context elements is an alternative strategy, and thus can be thought of as being linked by OR to any strategy developed using content terms. The exception to this is language which is used to restrict content term statements and thus the set of language terms is linked to the set of content terms by AND. In simple cases where a narrow topic is of interest to the inquirer, concepts identified in subproblem II are linked together by AND. As the search grows more complex, however, concepts are combined in various ways using all the operators—AND, OR, NOT.

Automatic solutions.

The discussion above suggests a simple rule which can be applied automatically, given solutions to subproblems II-V. The machine can create as a profile the set of concepts identified in subproblem II combined in an AND relationship, and all terms within concepts generated
in subproblem III are combined by OR's. If language codes have been selected in subproblem IV, they are linked by AND with the concepts from subproblem II; any other context elements to be searched (journals, citations, authors) are linked by OR's. Applying this rule to the cases examined in this study, 33% of the intermediaries' solutions could be replicated (including Case 2 in Appendix E). There are two major causes of failure for this rule: (1) the profile contained one single concept statement and one or more multiconcept (linked by AND) statements; (2) the profile contained several multiconcept statements. The first situation occurs when the inquirer wishes to see everything on a single concept (e.g. echolocation) as well as material on a related area (e.g. hearing and communication in any species of bat). The second situation occurs when the inquirer has a number of related interests, often clearly differentiated in the prose statement of his search request (item #12 on the IFU), as for Case 1 in Appendix C where the inquirer has labeled his interests "primarily", "secondarily", and "thirdly". The second situation also occurs where multiple data bases are to be searched and different combinations of concepts are judged to be appropriate for each.

Intermediary solutions.

The intermediaries seem to have as a working model for search logic development the rule discussed above under automatic solutions, except that they are able to apply it to concepts which have been placed in groups corresponding to separable subtopics. The machine application of the rule, on the other hand, assumes that all identified concepts relate to a single topic. In profiles where a single concept search was part of the strategy (e.g. retrieve anything that mentions "gummosis"), the intermediary discussed the approach with the inquirer to verify that
the latter would indeed be interested in any document surrogate containing that term. No mention of the use of NOT has been made as yet, and in fact it was seldom used in the profiles examined. Intermediaries would occasionally ask specifically about aspects of the topic which did not interest the inquirer, but the search was built up by using terms which were specifically of interest rather than trying to exclude terms which were not of interest. This cautious use of NOT reflects the operation of the retrieval rule used in systems with logical operators: a document surrogate either satisfies the search strategy or it does not, it is either retrieved or not retrieved. The problem with using NOT is that one risks missing documents of potential interest, as in the case where one tries to limit a search to animal studies by including "NOT human" in the profile which excludes not only studies on humans but also studies dealing with both animals and humans. While the inquirer did not actually generate the logical statements for his profile, in many cases the intermediary had the inquirer verify the statements, agreeing that the concepts and their combinations which the intermediary had developed should in fact retrieve what interested him.

Interactive solutions.

The proposal for interactive solutions of subproblem II presupposed that the inquirer could identify the concepts which could be used to represent his area of interest. Likewise interactive solution of this subproblem depends on the inquirer's ability to group the concepts into sets which can then be used by the machine to generate logical search statements. The remaining steps of search logic development can be accomplished automatically by machine--inclusion of context elements in the proper relations to other portions of the profile and linking
of searchonyms and variant word forms associated with a particular concept. Since an understanding of concept identification (subproblem II) and search logic development (subproblem VI) are closely related, any method of instruction of the inquirer should include both. Once the inquirer has identified and grouped concepts, the machine can combine the partial solutions implicit in solutions to subproblems III-V to generate a completed profile which is in a format acceptable for processing.

Intermediary expertise.

As was found for other subproblems, intermediary expertise in this case was a combination of knowledge of subject areas and knowledge of the search system. Knowledge of subject areas allowed the intermediary to distinguish various subtopics within an inquirer's area of interest, each of which could correspond to a separate logical statement in the profile. Knowledge of the search system allowed the intermediary to construct a profile processable by machine and containing the proper combinations of solutions to subproblems previously solved. Whether the final profile is developed by intermediary or interactively, it represents a combination of solutions to all the subproblems. Research directed toward developing a better understanding of solution of individual subproblems ultimately can contribute to a better understanding of the profile development process as a whole.

8. Tutorials

While the purpose of the analysis done for this study was the identification of processes used by the intermediary in solving the various subproblems in the task of profile development, at least a brief description should be given of the role of the intermediary as teacher. Since the inquirers in most of the cases examined were first time users of the
system, the intermediaries spent a significant proportion of their time giving information to inquirers rather than getting information from them, so consideration of the interview solely as an information getting activity is a misrepresentation. Some examples of tutorial activity are included in the Instructions for Content Analysis in Appendix A. The types of tutorial activity include:

(1) profile retrieval mechanisms—such features as logic, weighting, and truncation;

(2) search process—a narrow vs. a broad search, what parts of the document surrogate are used in searching (abstract, title, identifiers, etc.), the importance of accurate spelling given the way the machine matches terms for retrieval;

(3) data bases available—subject coverage, type of material coverage, time period covered;

(4) searching free text—need for abbreviations and variant spellings, special features of individual data bases (e.g. the form of chemical terminology in Chemical Abstracts, augmented titles in Biological Abstracts);

(5) searching controlled vocabularies and subject codes—structure of a thesaurus, uses of subject codes, need for different strategies in different data bases, indexing policy (e.g. depth of indexing, use of the most specific heading available);

(6) output format and timing—how to interpret a sample printout, frequency with which printouts should be received;

(7) related services—online systems, search centers elsewhere, commercial DOI services, translators, costs of other services, use of printed tools (e.g. principles of citation indexing, organization of indexes in issues of Biological Abstracts).

An important function of tutorial activity is to give the inquirer a good understanding of the capabilities and limitations of the computer-based system, so that his expectations about its output are realistic. Related to this is the need to stress the inquirer's role in the profile modification process, evaluating the output and making suggestions for changes in an attempt to improve profile performance. While tutorial activity has not been considered in detail in this study, it is an
important part of the interview between intermediaries and inquirers who are first time users of the system. Whether the need for tutorial activity decreases substantially as the inquirer has more experience with the system cannot be determined from the data available for this study. The types of tutorial activity listed above suggest, however, that much of the instruction is needed only by a new system user. Further study of the tutorial function of intermediaries, as well as their problem solving processes, is needed if one is to build retrieval systems with which the inquirer can interact directly.

9. Sequence of subproblem solution

The analysis in this study differed from that in the UGA-UCLA study in that the objective was not to describe the time sequence of events in the dialogue between intermediary and inquirer. The inquirer-machine dialogue need not mimic in every detail the inquirer-intermediary dialogue. What is necessary is identification of the knowledge and processes which are required to convert the interest statement to a profile which can be submitted to a retrieval system.

The discussion of subproblems has clearly illustrated that the solution of one or more subproblems may be needed before another can be solved, e.g. one cannot refine terms before they have been selected. Viewing the solution process as one of a sequence of subproblem solutions allows one to focus on the particular knowledge and processes needed to solve each subproblem individually. As Sussman observes (1975, p.117): "This idea of 'linear approximation' turns out to be a powerful complexity-limiting device." Some of the complexity evident in the dialogue between inquirer and intermediary is due to the fact that one may partially solve a particular subproblem, go on to another subproblem, and subsequently return
to the first to complete its solution. The UGa-UCLA study identified
two patterns of solving subproblems II and III (Carmon, 1975, p. 35):

(1) all major concepts in the query are first identified and tagged,
after which attention is then directed to vocabulary considerations;

(2) the first concept is identified and some vocabulary operations
are carried out, then the second concept, third concept, etc.
The first pattern corresponds to the sequential solution of subproblems
described in this study, while the second deviates from this by developing
partial solutions to subproblems II and III and alternating between them.
To identify whether different sequences of subproblem solution simply
reflect the intermediaries' personal preferences or whether in fact the
"linear approximation" is too rigid and not well suited to the solution
of some profile development problems, additional study is needed.

E. Suggestions for further research

This study, unlike that reported in Chapter VIII, is a Gedanken-
experiment. It involves a comparison of three systems whose task is to
develop SO1 profiles: a hypothetical machine system (with automatic
solutions), the inquirer-intermediary system (with intermediary solutions),
and a hypothetical inquirer-machine system (with interactive solutions).
The machine alone is likely to have limited success in solving the profile
development problem, but the inquirer-machine system is a possible alter-
native to the currently existing inquirer-intermediary system. Before
outlining areas for further study, it is helpful to illustrate how this
study fits within the framework of AI in IR by reviewing it in terms of
the engineering design process, the general method of investigation for
AI studies.

In Chapter II, the five stages of the process were identified:
problem identification, information gathering, idea generation, model
building, and testing.

Stage 1. Problem identification. Most inquirers cannot use computer-based IR systems directly, but must instead work with an intermediary who assists in translation of the initial query into a form which the system can process. If systems are to be made accessible to a wider range of people who do not have intermediaries available, it is necessary to study the profile development process in some detail to see what parts of the process could be accomplished by machine or through a dialogue between inquirer and machine.

Stage 2. Information gathering. The AI literature on problem solving suggests problem reduction as a framework for developing a model of the problem solving process in profile development. But as a source for identifying subproblems in profile development, it was necessary to examine process models of reference and profile development reported in the library and information science literature. The stages in these models were examined to select a set which could be viewed as subproblems, the solutions of which taken together would constitute a profile. The details of the subproblem solution processes were not given in sufficient detail in the literature to provide many ideas for the design of either machine or inquirer-machine systems.

Stage 3. Idea generation. Protocol analysis of interview transcripts was the tool used to generate ideas for both representations and procedures, the knowledge and processes used by the intermediary in solving the various subproblems. The results were suggestions for the types of knowledge which should be embedded in the machine and the types of processes, both automatic and interactive, which could be used to assist the inquirer in developing a profile. While the proposed representations
and procedures were not specified in the detail necessary to implement
them directly on the machine, they can provide guidelines for ways in
which existing systems can be extended from systems which simply accept
commands to systems which actively assist the inquirer not only in pro-
cessing his query, but also in formulating it.

Stage 4. Model building. The models in this study are the machine
system and the inquirer-machine system. Since this was a Gedankenexperiment,
they were not actually implemented. A Gedankenexperiment has been worth-
while, however, for it suggests that if one does actually implement the
proposed models, an inquirer-machine system is more likely to match the
performance of the inquirer-intermediary system than is the machine
acting alone.

Stage 5. Testing. In this study the only tests involved comparison
of the ability of a machine system or an inquirer-machine system to create
profiles identical to those created by an intermediary working with an
inquirer. The logical next step, provided one first actually implements
the inquirer-machine system outlined in the section "possible interactive
solutions" under each subproblem, is testing for human acceptance. This
testing has two aspects: determining whether inquirers will work directly
with an interactive system rather than working with intermediaries and
comparing the retrieval performance of strategies or profiles developed
interactively with those developed in cooperation with an intermediary.

1. Philosophy of system design

A: See, you have to know the language in order to
write a profile. So you can't just walk in and
get your information, see there has to be an
interface. (UCLA, #2122)

The attitude expressed by the intermediary in the above passage
is probably typical of many intermediaries working with existing opera-
tional systems. These systems are designed in such a way that they are
difficult to work with effectively unless one has a good deal of training.
The results of a survey conducted during 1974-1975 (Wanger, Cuadra, &
Fishburn, 1976) showed that in most of the organizations surveyed, search-
ing of online systems was done by an intermediary rather than the end
user of the output. Institutions currently using online IR systems in-
clude libraries in industry, government agencies, and universities. A
pilot study of online system use in public libraries (Firschain & Summit,
1977) showed that there is also a market for online searches in public
libraries. One of the major problems found in this study was the increased
demand on reference librarians' time, for to provide the service they had
to conduct presearch interviews, formulate search strategies, search on-
line, and assist the inquirer in utilizing the output. Until systems are
designed where one can "just walk in and get your information" in the
sense that the inquirer can successfully interact directly with the machine,
use of online systems will be limited primarily to those situations in
which a trained intermediary is available and has the time to work with
individual inquirers.

The outline of possible interactive solutions in this study suggests
a different philosophy of system design from that implicit in current
operational systems. Online systems should be designed so that they can
be used directly by the inquirer to develop SDI profiles, formulate search
strategies for retrospective searches, or perform whatever additional
tasks are found to make up the problem domain for IR systems in the
future. This interest in providing direct access to the inquirer is
found in some recent experimental systems. For example, Park (1977),
in developing her Inquiry Process Model, felt that the optimal approach
is one which combines the judgmental capabilities of the inquirer with the management facilities of the computer in an interactive man-machine system which capitalizes on the different strengths of each component (human and computer). In this model the computer mediates the inquiry process by presenting to the inquirer the decisions to be made, the choices (outcomes) which are available, and additional information which the inquirer may need to make the choice. The inquirer is the decision agent except in those cases where contingent relationships between two or more decisions predetermine the choices, thus allowing them to be made automatically.

There are two extremes in responsibility for query formulation (Gennett, 1973):

(1) automatic search—inquirer states his information need and the machine digests the statement, perhaps requesting additional data from the inquirer if it recognizes that needed information is missing. The machine develops a search strategy, compiles a data base subset, and delivers a bibliography of document references back to the inquirer.

(2) user-controlled—inquirer has the responsibility for issuing commands to control development of a search.

The interactive solution outlined above lies between these two extremes, considering which roles are best performed by inquirer and machine and how their activities can be combined. It tries to build on the particular knowledge and capabilities possessed by human and computer. The inquirer often has extensive subject knowledge as well as knowledge of English, while one can build into the machine knowledge of data base characteristics, indexing vocabularies, and rules for constructing profiles with acceptable formats. As one turns to the problem of search strategy
development for online systems providing access to a large number of data bases with diverse characteristics, it will probably be desirable to provide more automatic aids for such functions as data base selection and vocabulary mapping. But the inquirer should still have some control over the problem solving process, at those points where his skills can lead to more satisfactory and/or more rapid solution of certain sub-problems. The discussion of interactive solution exhibits a concern for external representations, the displays at the man-machine interface to aid in subproblem solution. This is an area which must be investigated in some detail before systems can be used directly by the inquirer.

The philosophy underlying the proposed inquirer-machine system can be summarized as follows: anyone who sits at a terminal interacting with a computer-based IR system should feel the presence of another active intellect who is using the computer as an effective agent and thus providing greater access to his retrieval skills. Investigations of AI applications in IR are directed toward exploring ways in which some of the skills currently held only by human intermediaries can be transferred to the machine so that access to computer-based systems can be made available to many more inquirers.

2. Training: Machines, Intermediaries, and Inquirers

A: And we don't have the training although we've been trained and we don't understand and we don't want to do it, so it's been a problem. (UCLA, #2122)

In the discussion above concerning the philosophy of system design, it was observed that the design of existing operational systems has implicit the assumption that they will be used by individuals knowledgeable about the command language and other system features, i.e. people who have had some training in the use of the system. With the rapid spread
of online systems, more librarians are serving as intermediaries to these systems. While the ultimate goal of AI studies is to identify knowledge and processes which can be incorporated into the machine, in effect training machines to be more intelligent partners in query formulation and search strategy development, in the short term one benefit of studies like that reported here will be ideas on how better to train intermediaries.

The passage cited above suggests the frustration poorly trained intermediaries can feel when confronted with problem solving tasks, such as profile and search strategy development for existing computer-based systems, which are admittedly complex. Viewing profile development as a solution of a sequence of subproblems is one way to reduce this complexity. The resulting profile development process is more systematic, than that used by some of the intermediaries in cases examined, but the outcome can still be tailored to the needs of individual inquirers.

While this study did not include cases where the same search request was handled by two different intermediaries, studies of online searching comparing strategies for the same questions developed by different intermediaries show substantial differences. A recent study comparing the performance of several searchers on two test questions showed differences in: (1) file selection, from two to four files per question with agreement neither on file choice nor on order of preference; (2) term selection, with differences in both the particular terms searched and the number of terms searched; (3) logic, from simple to extremely complex (Oldroyd & Citroen, 1977). While this study did not evaluate the outcome of the search strategies in terms of relevance of items retrieved, the authors noted their concern that the searchers seemed to pay little
attention to differences in data base characteristics and the adaptations necessary because of these differences. If a strategy was considered effective in one file, it was often continued in the next file that was searched, even if the files had completely different vocabulary structures. Commenting on the results of a SUPARS evaluation study which compared performance of different intermediaries using the same questions, Katzer (1972) observed that if differences between trained users of retrieval systems are so large in general, better training or a more adaptive interactive language might contribute more to improving cost performance characteristics of a system than costly software developments.

The cases examined in this study represent: (1) a number of different intermediaries (four at UCa and seven at UCLA); (2) a number of different data base characteristics (social vs. natural sciences, citation vs. subject indexing, controlled vs. free text); and (3) searches of one vs. multiple files. More extensive studies of the work of experienced intermediaries should provide more insight into the profile development process. While a long term goal is to build more capable machine systems, in the short term the utility of the problem reduction model as an aid in training intermediaries should also be investigated.

Viewing the type of study reported here as a source of ideas for training is consistent with an observation made by Shera (1964a) several years ago:

The advent of mechanization in the library, if it does nothing more, has been justified by encouraging a few students of the library process to break through the hard crust of conventional thinking about library operations, thus forcing an awareness of the total library system as a configuration of closely interrelated parts. A new understanding of librarianship may eventually prove to be the greatest single gift of automation to the library world. (p. 7)
In addition to training machines and intermediaries, one should also consider the training of inquirers. The understanding of the sub-problems of concept identification and search logic development which the inquirer would need to work with the proposed interactive system is also useful at present in situations where the inquirer works with an intermediary. Such training should provide a sense of the type of questions most appropriate for online IR systems. The findings of a current study on individualized instruction for data access (Meadow, Hewett, Rafsnider, Toliver, Epstein, Edelman, & Maher, 1977) should provide additional indications of the types of instruction needed by inquirers working with online systems directly. Their system has two parts: (1) CAI—to assist bibliographic data base users in learning to search without the benefit of an intermediary and (2) Monitor—to trace the progress of an inquirer performing actual searches and to provide active assistance when help is needed.

3. Conclusion

In order to put the study reported in this chapter in perspective, it is of interest to cite a passage from Dijkstra, whose work on structured programming has provided better understanding of the programming process (Dijkstra, 1968):

"One can remark that I have not done much more than to make explicit what the competent programmer has already done for years, be it mostly intuitively and unconsciously. I admit so, but without any shame; making his behaviour conscious and explicit seems a relevant step in the process of transforming the Art of Programming into the Science of Programming. My point is that this reasoning can and should be done explicitly. (p. 175)"

If one replaces "programmer" by "intermediary" and "Programming" by "Profile development" in the above passage, one has an apt characteriza-
tion of what studies like that reported here may contribute to IR. This is not to suggest that the human processes in profile development will be completely reducible to machine processes. The opposite should be evident from the emphasis on online systems and inquirer-machine interaction. Rather, studies can aid in the identification of those human information processing capabilities which can be implemented on the machine.
CHAPTER X

CONCLUSION

A. Summary of the study

This study reports results of research which has explored possible contributions of artificial intelligence to the design of information retrieval systems. It may be summarized in terms of the stages of the design process introduced in Chapter II and used to characterize the methods for the investigative studies. The five stages of the process are: problem identification, information gathering, idea generation, model building, and testing. As indicated in Figure 3 (p. 39), the study includes work at two levels: analysis of areas of application of AI in IR (Chapters III-VII) and investigative studies (Chapters VIII and IX). The first three stages of the design process are covered in the analysis of application areas; the investigative studies carry specific problems through all five stages.

To summarize this study, each of the five stages is described separately:

Stage 1. Problem identification. The problem is to explore possible areas of application of AI in IR. This problem as stated is too general and vague to be useful. Using a problem reduction approach, the problem is broken up into four subproblem areas, corresponding to four major concepts in AI: pattern recognition, representation, problem solving, and learning. These concepts are introduced in Chapter II together with an interpretation of their use in the context of IR. Each concept is
then discussed separately in Chapters III-VI, where it is further subdivided into application areas. Together there are ten: automatic indexing as a feature selection problem, pattern classification, measures of similarity, internal representations, external representations, question answering as a theorem proving problem, heuristics in IR, ill-structured problems, short term learning, and long term learning.

Stage 2. Information gathering. The discussion of each of the ten application areas begins with two sections which constitute information gathering: surveys of work representing (1) IR approaches and (2) AI approaches relevant to the application area.

Stage 3. Idea generation. The third section of the discussion of each application area involves enumeration of specific problems for research, questions to be answered to determine the applicability of AI concepts and techniques in IR. Chapter VII presents an overview of the results of the idea generation stage.

The investigative studies also go through the stages given above for the research problems "processing documents as queries" (Chapter VIII) and "query formulation as problem reduction" (Chapter IX). The first study focuses on activities of query processing by the machine and the second study focuses on the activities of formulating a query in such a way that it can be processed by the machine. This complementary character of the two studies is also evident in the last two stages of the design process:

Stage 4. Model building. In the first study the model is based on consideration of how the machine might identify documents "like" a query document, with little regard for how a human might perform the same task. In contrast, in the second study the model is largely based on the
methods which humans use to convert user interest statements into SDI profiles.

Stage 5. Testing. While both investigative studies do not include testing for human acceptance, they involve different types of comparison. The first study compares the results from different machine models, the sets of items retrieved when using alternative internal representations. The second study compares performance of a machine model with human performance, the SDI profile which a machine can produce from a given user interest statement vs. that produced by a trained intermediary.

Taken all together, the analysis and studies completed represent definition and initial exploration of possible artificial intelligence applications in information retrieval.

3. Significance of the study

The contributions of this study lie in three areas: identification of conceptual areas common to AI and IR, outline of a research program within this conceptual framework, and completion of two investigative studies.

1. Analogies

Study of AI in IR must begin with an identification of analogies, pointing out areas of common interest by establishing links. Such a study acknowledges the "opportunity for enrichment of research in librarianship through synthesis with other disciplines, some of which are themselves quite new and as yet not fully formalized" (Shera, 1964b, p. 148). Since both AI and IR are relatively new and developing areas, terminology is not yet standardized and similarities may be masked where a common vocabulary does not exist. While Licklider's description of a procognitive system (1965) did point to possible relations between AI
and IR, no work since has systematically identified overlapping areas of concern. In fact two recent papers in information science journals have questioned the desirability and feasibility of applications of AI in IR. Rosenberg (1974) questions the desirability of AI approaches, since he feels that AI is responsible for the "gestalt of the computer", the view that human information processing is analogous to machine procedures. In his opinion, people working in AI have developed a vocabulary and have given rise to popular ideas that suggest a basically anthropomorphic nature of the computer. A different view of AI research has been presented by Boden (1977):

Contrary to what most people assume, this field of research has a potential for counteracting the dehumanizing influence of natural science, for suggesting solutions to many traditional problems in the philosophy of mind, and for illuminating the hidden complexities of human thinking and personal psychology. (p. xi)

The discussion in this study has tried to demonstrate that applications of AI techniques, rather than posing a threat to individuality, can be used to enhance retrieval system performance under the direction of individual users. It is a view also held by Schneider (1974):

To many people the computer has become a symbol of all that is impersonal, mechanistic, and dehumanizing in our society... But their real target should be the unimaginative computer system designer. Contrary to the popular myth which holds that computers force us into rigid molds, if used constructively they actually allow for greater individualization of service. (p. 243)

Questions as to the feasibility of AI applications in IR can only be answered through experimentation. Barracough (1977) in a review of online searching in IR cites a critic of AI to support her position that AI is not likely to prove very useful to IR. She notes that potential for progress in AI has been questioned by Lighthill particularly where,
as in IR, understanding of wide areas of knowledge is necessary. The discussion in this study has suggested a number of different areas of AI of potential relevance to IR and Chapter VII in particular has tried to identify those areas of investigation likely to be most fruitful, given the state of the art in other AI applications.

The papers by Rosenberg and Sarracloough cited above illustrate that an awareness of and support for possible AI applications in IR are by no means as yet widespread in the information science community. The UNISIST study report (Unesco, 1971b) did list AI among subject areas for research and development, cautioning that AI methods "are sometimes presented somewhat naively, as the final answer to the more ambitious goals of information processing in science and technology" (p. 108). The report goes on to say that "UNISIST adherents should be aware of the issue, and develop their own doctrine as to the areas of information handling where such methods can be expected to yield useful results in a foreseeable future" (p. 108).

Though the UNISIST report appeared eight years ago, it has not led to many studies outlining "areas of information handling where such methods can be expected to yield useful results". A survey of information storage and retrieval by Minker (1977) does identify some relevant work, in particular natural language processing and question answering systems. But the author's identification of analogies beginning in an earlier article (Smith, 1976) and stated more clearly and concisely in Chapter II of this study, appears to be the most complete assessment to date. That others have not done similar studies is somewhat surprising, given the common interest of AI and IR in "symbol crunching" (Winograd, 1974) rather than "number crunching" applications. With the advancement of
online systems, IR becomes an issue of dealing mechanically with large
data bases in an intelligent way, thus making IR a legitimate application
area for AI. But an identification of analogies is only the first step,
for "an analogy can be more or less detailed and hence more or less
informative" (Lorenz, 1974, p. 230). Once recognized, analogies must be
explored in greater depth to identify meaningful research problems.

2. Images of potentiality

Scientists and technologists are guided by "images
of potentiality"—the untested theories, unanswered
questions, and unbuilt devices that they view as
their agenda for five years, ten years, and longer.
(Paisley & Butler, 1977, p. 42)

This study began in Chapter 1 with a discussion of two "images of
potentiality"—Bush and his memex for IR, Turing and the machine playing
the imitation game for AI. The contribution of Chapters III–VII takes
two forms: specifying characteristics of these "unbuilt devices" as goals
for AI applications in IR and identifying "unanswered questions" which
must be investigated in order to reach the specified goals.

The images as originally presented by Bush and Turing are not entirely
adequate as goals for research thirty years later. The image of the memex
as described by Bush is inadequate, given the new capabilities made pos-
sible by technologies not available in 1945 when memex was proposed. Two
such limitations include (Paisley & Butler, 1977):

(1) no networking or communication capability either with other
information resources, which may be valued for the trails that other
users have created through them, or with other users directly;

(2) no transformative power over symbols in text.

Turing's test as an image is similarly inadequate. For one thing
it may not be necessary to converse in natural language for many IR
applications. As Amarel (1968) has observed, it is not clear that "analysis of natural language text is as important an element in computer-based information systems (at least for the near future) as the analysis and manipulation of other artificial languages that will be used in such systems for describing documents and for formulating queries" (p. 96).

Another weakness of Turing’s test is more subtle. It may be misleading when applied to IR systems because the capability of fooling human intelligence may not be the same as the capability necessary to aid it. More pertinent tests or guiding ideals for research in this field would involve the ability of machines to react effectively to requests for help which humans typically address to each other, as in the case of the use of machines in lieu of human intermediaries in IR systems.

Long range goals for AI applications in IR include making systems accessible to a wider range of users, enabling the machine to take over routine chores while helping the user with more complex tasks. By nature long range goals—the ultimate, frequently idealized ends toward which research efforts are directed—are specifiable only through a hierarchy of successively more limited targets. The contribution of Chapters III-VII is the establishment of a research program, directed toward attainment of long range goals. Given the conceptual framework defined by the analogies described in Chapter II, the research program outlined directs emphasis to the study in depth of well chosen problems, such as the problems of document representations, search strategies, and their interactions.

3. Exemplars

This study would not have been complete without the inclusion of reports of the two investigative studies in Chapters VIII and IX. They
provide examples of studies addressing research problems identified in
the analysis phase and demonstrate the use of the general method for
the design process outlined in Chapter II. As noted in the study reports,
further study is needed, in that evaluation of the models built has stopped
short of testing for human acceptance. The suggestions for further research
which constitute the final section of each report outline approaches to this
phase of testing. The investigative studies can be thought of as what
Kaplan (1964) terms "heuristic", serving "to generate ideas, to provide
leads for further inquiry or to open up new lines of investigation" (p.
149). The contribution of the investigative studies can thus be judged
in two ways: as examples of subject matter and method for the investigation
of problems from the research program outlined in Chapters III-VII and
as sources of leads for further inquiry.
C. Practical applications

The focus throughout the study of AI applications in IR has been
online systems. What is proposed is a type of technology transfer.
Hardware technology for online systems is already available in terminals,
communication networks, etc., but software engineering for IR has in the
past concentrated primarily on data base maintenance. In order to suggest
the environments in which AI applications in IR may be investigated, three
elements (experimental mode, hardware, and software) are described briefly
in this section.
1. Experimental modes

Two alternative experimental modes are possible for the incorporation
of AI techniques in IR systems: modification of the existing large
operational systems or development of new (probably small-scale) systems.
In the short term it is not likely that operational systems will be
modified directly. Instead, modifications can be accomplished through approaches such as distributed processing, where the large operational system with its many data bases serves as a host system. In distributed processing a minicomputer can perform device simulation, emulating a normal user terminal so that the host data bases and operating system need not be modified to perform experiments. In the case of AI applications in IR, experimental programs can be written for the satellite minicomputer to process data retrieved from the host system. The feasibility of this approach has been demonstrated in an experiment by Carson (1977), who used a minicomputer to provide access from remote sites to MEDLINE, with additional facilities for local processing by programs written in MUMPS.

The alternative to modification of operational systems is the development of new systems with modular designs allowing easy changes in various components for testing. An example of such a system is SIRE, used in the study reported in Chapter VIII. Such systems are not limited only to conducting experiments, for they are likely to find use as both personal information retrieval systems and as IR systems for data bases built up by research communities. The latter case is exemplified by SIRE, which currently provides access to data bases belonging to individual researchers and research groups at Gallaudet College.

2. Hardware

The discussion above on distributed processing has already mentioned minicomputers as one type of hardware which will facilitate the investigation of AI applications in IR. Another is the intelligent terminal. Both can allow local processing of sets of items retrieved from the host data base(s), performing such functions as analysis and editing. While
minicomputers and intelligent terminals are in wide use now, data base
machines may also be important in the future. In a review of research
in this area, Berra (1977) has stated that "the objectives are to imple-
ment many of the traditional software data base management functions in
hardware in order to increase performance and capability" (p. 5). Other
types of hardware which may eventually be used for AI applications in
IR are devices at the man-machine interface, for commanding and inputting
(e.g. the touch screen) and for displaying and outputting (e.g. the plasma
terminal). A recent report provides brief descriptions of the devices
currently available (Paisley & Butler, 1977, pp. 64-75).

3. Software

It is likely that IR systems could benefit from some of the software
already developed for other AI applications. The set of unique capabili-
ties provided by a particular programming language make some types of
programs easier to write in that language than in any other. Programming
languages for AI store, access and manipulate lists of symbolic information
with capabilities for expression retrieval and pattern matching. They
permit representation of associations and complex symbol structures formed
from new data types such as sets. A review of AI languages is given by
Bobrow and Raphael (1974). One of these languages--SAIL (Stanford Arti-
ficial Intelligence Language)--has been used to program SIRE.

D. Code

Code--a passage of more or less independent character
introduced after the completion of the essential parts
of a movement, so as to form a more definite and
satisfactory conclusion. (Murray, 1893, v. 2, p. 582)

At the end of Chapter II, it was suggested that artificial intelli-
gence in information retrieval can be viewed as a new paradigm to be
considered by the research community. In this concluding section, the
intent is to characterize this paradigm and to make a final statement as to what may be gained through its adoption. In this author's view, the proposal to study AI applications in retrieval systems has intrinsic to it the two attributes which Rosenberg (1972) has proposed for a new paradigm in information science:

(1) computer viewed as a tool of man—"setting the basis for genuine man-machine interaction. The interaction will not be trivial human response to a computer dominated system, but a computer response useful to a human system" (p. 103).

(2) basic change in design philosophy—"A methodology is needed to deal with truly dynamic phenomena, changing needs, and conditions. This means adaptive flexible systems. The idea of the system design will give way to a concept where a system is continually redesigned and restructured to adapt to changing needs" (p. 103).

The first of these has been expressed from the system designer's viewpoint by Mooers (1959):

At all times, it is important to remember that it is the human customer who uses the information retrieval system who must be served, and not the machine (p. 81); as well as from the humanist's viewpoint by Bellow (1974):

A million years passed before my soul was let out into the technological world. That world was filled with ultra-intelligent machines, but the soul after all was a soul, and it had waited a million years for its turn and did not intend to be cheated of its birthright by a lot of mere gimmicks. It had come from the far reaches of the universe, and it was interested but not overawed by these inventions. (p. 59)

As the discussion throughout this study has tried to demonstrate, explorations of AI applications in IR systems have as an objective the development of IR systems as more effective tools both through delegation
of retrieval tasks to the machine and through acceleration and augmentation by machine of tasks performed by the human user. Research problems such as those cited under the category of "external representations" in Chapter IV are directly concerned with increasing the "user-orientatedness" of an IR system.

The argument in support of a change in design philosophy has likewise appeared in other contexts:

The memory institution cannot use a rigid, static structure for organization of its information but must provide as much basis as is practical for automatic reorganization when new information needs are recognized (Information Systems Panel, 1972, p. 20).

Information should not build up a dead structure; the body of knowledge is in continuous evolution and it is vital, in order to forecast and influence the future, that information should contain at least the seeds of tomorrow’s progress and discoveries. What distinguishes modern information from traditional documentation is precisely the introduction of this heuristic element (Pignoli, 1971b, p. 13).

It is one thing to advocate the desirability of dynamic systems and quite another to actually provide a mechanism with which to accomplish change. The discussion of learning with both short and long term effects in Chapter VI has begun to provide guidelines for how this dynamic quality might in fact be implemented in IR systems.

The suggestion that AI in IR be taken as a framework for research and system design is not made with unbridled optimism. For one thing, as the limits of discussion outlined at the end of Chapter I illustrate, it is a limited view, just as every world view is restricted. Information science is an interdisciplinary science derived from and related to such fields as mathematics, logic, linguistics, psychology, computer science, operations research, communications, library science, and management (Sorko, 1968). Many of these fields are relevant to the design of IR
systems. Where this is the case, there is a need for systematic review of overlapping areas of concern, just as this study has done for AI in IR. As another example, relations between linguistics and information science have already been explored in some depth (Montgomery, 1972; Sparck Jones & Kay, 1973).

Optimism is also tempered when one acknowledges the arguments of the critics of AI. Some, as personified by Dreyfus (1972), try to demonstrate the infeasibility of a computer system's capturing the entire range of intelligent behavior exhibited by humans. More recently Weizenbaum (1976) has questioned the desirability of certain AI applications, asserting that (1) there is a difference between man and machine and (2) there are certain tasks which computers ought not be made to do, independent of whether computers can be made to do them. There are thus likely to be limits to AI applications in IR, set both by what is found to be feasible and what is judged to be desirable. The proposal to study AI in relation to IR does not imply that ready answers to open questions will necessarily be forthcoming. Rather, awareness of the existence of common problem areas may increase the flow of new developments in both directions.

In advocating an investigation of AI in IR, this author agrees with Jackson's (1974) assessment in describing the transition from "toy" problems to practical applications that "the harvest of artificial intelligence may be for the good of humanity" (p. 398). This echoes Brewster's (1832) assessment of the benefits which the study of automata yielded:

Those mechanical wonders which in one century enriched only the conjurer who used them, contributed in another to augment the wealth of the nation; and those automatic toys which once amused the vulgar are now
employed in extending the power and promoting the
civilization of our species. In whatever way, in-
deed, the power of genius may invent or combine,
and to whatever low or even ludicrous purposes that
invention or combination may be originally applied,
society receives a gift which it can never lose;
and though the value of the seed may not be at once
recognised, and though it may lie long unproductive
in the ungenial till of human knowledge, it will
some time or other evolve its germ, and yield to
mankind its natural and abundant harvest. (pp. 258-259)

This analysis is followed by a description of "Mr. Babbage’s machine",
of which it is said:

The effects which it is capable of producing, and
the works which in the course of a few years we
expect to see it execute, will place it at an in-
finites distance from all other efforts of mechanical
genius. (p. 267)

The exploration of possible contributions of artificial intelligence
to the design of information retrieval systems represents yet another
step in the process of studying "the effects which [the computer] is
capable of producing".
APPENDIX A

INSTRUCTIONS FOR CONTENT ANALYSIS

The materials to be analyzed are transcripts of dialogues (interviews) between a user (U) and a search analyst (A). Analysis of the text of the transcript should lead to division of it into segments according to the following rules:

(1) Wherever discussion relates directly to profile development, the text is to be labeled with a number (between I and VI) denoting which of the subproblems in profile development is under consideration. Definitions of each subproblem and sample passages are given below.

(2) Segments of text in which the discussion is tutorial in nature should be so marked. Instruction can cover such things as features of the search system, characteristics of particular data bases or controlled vocabularies, the types of output available, and procedures for profile revision.

(3) Other segments of text may include discussion which is social, administrative, or general in nature. These should be marked as excluded from analysis.

Notes to aid analysis:

(1) Some segments of text may relate simultaneously to more than one subproblem and should be labeled accordingly with as many numbers as are required.

(2) The focus of discussion may shift back and forth so that consideration of a particular subproblem may occur at several points. Each occurrence should be labeled with the proper subproblem number.

(3) Consideration of a new subproblem often begins with a question by the search analyst. At other times the user makes a statement which introduces a new subproblem and the analyst follows up with a question. Both patterns should be looked for so that the entire discussion related to a particular subproblem is properly labeled.

(4) Some subproblems are discussed at length in the transcripts; others may not occur at all in particular interviews. The list of six subproblems is intended to be exhaustive, i.e. any discussion directly related to profile development should be classified in one of the six subproblem categories.

Subproblem Definition (label text with numbers I-VI)

I. Identify Data Bases to be Searched

Discussion relates to selection of the data bases to be searched from among those available on magnetic tape at the information center.

Ex. A: OK, now, for indexing sources you checked Chemical Abstracts
and Government Reports Announcements.

U: Right.
A: I would also strongly suggest looking at Engineering Index.

A: Do you have any feel for if there is any information in anything other than GEDREF?

U: I sort of wondered what there was in terms of direct mining subscriptions, things like Engineering Mining Journal and ... might include it, but I'm really not aware of ...

A: OK, now we could search Engineering Index, they do cover some mining engineering and some mineralogy.

A: You checked Government Reports Announcements as well as ERIC. I really think ERIC will cover it. What appears here will be in RIE. In fact they list from RIE using the same indexing terms.

II. Identify Concepts

Discussion relates to identification and choice of the concepts which will form the basis for profile development. The concepts are often the major nouns or groups of nouns in the user's description of his area of interest.

Ex. A: Basically as I see it we would want or would set up three groups of terms. The first group would contain the term lepidoptera; we would put the terms morphology and histology in group two; and then we put thermone glands in group three.

A: I definitely see two concepts: Dramatics and English instruction, English curriculum, English classrooms, something, but dramatics and English... now I'm asking should we put in a third concept.

A: OK, so, now we need to go back and take your statement and break it into concepts. Now something that has bothered me is you've got polishing and ruthenium in single crystals. Now, in other words, the statement indicates three concepts--RUTHENIUM and POLISHING and SINGLE CRYSTALS.

Note: Discussion of concept identification often contains implicitly the assumption that concepts in the final profile will be linked by AND's. If discussion of search logic is explicit (such as the decision to exclude a concept using NOT logic), it should also be classified under Subproblem VI: Develop Search Logic.

III. Develop Content Search Terms

Discussion relates to generation and selection of lists of subject terms which are to be associated with each concept. Such terms can either be selected from a controlled vocabulary or expressed as words and phrases to match free text. Terms can be synonyms, broader or narrower terms or related terms all used to develop concepts previously identified. This subproblem also includes selection of subject codes, such as taxonomic codes in Biological Abstracts. The need to identify controlled vocabulary terms and subject codes depends on which data bases have been selected for searching.
Ex. A: OK, now is there any other way to express it, is there a synonym for platform deposits?

A: You don't know of any other names of the disease? Common names?
U: Snail fever.

A: Why don't you look under problem solving and see, in the thesaurus, what kind of things are given there?

A: Let's start with ERIC and work on counselor. I like to look back in the rotating descriptors first and find counselor and it tells you every way the term is used, beginning there, this is counseling. If you want counselor specifically, read back through that list please and tell me all the terms that would be helpful to you.

U: Counselor attitudes, ...

A: You have attitude change, attitude measurement?
U: Yes, attitude measurement.
A: And that code is ...
U: 10900.

A: Do you want this to be college students or just students?
U: College students.

Note: Discussion of term selection often contains implicitly the assumption that all terms associated with a particular concept in the final profile will be linked by OR's. If discussion of search logic is explicit (such as the decision to exclude particular terms using NOT logics), it should also be classified under Subproblem VI: Develop Search Logic.

Discussion related to selection of context items (author, journal title, language) should be classified under Subproblem IV: Develop Context Search Terms.

IV. Develop Context Search Terms

Discussion relates to the use of context portions of a data base record as part of the search profile. Context attributes of a document include its author(s), the journal in which it appears, the language in which it is written, and the date of publication. Any of these may be included as elements in a search profile and are especially important when the database to be searched is a citation index.

Ex. A: Would you like to request certain authors' names?
U: Well. Yeah. These people there. Can you pick them up?

A: Would you be able to give me any references on this?
U: The article where I found them?
A: Yeah.

Note: Discussion related to inclusion or exclusion in the profile of particular attributes (e.g. journal titles, author names) should
also be classified under Subproblem VI: Develop Search Logic, if it refers to the associated profile logic.

Discussion related to selection of content items (subject terms, subject codes) should be classified under Subproblem III: Develop Content Search Terms.

V. Refine Terms

Discussion relates to such activities as truncation of terms, assignment of weights, and inclusion of variant spellings or abbreviations. It may occur following selection of a content or context search term (Subproblems III and IV).

Ex. U: On that spelling, the word sulfide is sometimes spelled with an F and sometimes spelled with a PH. I don't know whether that makes a difference or not.
A: OK, well let's spell it both ways.

A: Now one question about these authors. I see you've got say, Richard A. Dobbs. Would any of the others ever use their first name in publications, because if they do we have to code it both ways.

A: Dormitory. D-O-R-M; I'll truncate that and pick up dormitory and dormitories.

A: What I'll do is we can weight your expressions, well we can probably weight them so that the citations on Cumberland Island come first and then the others dealing with the coastal islands come next and clay mineralogy comes at the end and you'll be able to pick it up by the different weighting factors on there so it will help organize the material a bit.

VI. Develop Search Logic

Discussion relates to analysis of the logical relationships between concepts and terms within concepts using the Boolean operators OR, AND, NOT. This may occur either during or sometime following the identification of concepts (Subproblem II).

Ex. A: You want something from group 1 and group 2 to appear together. You only want references which have both those terms in them.
U: Right.

A: Alright, so we've got the analysis of combustion products or just combustion products and then authors, now I'm just defining what we're looking for.

A: Would that be anything that mentions catechol?
U: Not catecholamines.
A: All right. Everything with catechol but not catecholamines.
Tutorial (label text as tutorial)

If the discussion is both directly related to profile development and tutorial in nature, it should be classified under the appropriate subproblem number (I-VI). If discussion is strictly tutorial, it should be labeled as such. Tutorial passages often begin with a question from the user about something he does not understand or with a question from the search analyst such as "Do you understand ________?"

The following passages are given as examples of the many topic areas about which the search analyst may instruct the user. These would all be classified as tutorial.

System Operations

A: The computer is not thinking. It's just matching characters.
U: Is it possible to amend the search at any time?
A: Right, and in fact, this is very frequently desirable.
A: Have you heard very much about this service?
U: No.
A: Would you like a very fast overview?
U: Yeah.
A: This UC campus is able to use the Center at UCLA and they have about four or five of the major science data bases available.
A: Once your searches start, you will get something every day for two weeks til you have a total of 22 printouts, and this is what we call a printout, and you get that printout even if you have no answers.

Data Bases

A: Let me back up and explain ERIC. ERIC consists of actually two printed publications. One is RIE and that's Research in Education and the other is CIJE and that's Current Index to Journals in Education. Now RIE is basically a report file, it does have some books, some dissertations, and CIJE is your journal articles.
A: Well now let me explain. They keep updating ERIC and it's updated quarterly so you get fairly recent stuff.
A: Now in doing an ERIC search we don't use your terminology, we use the terminology of ERIC, for which there is a thesaurus.

Controlled Vocabularies

U: What does UF mean?
A: Use for, in other words, use this term, English Second Language instead of ESL or EFL.
U: Now what did you say this meant?
A: This is a scope note, it just defines it.
Exclusions (label text as excluded)

If discussion is neither related to profile development nor tutorial in nature, it should be labeled as excluded from analysis. Such discussion can take many forms, as suggested by the following examples.

Social

A: Hello, how are you today? I see you're back for another search question.

Administrative

A: All right, now we need to find out if your mailing address is the same as it was the last time you came in here.

General

U: What are we going to do with all this knowledge?
A: In the computers, itself?
U: In the society.
APPENDIX B

INFORMATION FORM FOR USER

1. Date______________________________

2. Name______________________________________________________

3. Telephone: Area Code_____ Number_______ Extension___________

4. Address______________________________________________________

_________________________________________________________________

5. Position or Title______________________________________________

6. Are you a ___first-time user ____previous user of this service?

7. If you have used any computer-based search service other than this one, please give the name of the service:

_________________________________________________________________

8. Number of searches prior to the present one:

_____ None ______ 1-5 ______ 6-10 ______ More than 10

9. The search should: ___begin today and continue (current awareness)

and/or ___go backwards from today through the collection (retrospective).

10. Check the appropriate blanks which describe the person receiving the bibliography.

_____ Faculty _____ Academic Researcher

_____ Staff _____ Government Researcher

_____ Graduate Student _____ Reference Librarian

_____ Undergraduate _____ Technical Librarian

_____ Post-doctoral Appointment _____ Other: _______________________

_____ Industrial Researcher
11. Profile Title (short descriptive phrase): ____________________

12. Prose statement of search request:

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________
13. **Suggested Terms:** List any terms, keywords, synonyms, index terms, words or phrases that might be used to describe your search request. Be sure to include both scientific and common terms. If necessary, consult appropriate thesauri, subject indices and bibliographies. If convenient, group related terms according to major ideas in your search request.

<table>
<thead>
<tr>
<th>TERMS</th>
<th>SYNONYMS, CLOSELY RELATED TERMS</th>
<th>EXCLUDED USES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
14. Check the languages of references that are acceptable:

- Any language
- Russian
- English
- Italian
- German
- Spanish
- French
- Japanese
- Other: ____________________________

15. List any particular authors/corporate authors whose writing always is of interest to you (Specify first and middle names, if possible):

- ____________________________
- ____________________________
- ____________________________
- ____________________________

16. List any journals which are always of interest to you.

- ____________________________
- ____________________________
- ____________________________
- ____________________________

17. List any journals or other periodicals related to your question that you regularly read and from which you do not need to be informed of the contents.

- ____________________________
- ____________________________
- ____________________________
- ____________________________

18. If you are aware of any articles relevant to your search request, please supply as much bibliographic information, e.g. titles, authors, references as you have available.

- ____________________________
- ____________________________
- ____________________________
- ____________________________

19. If you were performing the literature search manually, in which library publications would you expect to find pertinent references?

- Bibliography of Agriculture
- Government Reports Announcements
- Bibliog. & Index of Geology
- Index Medicus
- Biological Abstracts
- Nuclear Science Abstracts
- BioResearch Index
- Physics Abstracts
- Chemical Abstracts
- Psychological Abstracts
- Engineering Index
- Sociological Abstracts
- Research in Education (ERIC)
- Social Sciences Citation Index
- Current Index to Journals
- Other: ____________________________
- In Education (ERIC)
- Unknown
20. Number of pertinent references you would judge to have been published in the previous twelve months:

<table>
<thead>
<tr>
<th>Less than 50</th>
<th>300-499</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-99</td>
<td>More than 500</td>
</tr>
<tr>
<td>100-299</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

21. Search Requirements:

- A broad search designed to retrieve as many of the relevant citations as possible, but which might also retrieve many irrelevant citations.

- A narrow search designed to retrieve primarily relevant citations, with few irrelevant citations but which may not retrieve some relevant citations.

22. Check the subject area(s) to which the question is directly related:

- Agriculture
- Biology
- Business
- Chemistry
- Computer Science
- Ecology
- Education
- Engineering (Mechanical, Civil)
- Engineering (Electrical)
- Fine Arts
- Geology
- Humanities
- Librarianship & Information Science
- Mathematics
- Medicine
- Physics
- Psychology
- Social Science
- Zoology

23. Check all the applicable blanks which most accurately describe the ultimate purpose of the search results:

- Research Project
- Bibliography for Publication
- Seminar
- Personal Bibliography
- Class Project
- Patent Search
- Term Paper
- Instruction or Teaching
- Thesis (Master's Degree)
- Other: __________________
- Dissertation (Doctoral Degree)

24. If this form is completed by other than the user of the search results, please give the following information:

Name: __________________

Address: __________________

Position or Title: __________________
APPENDIX C

DATA FROM INFORMATION FORM FOR USER

Case 1 - UCLA

1. Date: 5/28/74
2. Name: UCLA - 2058
3. Telephone:
4. Address: Anatomy Dept. USC School of Medicine Los Angeles CA 90033
5. Position: Assoc. Prof.
6. First time user of the service
7. Does not recall the name of other computer-based search service used
8. 1-5 searches prior to this one
9. Current awareness and retrospective search
10. Person receiving bibliography: faculty, academic researcher
11. Profile title: EM studies of cation localization in the anterior pituitary gland (or secretory tissues).
12. Prose statement of search request:
   I'm primarily interested in being alerted to the literature on cation localization by electron microscopy in the anterior pituitary gland or other secretory tissues. Secondarily I'm interested in any literature on electron microscopy of the pituitary gland. Thirdly I'm interested in physiological studies of hormone release from secretory tissues or enzyme release mechanisms.
13. Suggested terms: electron microscopy ultrastructural studies microprobe analysis anterior pituitary gland pars distalis hypophysis cation localization calcium magnesium secretory tissues
14. Acceptable languages: English, German, French
15. Authors of interest: Marilyn Farquhar, Burton Baker, Paul Nakane
17. Journals that need not be included: Journal of Cell Biology
18. Citation of relevant article: (none given)
19. Publications consulted in manual literature search:
   Biological Abstracts, Index Medicus

20. Number of pertinent references in last 12 months: less than 50

21. Search requirements: a broad search

22. Subject areas to which the question is directly related:
   biology, medicine

23. Ultimate purpose of search results: research project, bibliography
   for publication, personal bibliography

Case 2 - UGa

1. Date: 2/11/74

2. Name: UGa - 042590-001

3. Telephone:

4. Address: Poultry Science Dept. UGa Athens GA

5. Position: (none listed)

6. First time user of the service

7. No use of any other computer-based search service

8. No searches prior to this one

9. Current awareness and retrospective search

10. Person receiving bibliography: graduate student

11. Profile title: (none given)

12. Prose statement of search request:
   What are the effects of dietary lipids on channel catfish
   (Ictalurus punctatus)

13. Suggested terms:
   lipids triglycerides diglycerides monoglycerides glycerol
   fatty acids polyunsaturated fatty acids essential fatty acids
   free fatty acids nonessential fatty acids saturated fatty acids
   linoleate linolenate
   CIS-4,7,10,13,16,19-docosahexaenoic acid
   CIS-9,12,15-octadecatrienoic acid
   tallow manhadan oil lard
   corn oil safflower oil rape oil linseed oil soybean oil
   cottonseed oil peanut oil olive oil palm oil sunflower oil
   sesame seed oil
   cholesterol chylomicrons
   serum albumin
   bila salts
   channel catfish ictalus punctatus rainbow trout salmon bass

15. Authors of interest: J D Castell, D J Lee, R O Sinnhuber, J W Andrews, W G Knipprath, J F Mead, Nicholas Nicolaides, A N Woodall, Masanichi Toyomizu, J H Wales, J Halver

16. Journals of interest: (none given)

17. Journals that need not be included: (none given)

18. Citation of relevant article:

19. Publications consulted in manual literature search:
Bibliography of Agriculture, Biological Abstracts, BioResearch Index, Chemical Abstracts, Government Reports Announcements, Chemical Titles, CSSC

20. Number of pertinent references in last 12 months: 100-299

21. Search requirements: a broad search

22. Subject areas to which the question is directly related: agriculture, biology, chemistry, zoology

23. Ultimate purpose of search results: research project, thesis (master's degree), personal bibliography
APPENDIX D

SAMPLE TRANSCRIPTS

Case 1 - UCLA

A: I'm just going to say a little about your search and I made up a whole bunch of things I need to ask here. Alright. Interested in being alerted to literature on cation localization by electron microscopy in the anterior pituitary gland or other secretory tissues. Secondly, I'm interested in any literature on electron microscopy of the pituitary gland. Thirdly, in physiological studies of hormone release from secretory tissues or enzyme release mechanisms. One of the questions I had is for the calcium magnesium localization. When you say in other secretory tissues, are you interested in the other endocrine glands or any specific?

U: That would be a good place to start. Other endocrine glands, yeah.

A: Uh. I mean, like the adrenals, gonads, pancreas, parathyroid...

U: With respect to cation localization?

A: Yes.

U: That'd be locally, yeah.

A: Pineal... Is placenta one you consider?

U: I sure would.

A: Thymus and thyroid.

U: Well, the thymus is not endocrine.

A: It's not.

U: No.

A: I took this list from... I was looking at some books. I don't remember now.

U: Yeah, but thyroid, yes.

A: Alright. So, all the ones besides the pituitary which you mentioned.

U: Right. Right.

A: O.K. And I presume... Are you interested in any methodology or it'll just come anyway?

U: I'm interested in methodology too, yeah.

A: O.K. Fine. I think.. Excuse me. And then.. Alright, so that's
the cation localization particularly in the anterior pituitary and then all those other glands generally. And then, you said EM of the entire pituitary also as your second sort of parameter.

U: Well, primarily anterior pituitary gland, yeah.
A: Alright. So then I wondered if you meant, you know, other EM studies, not necessarily calcium and magnesium localization but just EM in general.

U: Of anterior pituitary glands?
A: Of anterior pituitary.

U: Yes.
A: Alright. Alright, just the anterior. And then, what about all the other goodies, uh, like the other glands, just EM?

U: Of other endocrine glands, you mean?
A: Yeah.

U: No.
A: No. O.K. So...

U: Not on a general nature.
A: Alright. That will be good then. I think it'll narrow that down. I need to understand. Alright, so you don't want that, no. Alright. The physiological studies of hormone release.

U: Right.
A: Maybe you can explain to me what kind of studies which...

U: O.K. What is it pri... The reason I said physiological is that I'm thinking about people that are not doing any morphology. They might take some pituitary cells or thyroid cells or whatever--incubate them.
A: So, in vitro?

U: In vitro. Stimulate them with something and measure the release. They might, you know, or any number of things like that, that I just for a title said physiological studies.
A: I wonder if it's necessary for me than to start naming all the different hormones secreted by all these glands or just do a general kind of, um--'cause I started off with the pituitary and I found the list I had forgotten from my own studies, you know...

U: Yeah.
A: 6 or 8 different ones. So that's just pituitary and then we have all these other glands too, uh..

U: To follow in what way do you mean?

A: Uh... In other words, to specifically name hormones that are just generally--because when you do say, you know, you expect like 1 to 50, it's sort of, sort of medium retrieval for you. It's only, comes out every..

U: 1 to 50 for the primary object is what I meant there.

A: The calcium of the cation mic..

U: For electron microscopy of the pituitary gland, there will be many..

A: Yes.

U: But I, I began to envision a hierarchy of things..

A: O.K.

U: And that maybe we could go in a couple of directions at once. The more limited one is cation localization.

A: Right. Yeah. The more para..

U: And that, I would anticipate over a year is going to be up to 50 and I'd be surprised if there are that many that fit nicely into my research.

A: I looked manually in Biological Abstracts yesterday in preparation for this interview just to see what kind of articles and yes, there aren't that many.

U: Right.

A: I looked under pituitary and then I looked under EM and, you know, I can't coordinate too well by eye. Sometimes you can, but..

U: Right, yeah.

A: There weren't that many, yes. But, the broad subject..

U: But expanding it to electron microscopy and expanding it in another direction to just in vitro studies or physiological studies of hormone release mechanisms or protein release mechanisms should fill, make a much greater number.

A: O.K. So, would that physiological side of hormone release and enzyme release then be for not only the pituit--if, I don't know if I asked you this--not only the pituitary..

U: Oh, that's right. For any secretory tissue.
A: Alright.

U: Yeah, that's right. That would have to be because people are using all, all kinds of different tissues.

A: O.K., then, uh, sal...

U: Salivary glands, they've... You know, whatever.

A: Ooh, I didn't have that did I?

U: Well, that's not an endocrine, and you were listing endocrine.

A: Oh. I see. So, but you are inter...

U: So... When... That's why I said other secretory tissues. I didn't mean endocrine. I meant like salivary glands.

A: O.K. I need to... Alright. What else?

U: Like.. Namely, like pancreas.

A: That I had. O.K.

U: Uh, well... You meant the endocrine pancreas I suspect, but I meant the exocrine pancreas too.

A: I'll put that down.

U: And the release of digestive hormones, enzymes from the gut.

A: O.K. Of course, you understand, I may be calling you up a few times. Once I start formulating--it's very nice to have this interview, but then when I start to playing with it, formulation questions may come up again.

U: Yeah, sure.

A: Alright. I presume you want human, animal, everything.

U: Yes.

A: Good. So in vivo and in vitro and normal and diseased states.

U: Yes. Yes.

A: O.K. I need to ask you what microprobe analysis is, I don't know that...

U: O.K. Do you see what I did there. Let me show you. That's.. I put electron microscopy, another centered on ultra-structural studies, and then I realize that, that there is a number, a growing literature on microprobe which is a specialization of electron microscopy and this is an accessory to an EM that allows a person to analyze an element such as calcium.
A: Oh, yeah. Oh my gosh.
U: O.K., and that's called the microprobe.
A: At what stage... Yeah. At what stage do you analyze it? In the live animal? In the... 
U: Oh, no. In the section. In, in a section...
A: Oh, once it's embedded...
U: It's an EM with another goody on it...
A: Oh.
U: And you do an...
A: Is it a quan.. Is it a quantitative thing? Or...
U: Qualitative and quantitative.
A: Both?
U: Yeah.
A: Oh, that's interesting.
U: Yeah, it's incredible. So, you might come across microprobe studies...
A: O.K.
U: Or X-ray analysis studies that would also be the kinds of things I'm after.
A: O.K., I looked in Index Medicus in their subject headings to see but perhaps one would just have to look under electron microscopy in Index Medicus and then articles on microprobe. Uh...
U: Yeah. Well, there's a lot of little different ways of, of talking about it. They might call it electron microprobe or they sometimes just say microprobe studies of the pineal gland or, you know—that kind of thing.
A: O.K. O.K. The philosophy of searching is so different with the one we're.. Oh, by the way.. I discussed and I think you were right that the BA would probably..
U: Fit my needs better.
A: Yeah. So, that's why I assumed. You did ask for a retrospective. I'm sorry to say that for this free study, I mean we, I can include it but we can run one for you here on line. This is going to be--this current awareness service is going to be..
U: For this next year, yeah.

A: Yeah— is going to be... Uh, we're going to formulate the search here and then send it to UCLA and they just throw it on the computer and it's run automatically, every issue of SA that comes out.

U: Right, right.

A: But, the retrospective, we can run here at, uh...

U: O.K. I, I wondered about that. I, I... That's why I put, if possible, because I didn't think that...

A: O.K. I know. I, I was going to say. If you're, you're willing and it's according to computer connect time—it's not cheap like Medline is. We access a commercial service, the National...

U: I see. Then the user pays a fee for co...

A: Right. So we don't charge ahead of time. We just charge you after the search is run. It's for computer connect time and the number of citations retrieved.

U: Oh.

A: So, it's a post-paid basis. And we try... You know, they average about $20. We try not to, you know, have it formulated very tightly ahead of time so we don't play.

U: Right. Right. Yeah.

A: Yeah. O.K. I think everything else that you've written here is quite clear, at least, for me. What to exclude and include. Um... And as far as the synonyms for anterior pituitary—those, I found myself too.

U: Alright.

A: O.K. I don't have any questions. Do you have any yourself on the system?

U: No. No, no. I thought you'd have a whole lot.

A: Yeah. Well, you know, I understand what you want.

U: Yeah. Fine. (Laughter)

A: After we got over the, you know, that you wanted everything secretory and then the physiological studies and that microprobe—those were really the questions..

U: Yeah.

A: Mostly that I had. I think that's fine.

U: O.K. Now, you might call me if you discover that you need some other kind of input.
A: I'm sure I will. Yeah. Right, because once I start formulating it and choosing words and coordinating things, I might need your help. Or if I'm finding it's just too enormous, we may have to just...

U: Right. Right. O.K. So, this is what I had no feel for.

A: Yeah.

U: That's what I was trying at the begin--to manage...

A: O.K.

U: The things in terms of first level priority, second level...

A: Yes, O.K.

U: And thinking you might say to me, hey—that going this way is going to be everything...

A: Yes, yes. I will say to you that I haven't had a lot of experience so I'll be having help formulating this. So, that—I can't tell you that this is much too much for a search. That, I'll have to tell you later.

U: Right. O.K.

A: O.K.

U: When is this thing supposed to start or when—we'll get a monthly alert or a weekly alert or?

A: It'll come out bi-weekly. I think 3A comes out bi-weekly. So, twice a month, you'll be getting—if you, we retrieve citations, you'll or... I think you'll be notified that you got zero this time...

U: Yeah.

A: Or what, so, twice a month I think can expect... Now, it'll just come directly from UCLA. We won't...

U: I see. They'll just mail it to me.

A: It's possible... Yeah. It's possible the first couple might come to us so we can look and...

U: Yeah, I see.

A: And by the way, if you don't like what's happening, we modify the profile all along.

U: Oh, fine.

A: We'd like to monitor it and have feedback from you because otherwise it's of no value if you're getting irrelevant things.

U: Right. Right.
Case 2 - UGa

A: Alright, before we start trying to determine exactly what you want, do you have any questions on anything, the questionnaire, or anything else in connection with the service at all? Alright.

U: Not that I can think of.

A: Well, you have seen a form somewhat similar to this but not nearly as detailed, and I believe we explained this in Dr. Reed's class some, so you have, well you are not completely cold to this type of request form. Alright, you are interested in the effects of dietary lipids on channel catfish. And some or all of the particular lipids that you are interested in you have listed under #13. Do you think it will be necessary for us to list all of these? Some of them you have individual lipids or acids, fatty acids, for example C15-4,7,10,13,16,19-docosaheaxenoic acid, this is one you are specifically interested in, so it would probably be a good idea of course to include that— you don't want to miss.

U: If you just put the term fatty acids, it pulls anything with that term in the title?

A: Well, in the title or keyword list also there may be some specific numerical codes that will be assigned to a paper that deals with fatty acids or lipids, however, theoretically it will pull any of those, but it may not, there may be errors involved or this sort of thing on the part of the reviewer, so if you do have someone like this, some fatty acid that you are particularly interested in, then it would be a good idea to list it so that you cover every possibility presumably of picking it up and tallow, menhaden oil, lard, and corn oil and so forth, I think it would be well to include them, certainly won't do any harm. You have one here that I'm not familiar with and not sure about the spelling C-H-Y-L-O-M-I-C-R-O-N-S.

U: Chylomicrons—transports the lipids through the system.

A: O.K. That is a single term?

U: Yes.

A: I think it would be a good idea here as you've indicated to include corn oil, safflower oil, and so forth than just say oil—you might get motor oil, crank case oil and things of this sort, of course that, well you do have oil down here don't you, what about that, might pick up such things as oil as a pollutant, lubricating, lubricating oils, would this be of any interest at all?

U: No, sir. It might be a good term to take out. 'Cause I've covered the oils I'm interested in pretty well.

A: O.K. I think perhaps it might be best to eliminate that one. Of course, there isn't any point to make oil a single term along with corn oil, safflower oil, and so on. If we are going to put in plain
oils we might just as well leave out these others, this would represent duplication, but I would gather that it would be preferable to list the specific individual oils. This is based on my concept of what you want, which may be wrong, particularly at this point.

U: I think it may be a good idea to take that out, I was just putting them down as I came to them.

A: Right. Now this is O.K. Alright, the channel catfish— I see you added some other fish here: rainbow trout, salmon and bass. What about the scientific names for rainbow trout, salmon and bass?

U: I have them for the trout, I don’t have them for salmon and bass.

A: Are you interested enough in those species that you want to get everything on them or do you, just sort of secondary...kind of like a sampling of the literature?

U: I’d like to get a good bit on them, to show, quite comparable with the catfish, otherwise...

A: Well, let’s see if I have any reference material on the scientific names, afraid mine doesn’t, won’t carry us down to the generic level. Salmon, salmonoides, salmonoidian, salmo, I suppose that would be the family name perhaps—salmonoides is suborder.

U: I think salmo might be the genus, because rainbow trout is salmos.

A: You did say you needed one for rainbow trout...

U: I can’t recall it, but it is later on in the paper— I put it down in the bibliography.


U: Yes.

A: O.K. If we put in just salmo though in addition to rainbow trout, you’ll get all other kinds of trout and salmon, too. I presume that might be a little broad for you?

U: Yes.

A: Would you like to see if you could determine the specific generic name, species name for the salmon and bass you are interested in and give me a call or drop back?

U: Yes.

A: Alright. Let’s see, let me just check on the bass—there apparently is a generic name that would cover all bass, but that probably would be too broad. OK, what we are going to do here is to essentially as you have indicated set up these two groups of terms—your lipid and fatty acids terms we will designate as one set of terms and then the
fish terms as the second set. And specify that there must be one
term from each group present in the keyword list and so forth of
the paper on lipids in order for it to be printed out. I will
also add some of these class codes, classification codes that may
be appropriate that I mentioned a moment ago. Alright, now then
you want both the retrospective and current search and you wanted
to continue and in that regard we will appreciate it very much if
you would keep in mind these searches on a current basis continue
indefinitely until we get some notification to stop. So whenever
you have all the information you want, by virtue of having covered
the subject matter area to your satisfaction, or changing your interest
or graduating or whatever, we would appreciate your letting us know
that we should stop it.

U: On that, at the end of spring quarter I'll be going down to Savannah
to do the research and I'll still be enrolled, will it be possible
for me to pay postage or something to have it sent there? I think
it will be at Skidaway Institute and they have an account with ya'll.

A: Yes, we can do this either one of three ways. Perhaps the best way
from our viewpoint would be to have your major professor or secretary
of the office in the department here simply to cumulate a few of them
and put them in an envelope and mail them to you. I imagine you will
be getting other mail in this manner, I suppose that this will be
expected. I suppose as an alternative we can simply transfer your
profile, change the address on it, put it under, actually transfer
it to Skidaway, a Skidaway number, and then it would have to go to
someone down there in their name and get sorted out again and if at
the end of the summer you come back up here, this would be perhaps
a little too much effort for a short time, but if you are going to
be there at least 6 months...

U: I'll be there for at least a year. And when I leave here in June
I'll be through up here and my main professor is at Skidaway.

A: Under these circumstances we can very definitely just transfer it
down there; now we want you to come in and tell us and remind me
at the time or of course you could wait until you get down there,
something of this sort, rather than coming by if you happen to
forget it in the rush of getting away. So if you are going to be
there that length of time, why definitely I think it would be better
to transfer it. Now, let's see, you checked here that you want
English and Japanese would you object too strongly if you get some
German? If the original article is in German or French, actually
it still has an English title and English abstract. Actually the
point is, even though this question is in here, in most data bases
such as Biological Abstracts it is really not feasible to worry about
language, because they do not in all cases specify what the language
is that the primary journal article is written in and if we put in
an English article and they don't say that it is written in English
even though it may be, then we wouldn't get it regardless of how
pertinent it was. So, it's just not too accurate or desirable,
ordinarily you get a title translation and a summary or an abstract,
you get a little something out of it anyway. Alright, now #15, list
of authors, I presume that these are people who do write about the subject that you have here. Does this mean, the way you interpret this, that you want every paper that each of these people have written even though it may not be specifically on the subject of the main concepts of your interest? Or if we get all of the papers on fatty acids and lipids in fish will that automatically include those papers by these authors that you are interested in?

U: That should include papers by them.

A: And anything else that these people may have written that's not on that subject you are not particularly interested in, right?

U: Right.

A: OK, alright, and do you have any questions?

U: No.

A: You are familiar with the fact that the output comes through the campus mail to you and let's see what else, Oh! particularly on the first output that you get look over the list of keywords, the search terms, rather, in the profile because these will be printed on the output, and look over them very carefully with respect to spelling, because I can make a typographical error and if we misspell a word then this will very definitely affect the search results. I had another point that I wanted to bring up...can't think of it right now, slipped my mind...well one thing we haven't talked about is what data bases would be appropriate here, Bibliography of Agriculture, are you familiar with the fact that this comes out as CAIN?

U: Yes.

A: Biological Abstracts, BioResearch Index and Government Reports Announcements. Would you presume that probably any article on this subject which might be, get into a chemical journal, and also probably picked up in the Biological Abstracts?

U: I was thinking about that, it might be a good idea to search Chemical Abstracts.

A: There's a good bit of nutrition papers cited in Chemical Abstracts, O.K., let's do so, what about the essential necessary work which may have been done prior to 1969, which is the, as far back as we go, in these which you have checked here.

U: Yes, there has been a good deal of work done before then.

A: Well, we can get back to 1965, Chemical-Biological Activities of CBAC and we can also jump back to 1962 in Chemical Titles. CBAC will carry a good bit of nutritionally-oriented material and consequently would be an appropriate source if you are interested in as far as 1965, the additional four years that we could get back in—Chemical Titles covers just about the entire chemistry field.
and search only on title terms. Consequently it might not be too thorough or too accurate, and your recovery may not be too great, but by virtue of the fact that it does enable us to get back four more years to 1962. Want to give it a try?

U: Yes.

A: O.K. on any of these, this Chemical Titles, you get one or two volume searches there and there's nothing in it apparently, if you let me know, we can stop it, of course, without any problems. Alright, will that take care of it, still no questions?

U: No.
APPENDIX E
SAMPLE PROFILES

Case 1 - UCLA

IN TITLE/TERMS/ABSTRACT
CHEMTERMS: CALCIUM*/MAGNESIUM*
MICROTHERMS: (ELECTRON MICROSCOP*/ULTRASTRUCTUR*/(MICROPROBE ANALYS*)/
(X RAY ANALYS*)
PITTERMS: (ANTERIOR PITUITARY*/HYOPHYSI*/(PARS DISTALIS)
GLANDTERMS: PANCREA*//(BETA CELLS*/(ALPHA CELLS*/GUVAS*/TESTIS*/
TESTICULAR/ADRENAL*/PARATHYROID*/ PLACENTA*/CHORION*/
THYROID*//(SALIVARY GLAND*/PAROTID*/SUBLINGUAL*/
SUBMANDIBULAR*/MAMMARY*/(SECRETORY TISSUE*)

IN AUTHOR
A1: FARQUHAR MARILYN/M*
A2: BAKER BURTON/B*
A3: NAKANE PAUL/P*

IN CROSS-CODES
C1: 01058$ "ELECTRON MICROSCOPY"
C2: 17000$ "ENDOCRINE SYSTEM"

IN CODEN
J1: ENDOAO/AJANA2/JHCYAS
J2: JCLBA3;
"ENDOAO ENDOCRINOLOGY/AJANA2 AMERICAN JOURNAL OF ANATOMY/JHCYAS JOURNAL
OF HISTOCHEMISTRY AND CYTOCHEMISTRY/JCLBA3 JOURNAL OF CELL BIOLOGY"

SELECT IF CHEMTERMS AND MICROTERMS AND PITTERMS;
SELECT IF CHEMTERMS AND MICROTERMS AND GLANDTERMS;
SELECT IF CHEMTERMS AND C1 AND PITTERMS;
SELECT IF CHEMTERMS AND C1 AND GLANDTERMS;
SELECT IF CHEMTERMS AND C1 AND C2;
SELECT IF MICROTERMS AND PITTERMS;
SELECT IF C1 AND PITTERMS;
SELECT IF A1 OR A2 OR A3;
SELECT IF J1 AND NOT J2;

Notes

Related terms are preceded by a label (e.g. CHEMTERMS).

Term type is indicated by the expressions beginning with "IN". For this profile term types include subject (i.e. title/terms/abstract), author, cross codes, coden.

Comments are enclosed in quotation marks (e.g. interpretation of the meaning of cross codes).

Truncation is of two types: * indicates that 0 or more letters may appear in this position; $ indicates that a single letter or blank may appear in this position.

The logic is given in the SELECT IF statements.
<table>
<thead>
<tr>
<th>GROUP</th>
<th>NO.</th>
<th>TYPE</th>
<th>WEIGHT</th>
<th>TERM</th>
</tr>
</thead>
<tbody>
<tr>
<td>G001</td>
<td>1</td>
<td>TXT</td>
<td>00000</td>
<td>CHANNEL CATFISH</td>
</tr>
<tr>
<td>G001</td>
<td>2</td>
<td>TXT</td>
<td>00000</td>
<td>ICTULURUS PUNCTATUS</td>
</tr>
<tr>
<td>G001</td>
<td>3</td>
<td>TXT</td>
<td>00000</td>
<td>I. PUNCTATUS</td>
</tr>
<tr>
<td>G001</td>
<td>4</td>
<td>TXT</td>
<td>00000</td>
<td>RAINBOW TROUT</td>
</tr>
<tr>
<td>G001</td>
<td>5</td>
<td>TXT</td>
<td>00000</td>
<td>SALMO CAIRDONERI</td>
</tr>
<tr>
<td>G001</td>
<td>6</td>
<td>TXT</td>
<td>00000</td>
<td>S. CAIRDONERI</td>
</tr>
<tr>
<td>G001</td>
<td>7</td>
<td>TXT</td>
<td>00000</td>
<td>BROWN TROUT</td>
</tr>
<tr>
<td>G001</td>
<td>8</td>
<td>TXT</td>
<td>00000</td>
<td>SALMO TRUTTA</td>
</tr>
<tr>
<td>G001</td>
<td>9</td>
<td>TXT</td>
<td>00000</td>
<td>S. TRUTTA</td>
</tr>
<tr>
<td>G001</td>
<td>10</td>
<td>TXT</td>
<td>00000</td>
<td>SALMON</td>
</tr>
<tr>
<td>G001</td>
<td>11</td>
<td>TXT</td>
<td>00000</td>
<td>ONCORHYNCHUS NERKA</td>
</tr>
<tr>
<td>G001</td>
<td>12</td>
<td>TXT</td>
<td>00000</td>
<td>O. NERKA</td>
</tr>
<tr>
<td>G001</td>
<td>13</td>
<td>TXT</td>
<td>00000</td>
<td>ONCORHYNCHUS TSHAWYTSCHA</td>
</tr>
<tr>
<td>G001</td>
<td>14</td>
<td>TXT</td>
<td>00000</td>
<td>O. TSHAWYTSCHA</td>
</tr>
<tr>
<td>G001</td>
<td>15</td>
<td>TXT</td>
<td>00000</td>
<td>CARP</td>
</tr>
<tr>
<td>G001</td>
<td>16</td>
<td>TXT</td>
<td>00000</td>
<td>CYPRINUS CARPIO</td>
</tr>
<tr>
<td>G001</td>
<td>17</td>
<td>TXT</td>
<td>00000</td>
<td>C. CARPIO</td>
</tr>
<tr>
<td>G002</td>
<td>18</td>
<td>TXT</td>
<td>00000</td>
<td>LIPID*</td>
</tr>
<tr>
<td>G002</td>
<td>19</td>
<td>TXT</td>
<td>00000</td>
<td>TRIGLYCERIDE*</td>
</tr>
<tr>
<td>G002</td>
<td>20</td>
<td>TXT</td>
<td>00000</td>
<td>DIGLYCERIDE*</td>
</tr>
<tr>
<td>G002</td>
<td>21</td>
<td>TXT</td>
<td>00000</td>
<td>MONOGLYCERIDE*</td>
</tr>
<tr>
<td>G002</td>
<td>22</td>
<td>TXT</td>
<td>00000</td>
<td>GLYCERIDE*</td>
</tr>
<tr>
<td>G002</td>
<td>23</td>
<td>TXT</td>
<td>00000</td>
<td>GLYCEROL</td>
</tr>
<tr>
<td>G002</td>
<td>24</td>
<td>TXT</td>
<td>00000</td>
<td>FATTY ACID*</td>
</tr>
<tr>
<td>G002</td>
<td>25</td>
<td>TXT</td>
<td>00000</td>
<td>LINOLEATE*</td>
</tr>
<tr>
<td>G002</td>
<td>26</td>
<td>TXT</td>
<td>00000</td>
<td>LINOLENATE*</td>
</tr>
<tr>
<td>G002</td>
<td>27</td>
<td>TXT</td>
<td>00000</td>
<td>CIS-4,7,10,13,16,19-DECOSAHEXAENOIC ACID</td>
</tr>
<tr>
<td>G002</td>
<td>28</td>
<td>TXT</td>
<td>00000</td>
<td>CIS-4,7,10,13,16,19-DECOSAHEXAENOIC ACID</td>
</tr>
<tr>
<td>G002</td>
<td>29</td>
<td>TXT</td>
<td>00000</td>
<td>CIS-9,12,15-OCTADECATRIENIOIC ACID</td>
</tr>
<tr>
<td>G002</td>
<td>30</td>
<td>TXT</td>
<td>00000</td>
<td>CIS-9,12,15-OCTADECATRIENIOIC ACID</td>
</tr>
<tr>
<td>G002</td>
<td>31</td>
<td>TXT</td>
<td>00000</td>
<td>TALLOW</td>
</tr>
<tr>
<td>G002</td>
<td>32</td>
<td>TXT</td>
<td>00000</td>
<td>MENHADEN OIL</td>
</tr>
<tr>
<td>G002</td>
<td>33</td>
<td>TXT</td>
<td>00000</td>
<td>LARD</td>
</tr>
<tr>
<td>G002</td>
<td>34</td>
<td>TXT</td>
<td>00000</td>
<td>CORN OIL</td>
</tr>
<tr>
<td>G002</td>
<td>35</td>
<td>TXT</td>
<td>00000</td>
<td>SAFFLOWER OIL</td>
</tr>
<tr>
<td>G002</td>
<td>36</td>
<td>TXT</td>
<td>00000</td>
<td>SESAME SEED OIL</td>
</tr>
<tr>
<td>G002</td>
<td>37</td>
<td>TXT</td>
<td>00000</td>
<td>RAPE OIL</td>
</tr>
<tr>
<td>G002</td>
<td>38</td>
<td>TXT</td>
<td>00000</td>
<td>LINSEED OIL</td>
</tr>
<tr>
<td>G002</td>
<td>39</td>
<td>TXT</td>
<td>00000</td>
<td>SOYBEAN OIL</td>
</tr>
<tr>
<td>G002</td>
<td>40</td>
<td>TXT</td>
<td>00000</td>
<td>COTTON SEED OIL</td>
</tr>
<tr>
<td>G002</td>
<td>41</td>
<td>TXT</td>
<td>00000</td>
<td>COTTONSEED OIL</td>
</tr>
<tr>
<td>G002</td>
<td>42</td>
<td>TXT</td>
<td>00000</td>
<td>PEANUT OIL</td>
</tr>
<tr>
<td>G002</td>
<td>43</td>
<td>TXT</td>
<td>00000</td>
<td>OLIVE OIL</td>
</tr>
<tr>
<td>G002</td>
<td>44</td>
<td>TXT</td>
<td>00000</td>
<td>PALM OIL</td>
</tr>
<tr>
<td>G002</td>
<td>45</td>
<td>TXT</td>
<td>00000</td>
<td>SUNFLOWER OIL</td>
</tr>
<tr>
<td>G002</td>
<td>46</td>
<td>TXT</td>
<td>00000</td>
<td>CHOLESTEROL</td>
</tr>
<tr>
<td>G002</td>
<td>47</td>
<td>TXT</td>
<td>00000</td>
<td>CHYLOMICRON*</td>
</tr>
<tr>
<td>G002</td>
<td>48</td>
<td>TXT</td>
<td>00000</td>
<td>SERUM ALBUMIN</td>
</tr>
<tr>
<td>G002</td>
<td>49</td>
<td>TXT</td>
<td>00000</td>
<td>BILE SALTS</td>
</tr>
<tr>
<td>G002</td>
<td>50</td>
<td>CXC</td>
<td>00000</td>
<td>10056*</td>
</tr>
</tbody>
</table>
GROUP  TERM  TYPE  WEIGHT  TERM
       NO.
G002   51    CXC  00000  10066*
G002   52    CXC  00000  13006*
G002   53    CXC  00000  13222*

BOOL EXP. G001&G002

Notes

Related terms are assigned the same group number (G001 or G002).

Each term is assigned a unique term number.

Term type is either TXT=text term or CXC=cross code (from Biological Abstracts).

Weights are all equal to zero.

An * appearing in a term indicates that 0 or more letters may appear in this position.

The logic is given following the labelBOOL EXP.

Cross codes in the profile have the following interpretations:
10056 = Lipids (biochemical methods)
10066 = Lipids (biochemical studies)
13006 = Lipids (metabolism)
13222 = Lipids (nutrition)
BIBLIOGRAPHY

Amarel, S. Response to The conceptual foundations of information systems. In E. B. Montgomery (Ed.), The foundations of access to knowledge. Syracuse: Syracuse University, Division of Summer Sessions, 1960, 94-99.


Atherton, P., Cook, K. H., & Katzer, J. Free text retrieval evaluation (RADC TR-72-159). Syracuse: Syracuse University, School of Library Science, July 1972. (NTIS No. AD-748 218)


Carman, J. L. Model the user interface for a multidisciplinary bibliographic information network (UC1-002-75-01). Athens, Ca.: University of Georgia, Office of Computer Activities, May 1975. (NIS No. 82-242 964)


Green, C. C. The application of theorem proving to question-answering systems. Stanford: Stanford University, 1969. (NTIS No. AD-695 394) (a)

Green, C. C. Theorem-proving by resolution as a basis for question-answering systems. In *Machine Intelligence* 4, 1969, 183-205. (b)


Moore, C. N. The next twenty years in information retrieval: some goals and predictions. Western Joint Computer Conference Proceedings, 1959, 15, 81-85.

Moore, C. N. Moore’s law, or why some retrieval systems are used and others are not. American Documentation, 1960, 11(2), 11.


Nourael, T. *User-directed relevance feedback*. Syracuse: Syracuse University, School of Information Studies, 1979. (Unpublished dissertation proposal)


Piganiol, P. Science and information in prospect. In UNISIST intergovernmental conference for the establishment of a world science information system. Final report. Paris, Unesco, 1971, 30-35. (a)


Preese, S. E. Clustering as an output option. Chicago: Illinois Institute of Technology Research Institute, 1976.


Robertson, S. E. The role of theory in the testing of information retrieval systems. In Informatics 3. London: Aalib, in press.


Shere, J. H. Darwin, Bacon and research in librarianship. *Library Trends*, 1964, 13, 141-149. (b)


Smith, L. C. Artificial intelligence in information retrieval systems. Information Processing and Management, 1976, 12, 189-222.


Unesco. UNISISTI: Study report on the feasibility of a world science information system. Paris, Unesco, 1971, (b)


Vickery, B. C. Document description and representation. Annual Review of Information Science and Technology, 1971, 6, 113-140. (a)


BIOGRAFICAL DATA

Name: Linda Cheryl Smith

Date and Place of Birth: January 27, 1949; Rochester NY

College: Allegheny College, Meadville PA
B.S., 1971 (Physics and mathematics)

Graduate Work: University of Illinois, Urbana-Champaign IL
(University Fellowship, 1971-1972)
M.S., 1972 (Library science)

Georgia Institute of Technology, Atlanta GA
(President's Fellowship, 1973-1974)
M.S., 1975 (Information and computer science)

Syracuse University, Syracuse NY
(Graduate Fellowship, 1975-1977)

Work Experience: Medical Library, Washington University, St. Louis MO
Trainee in Computer Librarianship, 1972-1973

Library, Georgia Institute of Technology, Atlanta GA
Literature Searcher, Information Exchange Center, 1974-1975

University of Illinois, Urbana-Champaign IL
Assistant Professor, Graduate School of Library Science, 1977-Present